UiT The Arctic University of Norway

Faculty of Science and Technology

Developing the MAR databases – Augmenting Genomic Versatility of Sequenced Marine Microbiota

Terje Klemetsen

A dissertation for the degree of Philosophiae Doctor – December 2021

Developing the MAR databases – Augmenting Genomic Versatility of Sequenced Marine Microbiota

Terje Klemetsen

A dissertation for the degree of Philosophiae Doctor

THE ARCTIC UNIVERSITY OF NORWAY • UIT •

Table of Contents

Acknowledgments

The work presented in this thesis has been carried out at the University of Tromsø – The Arctic University of Norway, Faculty of Science and Technology, at the research group Molecular Biosystems and Bioinformatics.

The period spent as a PhD has been exciting for me, in particular thanks to the engagement in the international ELIXIR project and database development. A situation where my creativity and sense of order could be challenged. For this I am glad to have had a supervisor as Nils P. Willassen to introduce me for these responsibilities. Several years with talks, database teamwork, and international interactions have inspired me. I am also happy to have had the chance to learn about the fish farming industry and contribute to analyzing related datasets, thanks to Christian R. Karlsen. My teaching duties in the introductory course in bioinformatics, lead by Peik Haugen, has also been interesting to me. Students can be challenging to work with, but having had the opportunity to advance the course content, to help students learn and understand concepts, has spurred my interests as a teacher. The final stage of my PhD and writing of this thesis has, nonetheless, been a challenge much due to the corona outbreak and following uncertainty.

I like to thank my colleagues, especially Christian R. Karlsen and Erik Hjerde, for commenting and proofreading the thesis manuscript.

Thanks to my friends and family for being there and supporting me through my time as a PhD student. A special mention to my South-American connection with Peru, which has given me many adventures and a whole lot of experiences. Thanks Nieves. I have ascended one top with this thesis, but just like the Andes, there will always be more hills and mountains to climb.

Terje Klemetsen
October 2021

Abstract

Interactions with databases are happening globally on a perpetual scale around the world wide web. In this web, databases are the invisible cornerstone of online search engines and data resources. Databases concerning genomic data play a significant part in scientific advances for diagnostics and classification purposes. Scope and focus vary from database to database. Some being all-encompassing, containing everything between the deep ocean trenches to the atmosphere. Others are focused on a specific topic like taxonomic classification of species. Nevertheless, the annual growth of these databases can reach logarithmic scales as sequencing has become cheaper and mainstream. Databases often find their application in analytical work. But in this context "more" does not necessarily mean "better", because it can imply additional compute time and redundancy. Key issues can be solved by balancing specific content, improve quality, and computation time to achieve favorable outcomes. The identification of bacterial communities and isolates, for example, are of greatest interest for researchers and substantiates a demand for accurate taxonomic classification. Databases today provide a crucial role in this by providing reference sequences for classification, either it is a single gene, multiple genes, a genome or a metagenome.

This thesis introduces the MAR databases as marine-specific resources in the genomic landscape. Paper 1 describes the curation effort and development leading to the MAR databases being created. It results in the highly valued reference database MarRef, the broader MarDB, and the marine gene catalog MarCat. Definition of a marine environment, the curation process, and the Marine Metagenomics Portal as a public web-service are described. It facilitates scientists to find marine sequence data for prokaryotes and to explore rich contextual information, secondary metabolites, updated taxonomy, and helps in evaluating genome quality. Many of these database advancements are covered in Paper 2. This includes new entries and development of specific

databases on marine fungi (MarFun) and salmon related prokaryotes (SalDB). With the implementation of metagenome assembled and single amplified genomes it leads up to the database quality evaluation discussed in Paper 3. The lack of quality control in primary databases is here discussed based on estimated completeness and contamination in the genomes of the MAR databases.

Paper 4 explores the microbiota of skin and gut mucosa of Atlantic salmon. By using a database dependent amplicon analysis, the full-length 16 rRNA gene proved accurate, but not a game-changer in taxonomic classification for this environmental niche. The proportion of dataset sequences lacking clear taxonomic classification suggests lack of diversity in current-day databases and inadequate phylogenetic resolution. Advancing phylogenetic resolution was the subject of Paper 5. Here the highly similar species of genus *Aliivibrio* became delineated using six genes in a multilocus sequence analysis. Five potentially novel species could in this way be delineated, which coincided with recent genome-wide taxonomy listings. Thus, Paper 4 and 5 parallel those of the MAR databases by providing insight into the inter-relational framework of bioinformatic analysis and marine database sources.

Abbreviations

16S rRNA	16S ribosomal RNA
AAI	Amino-Acid Identity
ANI	Average Nucleotide Identity
CV	Controlled Vocabulary
DDBJ	DNA Data Bank of Japan
DNA	Deoxyribonucleic acid
DOI	Digital Object Identifier
ECO	Evidence and Conclusion Ontology
ELIXIR	European Life-Sciences Infrastructure for Biological Information
ENA	European Nucleotide Archive
ENVO	Environment Ontology
FAIR	Findable, Accessible, Interoperable, Reusable
GAZ	Gazetteer Ontology
GO	Gene Ontology
GSC	Genomic Standards Consortium
GTDB	Genome Taxonomy Database
INSDC	International Nucleotide Sequence Database Collaboration
LPSN	List of Prokaryotic names with Standing in Nomenclature
MAG	Metagenome-Assembled Genome
MixS	Minimum Information about any (x) Sequence
MLSA	Multilocus Sequence Analysis
MMP	Marine Metagenomics Portal
NCBI	National Center for Biotechnology Information
NGDC	National Genomics Data Center (China)
NGS	Next-Generation Sequencing
OTU	Operational Taxonomic Unit

PATRIC	The Pathosystems Resource Integration Center
Q50	Phred Quality Score (99.9% probability of correct base)
RAS	Recirculating Aquaculture System
RNA	Ribonucleic acid
SAG	Single Amplified Genome
SARS-COV-2	Severe Acute Respiratory Syndrome Coronavirus 2
WGS	Whole Genome Sequenced
WoRMS	World Register of Marine Species

Part I – Thesis

1 Introduction

1.1 Bioinformatic databases

1.1.1 Primary infrastructure and the sharing of sequence data

Data and information availability are crucial for the scientific advance and ability to reproduce experiments. The vast amount of genetic data generated by present-day sequencing equipment challenge the way we archive meaningful biologic information. Current technology offered by the leading company Illumina, for example, is capable of sequencing 3000 gigabases in less than two days with their NovaSeq 6000 instrument [1]. Ordinarily, sequencing experiments generate computerized genome data that relate to the central dogma of molecular biology [2]. That is the organism's DNA, RNA and proteins involving the transcription and translation processes in the cell. Most sequencing data produced globally become deposited to partners of the International Nucleotide Sequence Database Collaboration (INSDC) [3]. This collaboration of interconnected, synchronized bioinformatic databases has been operational for nearly forty years and has contributed to sequence storage for a growing community of scientists. Institutions involved in the network include the European Nucleotide Archive (ENA) [4], the National Center for Biotechnology Information (NCBI) [5], and the DNA Data Bank of Japan (DDBJ) [6]. Another archiving institution is the China National Center for Bioinformation (CNCB) with repositories held in the National Genomics Data Center (NGDC) [7]. Currently, the NGDC forms the core resources of a unified Chinese collaboration, but does not constitute a partnership with the INSDC.

Nonetheless, essential aspects of the INSDC partnership involve the global synchronization of sequence data from public experiments. Along with access points for data submission and free admittance to published sequence data, the INSDC collaboration provides the basic fundamentals for archiving genetic material. Data from sequencing and particularly raw sequence data from next generation sequencing (NGS) are highly demanding towards storage requirements [3]. In 2020 the NCBI reported a ten-time growth over the last four years, now exceeding 16 petabytes of data in their repositories. This number equals 16 million gigabytes or approximately 32,000 average laptop hard drives of today. With the vastness of sequence data, the partner institutions become hubs for scientists engaging stored data. On an average day in 2019, the European Bioinformatics Institute (EMBL-EBI) resources experienced 62 million requests and throughout the year received page visits from 24 million unique IP addresses [8]. Graphs like these have pointed steeply upwards since NGS technology became accessible as a mainstream method. Cost effective NGS has, among others, deepen our understanding and expanded the known microbial diversity in less than a decade [9].

However, while the INSDC and NGDC partners are centralized and all-encompassing, they do not represent targeted genomic resources for specialized study topics, and neither provide harmonized infrastructures for sequenced organisms of directed environmental origin. This also affects the prokaryotic (bacteria and archaea) marine domain. Facts about the sampling environment play a decisive role as descriptive contextual data (synonymous with metadata) associating any sequence data with the sampling site [10]. Roots to this limitation by INSDC partners include the flexibility given the myriad of submitters, either individual scientists or institutions. The ability to submit unique attributes, provide text where numbers should be, and limited use of controlled vocabularies and ontologies has lead to inconsistency in the main repertoire of contextual data – limiting sophisticated probing of sequence data. Studies made in marine environments have linked unique bio-molecules and products useful for various biomedical research and product development. For example was the marine environment a source of 1277 novel chemical compounds published in 2016 [11]. Some compounds represent additional secreted substances, as secondary metabolites, aiding the organism in its survival. In this context, marine bacteria and fungi have been associated with unique molecules having potential bioactivities beneficial for drug development [12]. Advancing the availability and accessibility of marine genomic data can further support discoveries in the field and improve the taxonomic representation of marine microbiome diversity.

1.1.2 Current targeted resources for marine sequence data

From small virus sequences to whole chromosomes in eukaryotes, all require specialized and complex frameworks to manage and project their composite and layered information. Over the years, the development of dedicated resources has advanced the availability of specific topics concerning bioinformatic data. These complementary initiatives of the INSDC model have, among others, promoted the organization of biological knowledge for model organisms, taxonomy, proteins, and pathways. The catalog of biological databases, Database Commons (https://bigd.big.ac.cn/databasecommons/), lists many of these while providing an overview of worldwide repositories [7]. Accordingly, there exist over 5100 biological databases, distributed into thirteen categories, from 70 countries. In March 2021, sixteen of these host content specific to marine initiatives, nine of these are operational and accessible. The focus of remaining marine resources relates to documentation of climate, biodiversity, environmental data, contextual data, geographical mapping, anatomic imagery, and marine-derived chemicals.

Considering the marine sequence resources listed in Table 1, EBI metagenomics (later MGnify) exemplifies the broadest in terms of content and provides metagenomic datasets with analysis results from selected biomes, including the marine domain [13], [14]. Toolkits of the MGnify not only provide a metagenomic resource but a service pipeline performing assembly, analysis, and archiving of microbiome data in connection with ENA. Another database considering commercial marine fish species is the FishTrace catalog [15]. However, the focus is limited to European fish species, but provides marker gene sequences of mitochondrial cytochrome b for recorded organisms. Reefgenomics and Marine sponge compounds interactions (DESMSCI) are databases targeting organisms of marine reef biomes and provide genomic, transcriptomic, and chemical compound data from invertebrates like sponges, soft corals, and anemones [16], [17]. Not explicitly marine, however, the SalmoBase focuses on salmonid fishes and represents a genomic knowledge database contributing with annotation and expression data for reference genomes of Atlantic salmon (*Salmo salar*) and rainbow trout (*Oncorhynchus mykiss*) [18]. Similarly, while CrustyBase provides a BLAST database for a selection of ten crustacean species [19], the ConoMode is a sequence resource for conopeptides of venomous marine snails [20]. Other resources have utilized the vast data accumulated from the sampling initiative of the Tara Ocean Project. The GLobal Ocean 16S subunit web-accessible resource (GLOSSary) represents one such development [21]. It utilizes raw sample sequences from the Tara Ocean expeditions to organize and provide a marine prokaryotic marker-gene resource constituting the 16S ribosomal RNA (16S rRNA). The global ocean atlas followed up on this initiative with a gene catalog of microbial eukaryotes and zooplankton [22]. Here, the processing and compilation of metatranscriptomic data provide an atlas with over 116 million unigenes sampled from the ocean's euphotic zone.

Some resources covered here target broadly, like MGnify and GLOSSary, and some focus narrowly like Salmobase and ConoMode. What makes resources of a broad character possible lies in well-structured and standardized metadata. Examples can be drawn from the metagenomic samples of MGnify. The intervention from curators or the consistent registration of metagenomic datasets by submitters provides shape for this advantageous data structure [14]. Such harmonized descriptive data enables record-keeping of samples under a unified and strict system that can be further applied in super studies –

combining datasets. These data types can range from environmental classification to coordinates and depth measurements but remain confined to pre-specified formats. However, lack of uniformity is currently a considerable challenge for prokaryotic genomics and may explain why database initiatives have not undertaken the creation of a marine-specific resource. As illustrated by the listed resources, their conception demands a considerable effort in determining scope, handling contextual and sequence data as well as establishing auxiliary services for processing data and online publishing.

Table 1. Operational marine targeted resources providing sequence data.

Database	Scope	Data type	Reference
MGnify	All biomes	Metagenomic	[14]
FishTrace	European fish species	Cytochrome b marker genes	[15]
Reefgenomics	Invertebrates	Gene and protein sequences	[16]
DESMSCI	Invertebrates	Knowledge data, externally linked sequence data	[17]
Salmobase	Salmonid fishes	Genome knowledge database	[18]
CrustyBase	Ten crustacean species	Transcriptomic	[19]
ConoMode	Marine snails (conopeptides)	Protein sequences and 3D models	[20]
GLOSSary	Tara Ocean dataset (prokaryotes)	16S rRNA genes	[21]
Global ocean atlas	Tara Ocean dataset (eukaryotes and zooplankton)	Gene sequence data	[22]

1.1.3 Sequence data and taxonomic systems

When the submission of sequence data is deposited into the primary archives of the INSDC, it formally requires taxonomic labeling to describe the originating organism [23]. By describing the organism name, taxonomic lineage, and identifier, any prior knowledge and reference data on the organism becomes accessible. Awareness of this taxonomic information in context with the sequence data is pivotal for practically all research topics within life sciences. Thus, keeping track of sequences by taxonomic labeling is a significant effort of contemporary sequence databases. Still, the contents of bioinformatic databases only hold taxa represented by sequences, and thus comprise a sub-selection of known species in encompassing registers like the

List of Prokaryotic names with Standing in Nomenclature (LPSN) [24] or the World Register of Marine Species (WoRMS) [25].

Since the 1990´s the NCBI taxonomy database has been a repository linking sequences held in the INSDC with a standard taxonomic nomenclature [26]. Moreover, by the end of 2020, the NCBI taxonomic browser lists a total of 21,716 bacterial and archaeal species. The NCBI taxonomy is literature-dependent rather than sequence-based and labels each taxonomic node using scientific naming [23]. However, restrictions on formal naming obstruct sequences originating from uncultured species. Most prokaryotic species in the NCBI taxonomy are represented with at least the 16S rRNA marker gene – constituting a part of the cell´s ribosome. Due to its conserved nature among species and the practical composition of variable regions, the 16S rRNA gene has been applied frequently in taxonomic affiliation through phylogenetic analysis since 1985 [27]. The popularity of ribosomal marker genes for phylogenetic and classification purposes resulted in the establishment of targeted databases. These include NCBI [23], SILVA [28], Greengenes [29], and the Ribosomal Database Project (RDP) [30]. These represent general-purpose databases covering the broadness of kingdoms in Bacteria, Archaea, and Eukarya from all sources ranging from human skin, the rumen of bovine, air filters, lab surfaces, seafloor sediment, and so forth – causing a taxonomic influx in databases. This growth and increase in size challenge repositories by making them less practical for common use and has spurred the demand to cluster similar sequences in an ad hoc approach to improve their versatility [31]. By further considering the lack of resolution due to the conserved nature of the 16S rRNA gene and the taxonomic node inconsistencies within and between the ribosomal databases, there are considerable compatibility issues concerning the use of taxonomic systems [32]–[34].

The standardized taxonomy (synonymous with the Genome Taxonomy Database (GTDB)), in parallel to the NCBI taxonomy (Figure 1), attempts to harmonize the available sequence data as the baseline for a genome-based taxonomic classification [35]. Instead of using a single ribosomal gene for bacteria, it applies 120 ubiquitous single-copy genes (about 4% of an average bacterial genome) to construct a taxonomic rank system based on relative evolutionary divergence. Further use of average nucleotide identity (ANI) for estimating genome distances laid the foundation for separating species by a circumscription radius [36]. Consequently, the GTDB (Release 202) (https://gtdb.ecogenomic.org/) contains 47,894 representative species, where approximately 74.1% have placeholder names and are particularly recurrent for genomes of uncultivated species derived from metagenomes or single-cell sequencing. This approach, utilizing larger sections of genomes for classification, adjusted 58% of prior NCBI taxonomy classifications above the rank of species [35]. However, representing general purpose classification for prokaryotes, neither of the two independent taxonomies provide specific support for marine studies. Improvements in classification accuracy, coverage, and speed are conceivable by targeted taxonomic resources – avoiding potentially adverse interference from unrelated non-marine species.

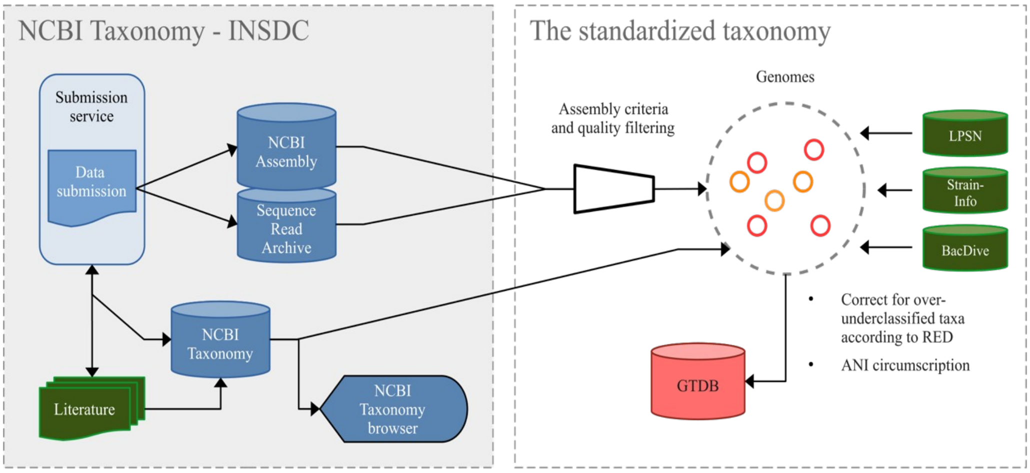


Figure 1. Brief overview considering the literature dependent NCBI and its genome dependent counterpart, the standardized taxonomy [23], [35], [36]. The latter circumscribe species based on genome ANI and draws data and taxonomic information from the NCBI and secondary sources like LPSN [24], BacDive [37] and StrainInfo [38].

1.2 Fundamental concepts in open bioinformatic databases

1.2.1 Find, Access, Interoperate and Reuse data

The intent and design of bioinformatic databases are to organize sets of sequence data accompanied with relevant information concerning its biological context, state of being, and provide unique and persistent accession identifiers. Further descriptions relating to the data collection, how it became processed, and finally analyzed can additionally be provided. Sequence data normally start as raw unprocessed reads and may be processed into genes, contigs and genomes, to mention a few applications. When a program process the raw reads, like performing an assembly, it advances the sequence data a step further to another level of complexity. The assembled data is not only altered in terms of sequence representation, as it conforms to other descriptive metadata. These may be attributes describing assembly statistics, software configurations and versions, and when the process was conducted. All attributes holding this information additionally requires clear definitions. For instance, an elementary attribute like the assembly date should explicitly detail how the value is expected to be inputted. Providing concise attributes can improve the data structures and enable meta-analysis across studies and sources [10]. Considering that data at these levels exist, how can it be found and further used? Normally, individual scientists and institutions provide their sequence and contextual data through INSDC submission systems, but guidance through submission brokering systems can aid in the process [37], [38]. Once made public by the authors, the INSDC database partners provide specific database designs for freely accessing study data like raw sequenced reads, assemblies, and gene sequences. ENA, NCBI, and DDBJ have committed to adopting the FAIR (Findable, Accessible, Interoperable and Reusable) data principles that encourage data to be found, accessed, interoperable, and reused [3]. These principles form guidelines for presentation and sharing data and accompanying metadata (see list below) [39]. The FAIR principles, first published in 2014, effectively extend earlier sharing policies stating that all listed data records must be freely accessible without restrictions or licensing requirements [40], [41]. The significance of the principles governs both the sequence and its relevant contextual data and its handling according to published criteria, particularly towards domain-relevant standards [39]. While the principles have firmly been established, the process of implementation and refining infrastructures is an ongoing process yet to be commonplace in life sciences. An overview of current principles can be explained through four points:

- *Findability* concerns the presence of descriptive metadata of any entry and particularly the unique, persistent, accession number or identifier associated with the sample. Another key factor is the discoverability of data for computer automation, also termed machine-attainable data.
- Gaining access to data and its descriptive metadata, for example, using a unique identifier represent its *accessibility*. However, authorization may still be required to access sensitive data.
- *Interoperable* data and metadata are represented by the use of accepted standards by the wide community, or globally accepted standards. This includes data formats, attributes, and ontologies that facilitate the exchange of data and contribute to abridge input/output between programs.
- *Reusable* data embody clear licensing for its use and present accurate metadata signifying its domain and provenance. In short, the data must be reliable to what extent it is described.

Interoperability of contextual data is an additional element of consideration by the FAIR project. The sample descriptions of individual projects are largely dependent on the domain in which they are relevant [42]. Accordingly, contextual information, particularly sampling data, is problematic in terms of achieving harmonized content structure and not expected to compare well amid projects. For example, the attribute field “depth” has independent interpretations based on its collection sites, like seawater, seafloor, soil, or other biomes. The expected field value as a number, text, boolean, and what unit may characterize the value is neither self-explanatory. Therefore, publishing defined attributes as community standards consequently endorse the advancement of interoperable contextual data – providing analogous database systems [40]. The checklists developed by the Genomic Standards Consortium (GSC) and their minimum information standards (MI) are examples of published biome-specific attribute collections [43]. INSDC partners have started to employ checklists on a project-basis in the submission process to describe the type of sequence data submitted. However useful, the checklists are scarcely implemented and enforced throughout life sciences. Checklists by the GSC include the Minimum Information about any (x) Sequence (MIxS), for marker sequences, for genomes, for metagenomes, and the more recent checklists for uncultivated samples [43]–[45]. These describe single amplified genomes (SAG), metagenome-assembled genomes (MAG), and uncultivated virus genomes. Checklists also support different environmental attribute packages. While the checklists are experiment-dependent, the attributes of the environmental packages cover the domain from which the sample originates. These include air, soil, water, sediment, host, human and plant-associated, but also specific packages for unique environments such as microbial mats and artificial environments. Packages have the additional effect of excluding irrelevant metadata by providing an extensive number of accordant attributes for the environment in question. Additional information explains fields deemed mandatory or optional for sample description. For instance, the water package for the MIxS checklist specifies 136 attributes relevant to the aquatic biome. Of these are 12 mandatory and 108 dedicated to the specific water environment where the sequence material originated. The remaining attributes constitute what makes up the experiment and sequencing-related data. The checklists further impose restrictions on some attributes by limiting on certain fields, like a set of choices governed by controlled vocabularies (CV). An example of the GSC MIxS water is the optional field ‘relationship to oxygen’ in which has a CV of seven choices like ‘aerobe’, ‘anaerobe’, and so forth based on known principles. Besides, the result of applying CVs can improve metadata consistency by limiting misspelling and non-standardized terms. Similar to CVs in the GSC checklists are subject-specific ontologies that represent improvements in data interoperability and machine automation for describing attributes [46].

1.2.2 Ontologies in databases as a means for standardization

Central to a database functionality is enabling human interaction by browsing the content. However, processing listed entries one by one is tedious, time-consuming, and inefficient to retrieve information from large-scale studies. Conversely, enabling machine-attainable data for automated systems significantly speed up database interactions. For instance, the field of proteomics has implemented ontologies as an initiative to standardize data content through CVs [47]. Ontologies share similarities with taxonomies by holding connected classes. However, the complexity of an ontology can be greater than the taxonomy due to the flexible relationships between classes and that classes may govern related but different domains (Figure 2**)**. These classes are particularly targeted for machine reading and enables data crawling. Classes themselves can act as defined vocabulary terms for databases and hold detailed properties. Interrelation with other classes in the ontology is not uncommon, and mapping between similar terms in different ontologies is also possible. In addition, the construction and maintenance of ontologies form under a feedback-oriented collaboration effort and their use is unrestricted, making ontologies dynamic [48]. The Gene Ontology (GO) [49], as one example, has become a prevalent ontology in bioinformatics for performing GO enrichment analysis between biological states [50]. From a database perspective, the MIxS water checklist has implemented the following six ontologies; Experimental Factor Ontology (EFO) [51], Ontology for Biomedical Investigations (OBI) [52], the Gazetteer ontology (GAZ) (environmentontology.github.io/gaz/), the Environment Ontology (ENVO) [53], Phenotypic Quality Ontology (PATO) [54], and Chemical Entities of Biological Interest ontology (ChEBI) [55]. This utilization of ontologies through the checklists underscores findability and interoperability of metadata data by following standardized sets of classes, which in terms facilitate improved machine autonomy [39].

As a practical example, the ENVO ontology standardizes environmental classes to systematically track sample origin [53]. Shaped from the upper-level Basic Formal Ontology (BFO) (https://basic-formal-ontology.org/), the ENVO defines an environment based on biomes, features, and material entities. These three complementary dimensions describe the site of sampling on a progressively narrower scale while ensuring a non-redundant characterization of the environment. Here the biome represents a locality where ecological communities are capable of adapting. Next, the environmental feature represents a landscape feature contributing considerably to its locality by having a causal effect on its surrounding setting. Lastly, the material entity defines the type of sample volume in which the collected sample is the most concrete class of the ontology. However, semantic systems as represented by ontologies rarely find its way to the contextual data of prokaryotic genomes. The utilization of ontologies complies with and advances the FAIR principles by providing findable and interoperable data, but still lacks uniformity for efficient use. Guidelines and checklists improve the quality of database content, but older data still linger without adhering to new principles. Still, with the increased application of guidelines metadata has largely remained unstructured. The management, authoring, and submission of contextual data can be cumbersome and time-consuming for scientists and is one reason for the contextual disorder [56].

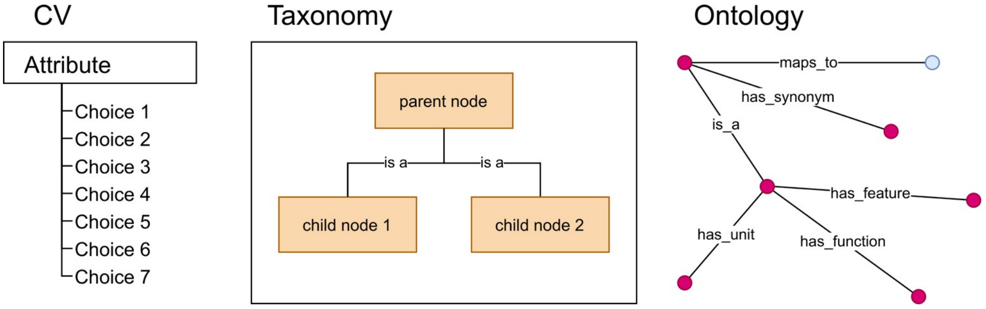


Figure 2. An attribute governed by a CV limits the number of possible choices and might be useful to exert control over certain attribute values. Such attributes can be less complex than those utilizing taxonomies or ontologies. While both descriptive and represent connected classes (synonymous with nodes and terms) of given domains, a taxonomy is limited to 'is a' relations between parent and child nodes. The increased complexity by node edges makes ontologies better optimized for computerized interpretation and enables mapping between ontologies.

1.2.3 Curation – adding value to databases

The structured systems comprising checklists, CV and ontologies introduced above form some of the instructional foundations for curation efforts made to databases. The data flow, nevertheless, originates from individual projects, where data become submitted into public repositories, as illustrated in Figure 3. Retrieval of targeted sequence data is often straight forward due to its availability in primary databases. However, the contextual data remain limited and minimal in most cases and require the curator's attention to accrue contextual information. Curation of metadata involves the collection of such contextual data values from sources like literature, authors, other databases, and registries. This information further supplements prior data of the given entry, add value where they are missing, and corrects inaccurate or faulty values. As has been noted, the absence of data values often results from contextual information not submitted in the submission systems in the first place [56]. Data submitters themselves are responsible for the original contextual information in main repositories like the INSDC BioSamples [57]. Biocurators however, from various life science domains and titles perform curation to enrich databases related to their project [58]. At any rate and depending on the database, the biocurator may further review content by administering attributes and oversee metadata to ensure harmonized and clean data in repositories. This may include changing the unit of values to fit the metric system and correct misspelled text entries. Where CVs and ontologies are implemented, the curator may also manage the terms related to entries in the database. A skilled curator can, in this way, improve the quality and increase the database value. To demonstrate, adding metadata to genomes of marine viruses has illustrated the effectiveness of manual curation. Using the MIGS checklist for viruses the authors succeeded by increasing the contextual content from covering only 21% of checklist values to 66% [59]. In this case, curating entries improved the amount of stored contextual data. Larger databases focusing on metadata, like the BioSamples, have also seen the value in curated entries. Normally, the BioSamples display author-dependent metadata, but have later enabled the overhead projection of externally curated information – promoting curated contextual data [57].

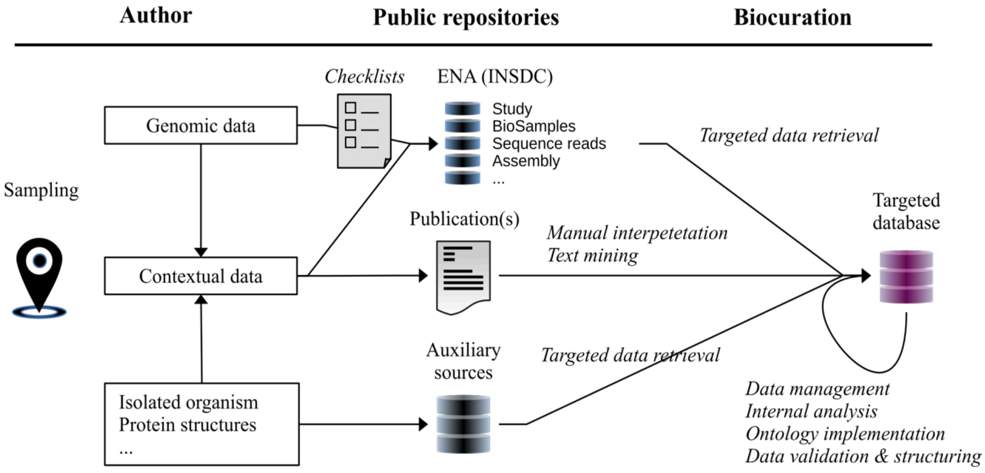


Figure 3. General data and information flow from the sampling, bioinformatic and data submission of authors to public storage and the implementation in targeted databases by biocurators.

Nonetheless, the curation effort is often demanding and time-consuming. Filtering and searching sources for relevant information are challenging biocurators due to unsurmountable large corpora [60]. Additional time consumption arises from the transfer of collected information into curation workflows and subsequent storage. However, with the amendment of existing data values or the addition of external data comes a responsibility to document information from third-party sources. Contextual data not provided by the relevant authors may potentially contain incorrect information. Providing the source material, as links or digital object identifiers (DOI), and modification details like corrected misspellings help substantiate data credibility but are generally lacking in bioinformatic databases. Nevertheless, the UniProt database [61] is a prime example of source documentation by having implemented the evidence and conclusion ontology (ECO) [62]. The ECO ontology represents a tool for advanced biocuration to systematize the annotation provenance and link supporting evidence [63]. In addition, as an ontology, the ECO is human and machine-readable and function as a labeling system for the source link. For instance, a human biocurator may extract information from a publication before adding it to a database entry. Instead of simply inputting the value, tagging it with the ECO:0007645 code, indicates that the inference was made by a curator from a published work. The value can further link with the relevant publication DOI. To point out, with full utilization of ECO, databases have the means to display the attribute value, its source, and by what means it became asserted. ECO tags also enable link assertions from analysis output results, as with the functional annotation of an unknown protein sequence. If applying BLAST [64] results in a successful annotation the ECO:0000044 can be associated with the given sequence. In this case, the code refers to sequence similarity as the evidence in which the assertion was made. This additionally exemplifies a process which can be automated for documentation and reproducibility purposes. For databases and their users, the practical application of ECO means extended potential when querying data for selective evidence and confidence in stored values [65]. As briefly mentioned, the implementation and use of ECO terms persist in database sources as UniProt, GO, and selected sources for model organisms [63], but remains less common elsewhere as in prokaryotic genomics.

1.2.4 Challenges in data storage: redundancy and contamination

A worldwide scientific community relies on bioinformatic sequence databases for analytic purposes like taxonomy, genomics, and metagenomics to answer critical research questions [66]. It is therefore of paramount importance that databases provide faultless sequence and contextual data. In a perfect-world situation flawless data would ensure any user absolute confidence in the repository content. This is, however, not the current situation. Various users and institutions deposited datasets on a daily basis to the centralized repositories of INSDC. The amount of submitted sequence data can prove demanding for database managers in respect to review and validation. Under these circumstances, faulty entries tend to linger in databases for a month on average before receiving amendments [67], [68]. In fact, approximately 70 prokaryotic genomes are found misidentified every month [23]. These require intervention in order to be revised. Leaving incorrect entries without attention can potentially compromise database content and further impact the users confidence in its data validity. Given that faulty data solidifies within INSDC databases, it may contribute to error propagation (Figure 4). This can affect projects utilizing the public data or become inherited within targeted or specific databases like UniProt [2]. The further the errors cascade from the original storage location, the more challenging they are to eliminate. Errors can be present in both sequence and contextual data, but need not represent critical problems for the database entry.

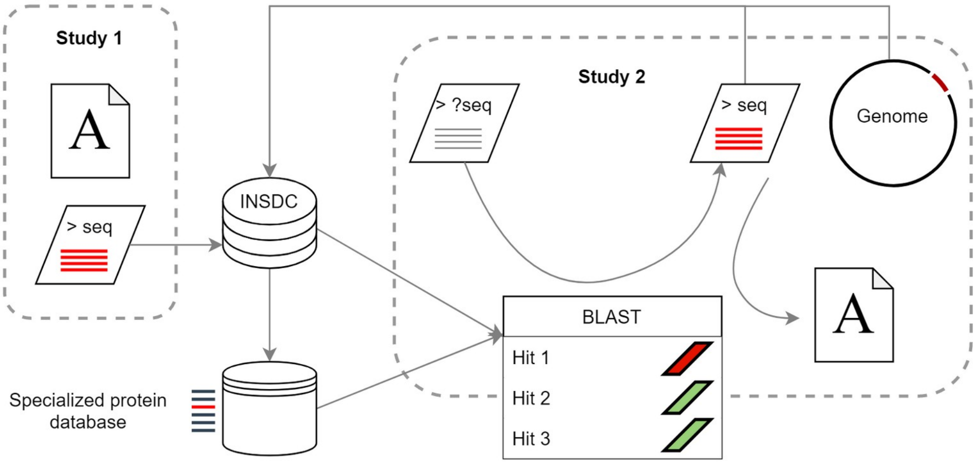


Figure 4. Error propagation of bioinformatic data. By publishing a faulty protein sequence (marked in red) in the INSDC repository (study 1) may result in its propagation to other sources like specialized protein resources and further act as reference data for auxiliary tools – here illustrated with BLAST. Their subsequent usage may further become embedded in the data of downstream studies and publications (marked here as A).

Contextual data describing sequences may contain errors or ambiguous information. This can be incorrect sampling details, organism, and taxonomic description, as well as genome metrics. While errors in the contextual sample description may not greatly compromise the entry, a taxonomic error e.g. providing incorrect labels for a sequence entry, can result in incorrect conclusions in forthcoming studies. This type of error has been documented for the Greengenes [29] taxonomic database. Its usage has lead to incorrect assignments between the orders Vibrionales and Alteromonadales, over representing the Alteromonadales, and could have affected 68 publications [34]. However, sequence data is subject to updates and can be resubmitted to correct or improve its status. For instance, the continuous work on the sequenced human genome has reached build 38. This version of the complete genome was published in 2013, 12 years after its initial release by the Genome Reference Consortium [69]. Because of this, sequence databases deal with dynamic content expected to change and update over time. While updates are beneficial for improving contextual and sequence data representation, problems linked with sequence data in databases still occur.

One such issue is redundancy in databases as the presence of duplicate or highly similar nucleotide or protein sequences. Complications can lead to unmanageable database sizes, longer querying time during searches, and subsequent longer manual assessment time. In addition, duplicates has been shown to impact conclusions and analysis for model organisms through bias in GC content and estimated DNA melting temperatures [70]. However, the deduplication of databases can alleviate the size issue. Clustering similar protein sequences reduced the NCBI non-redundant database to 56% of original size using a 90% identity threshold [71]. Then again, the challenge of removing redundancy is more complex than bluntly perceived. Simply deleting redundant sequences in an attempt to deduplicate databases may also interfere with natural redundancy in genomes. Thus, the definition of duplication may not directly relate to sequence content in databases, but in what context the duplicates exist [70]. However, sequence data can be undesired for other reasons with different implications.

Contamination in bioinformatic databases is emphasized as a growing problem in the last decade. In the case of metagenome-assembled genomes, the sequence material in question is considered contaminated if not correctly represented by essential single-copy genes [72]. Incorrect representations include the occurrence of sequence material from multiple species melded into one dataset, or if assembled sequence data is redundant and the given genome appears larger than expected. Still, contamination is not uncommon and can be present in most data types like genome assemblies, amplicon data, metagenomic data, and transcriptomic data [73]. Unintentional contamination can be caused by the actual sampling event, sample preparation, technical methods, and hardware related to sequencing, software, and data transfer [74]. Yet, for certain data types, contamination is expected. Environmental sampling, e.g. host-microbiota, is expect to contain some degree of contamination from the given host species. Cleansing of the sequence data is consequently an integrated part of the data process where applications remove host-related sequence material. Correspondingly, the bioinformatic removal of host sequence data is achieved by mapping against a host reference genome using tools like SortMeRNA and Bowtie2 [75], [76], or pre-sequenced chemical depletion of host cell material using lab kits [77]. Human-related sequences in sequenced prokaryotic genomes are one example of contaminants that require reference data for removal. In fact, sequence repeat regions from the human genome have been found in as many as 2250 bacterial and archaeal genomes of primary databases [78]. The study further linked these specific contaminants to incomplete reference databases for the human genome and, in particular, gaps caused by repeat sequences. In cases where prokaryote contaminant sequences represent sizable contigs, the prediction of open reading frames enables further annotation. Here annotations may specify gene locations, protein function, and to various extents, details as sequence domains and protein family information. With cost-effective determination of gene homology based on sequence similarity, the protein function becomes a target for error propagation since it bypasses experimental verification [2]. Thereby potentially emitting annotations from contaminated sequences to genomes if used as reference – aggravating the spread of incorrect information. In this way, proteins and coding genes additionally find ways inside secondary or specialized databases, making the contaminant data challenging to completely eliminate. With this in mind assessing and gaining knowledge on quality and sequence contamination is essential to maintain and strengthen database content.

1.2.5 Evaluation of genome quality and classification

The quality of a genome is linked with how complete or fragmented the representation is, and the potential contamination that resides in sequence data. Given that the finalization of a genome assembly is made by closing all gaps, there should theoretically be no, or minimal contamination in the genome [79]. However, the process of closing a genome requires considerable effort [80]. The added cost and time-consumption are unlikely adopted by all studies in the near future, but long-read DNA sequencing technologies, such as PacBio RS II and Oxford Nanopore are promising for mitigating the process [81]. Genomes are therefore frequently submitted in a draft state – fragmented into contigs or scaffolds with multiple gaps as a result of the assembly process. Draft genome assemblies greatly surpass the number of finished genomes in public repositories and pose a potential source of contamination. In 2015 the reported number of draft bacteria genomes in databases was six-fold to that of finished genomes [82]. Several metrics for determining the degree of quality in draft genome assemblies exist, and some essential statistics like raw reads, assembled contigs or scaffolds, are directly accessible from the assembly process. One of these is the sequence coverage which reflects the number of reads contributing and substantiating any given contig but may vary greatly within regions of a genome [83]. Low coverage caused by insubstantial sequencing depth may negatively impact the credibility of true nucleotides constituting the contig. At only two-times genome coverage there can occur up to four errors for every kilobase pair of DNA [84], which is roughly between 100- and 400-times the Q50 error rate precondition for finished genomes [44]. Numerous and very short contigs resemble low quality by leaving an excessive number of gaps, fragmenting the genome as a whole. Conversely, with the genome size roughly known, an assembly resulting in a few large contigs may indicate better quality with little fragmentation. The application of N50 and L50 contig lengths as metrics in bioinformatic assemblies represent approximations of draft genome quality. While the N50 represents the minimum contig length covering half the genome, L50 is the number of contigs constituting the N50. These represent a weighted median and prove less biased than the ordinary average or median, but still remain unreliable as a singular measure of draft quality [85]. However, these metrics are crude sequence measurements and are not suitable for detecting contamination.

The use of single-copy marker genes, as briefly mentioned above, are components in the later development of new tool sets for advancing genome quality estimates as well as phylogenetic inference [35], [86]. Any prokaryotic genome contains a set of genes contributing to the survival of the cell in its environment. There are housekeeping genes that constitute the basal subsistence of the cell and there are accessory genes providing extended endurance with a greater chance of being laterally transferred [87]. The selection of single-copy markers for tool analysis relies on universally conserved orthologous genes that rarely transmit via lateral transfer between cells [88]. Assuming that the presence of single-copy marker genes is unique in any genome, they have the advantage of uncovering contamination when detected in numbers. The emergence of metagenome-assembled genomes (MAGs) and single-cell amplified genomes (SAGs) from environmental sampling initiatives have spurred the need for rigid quality estimates as the sequencing techniques shaping these types are prone to contamination and incompleteness [89]. MAG genomes result from binning of metagenome assemblies and may inherit incorrect contigs leading to contamination. Contrarily, SAG genomes tend to have shortcomings from the limited amounts of available DNA material gained from a single cell and be less complete [44]. Several tools were developed for automatic quality assessment and include the Analysis and Visualization Platform for 'Omics Data [90], Protocol for Fully Automated Decontamination of Genomes [91] and CheckM [86]. The latter assesses the occurrence of single-copy genes in MAG and SAG type genomes to perform estimates of completeness and contamination. The most commonly used statistic from analyzing genomes or genome assemblies are completeness, contamination, and strain heterogeneity. The former two being the presence and duplication of single-copy markers. The latter represents a similarity measure based on estimates of amino-acid identity (AAI) between the gene-set material making up the contamination and compared gene-sets. Thus, it enables the interpretation of the phylogenetic distance between the analyzed genome and the origin of the contaminant species [86]. Endorsed as fairly comprehensible estimates, the completeness and contamination are included in classification schemes to make CVs for a summed quality assessment of genomes. By separating these as individual attributes Parks et al. suggested a set of four binate classes as listed in Table 2. The classes account for all possible combinations, from the detection of low completeness to high contamination in genomes. Later in 2017, Parks et al. introduced the term *Near-complete* representing the combined scores superior to 90% completeness and 5% contamination for a subset of 3438 recovered MAG genomes [92]. However, the GSC further elaborated the score metrics in an attempt to standardize the quality assessment through a checklist of controlled vocabularies [44]. Here the term *Near-complete* was not included. Exclusion of this term by GSC has not limited its use as it became a descriptive part of uncultured human gut bacteria [93]. The chief contrast between a *High-quality draft* and a *Near-complete* genome is the presence and absence of RNA genes, respectively. Without the ribosomal 5S, 16S, 23S RNA, and 18 ≤ tRNA; the *Near-complete* classification tends to fall on genomes otherwise considered *Medium-quality* drafts according to GSC checklists [44].

In contrast to the CheckM CV the GSC checklist categories hold no class associated with genomes having greater than 10% contamination. The INSDC repository states that the level of contamination in MAG/SAG genomes must be lower than 5% prior to submission [44]. However, there is currently no overview or knowledge regarding genome quality in primary databases using CheckM metrics (completeness and contamination). Completeness, contamination, and heterogeneity have not become integrated metrics of the INSDC databases for all genome types. Thus, little is known about the general condition of genomes in repositories and whether latent contamination is a cause for concern. Likewise, whole genome sequenced (WGS) genomes are prone to hold contamination but were not in the target group of the quality assessment by Bowers et al., but are in no way stopped from being assessed. Current state databases are potentially housing unattended contamination. Without performing quality assessment, particularly on contamination issues, public genomes remain unchecked and bypass user awareness.

Table 2. Vocabularies for classifying genome quality based on scores of completeness and contamination used and proposed by selected articles.

CheckM [88] controlled vocabulary

Completeness classification		Contamination classification	
90%	Near	≤ 5%	Low
70%	Substantial	≤ 10%	Medium
50%	Moderate	≤ 15%	High
< 50%	Partial	> 15%	Very high

GSC checklists [46]

Completeness		Contamination	Classification
NA	-	NA	Finished[1]
> 90%	and	< 5%	High-quality draft (SAG/MAG)[2]
50%	and	< 10%	Medium-quality draft (SAG/MAG)
< 50%	and	< 10%	Low-quality draft (SAG/MAG)

Classification as introduced by Parks et al. [94] and Almeida et al. [95]

Completeness		Contamination	Classification
90%	and	≤ 5%	Near-complete[3]

[1] Genome is represented as one contig with a base error rate of 1 in 10^5 (Q50) or better.
[2] Genome harmonize with completeness/contamination while having 5S, 16S, 23S rRNA and 18 ≤ tRNA.
[3] Genome harmonize with completeness/contamination.

1.3 Molecular systematics

1.3.1 Prokaryotic taxonomy and classification

In biology, the study of organisms largely depends on a backbone system describing the ordering and relation of organisms. System of rules like the *Systema naturae* developed by Carl von Linné during the first half of the 17th century introduced the initial concept of taxonomy by establishing a hierarchy of classes. It differs from the concept of phylogeny, which applies statistical measures to estimate evolutionary descent and relationship between samples. In an attempt to replicate natural order, a taxonomy tries to resolve biological classifications by introducing a taxonomic hierarchy following a system of rules and nomenclature [94]. These categories can serve as knowledge nodes operating reference points for classification purposes. Nonetheless, taxonomies are not rigid. They represent dynamic systems responding to changes in held categories, like during the introduction of novel taxa, or when studies gain knowledge that results in taxa being updated. Revisions including descriptions, nomenclature, and type material for prokaryotes are prepared and managed by taxonomists before reaching the scientific community [95]. As previously introduced in section 1.1.3 the database-stored 16S rRNA marker gene operates as a baseline for prokaryotic taxonomy [28]. It represented a breakthrough in the study of prokaryotic communities [96], where it has proven effective for taxonomic assignment of data from various environments [97]. Accordingly, utilization of the 16S rRNA gene sequence has accelerated our knowledge of prokaryotic community structures. However, the practical use of 16S is inconsistent and rarely applies the full-length of the gene despite the advantages it provides in terms of accuracy [98].

Utilization of the roughly 1500 bp 16S rRNA marker gene, however, is generally restricted to gene sections rather than its full sequence length during taxonomic classification [27]. The above-mentioned limitations can be ascribed to sequencing technologies using PCR primer-pairs resulting in short sequence stretches maxing at a few hundred bases. Consequently, studies of prokaryotic diversity utilize nine distinct hypervariable regions distributed over the gene length [99]. Considering the use of non-overlapping regions, it is likely that the outcomes of studies may become inconsistent [100]. These gene regions harbor variation in molecular stability and resolving power leading to inconsistent classification outcomes when originating from the same 16S rRNA gene. In comparison, the utilization of the full 16S rRNA gene sequence is a conceivable solution to alleviate the adversity of shorter, inconsistent regions. However, cost-effective methods in combination with satisfactory sequencing technologies to output the full sequence length at a desirable sequencing depth are capable of at least explaining major trends. One potential way of achieving near full-length sequencing of the 16S marker gene is using the PacBio sequencing technology, which has proven competitive to short-read sequencing in terms of error rate [101], [102]. The PacBio RS II technology is capable of sequencing longer stretches of DNA, with more than half the data being reads longer than 20 kilobases and reaching maximum read lengths up to 60 kilobases [103]. Further use of circular consensus sequencing, the PacBio can attain additional accuracy in its output potential. However, usage of the PacBio in 16S sequencing for the study of prokaryotic compositions remains infrequent compared to that of sequencing hypervariable regions and neither applied in the microbiota context of Atlantic salmon (*Salmo salar*).

1.3.2 The Atlantic salmon and its fluctuating microbiota

The farming of Atlantic salmon has been a rapidly expanding industry for the last two decades with a significant footprint in the commercial export for countries like Norway, Chile and Canada, but production has also increased in the Faroe Islands and the United Kingdom [104]. In Norway, for the full year of 2019, the export of salmon food produce reached 1.364 million tones achieving a first-hand export value of 68 billion NOK [105]. Thus, the fish health in farming facilities and hatcheries is of great concern in terms of animal welfare and counteracting production loss. Vaccination is one treatment known to positively impact survival rates [106]. Health additionally relates to microbiota, the microorganisms residing on external surfaces like the skin and mucosa, but also in relation to excreted feces. For instance, in fishes suffering from lice parasitism and showing disease symptoms, including infections of *Aliivibrio salmonicida*, intestinal microbiota has proven dissimilar to those of healthy individuals [107], [108]. Hence, knowledge of the composition making up the prokaryotic microbiota of farmed animals can be an important tool for research and development. But the microbiota composition is neither constant nor trivial. The plasticity of microbiota composition is influenced by the many life-stages of the Atlantic salmon; from egg to alevin, fry, smolt, and adult, which adds to the complexity of its microflora [109]. Fish at different locations including both wild and farmed seem to carry unique compositions of microbiota but tend to be consistent between populations [110]. During smoltification, the salmon undergo physical changes from life in freshwater (river-system) to adapt to a seawater lifestyle. This transition has proven to cause a destabilizing effect on the skin microbiota in artificial lab-regulated transition trials [111]. The study further noted an increase in microbial diversity after reaching the seawater phase and linked it to reduced levels of opportunistic bacteria. However, rearing fishes in artificial environments like hatcheries can also impact microbiota. Indeed, wild fish in their natural river habitat has indications of healthier microbiota than their counterparts in fish farms [112]. Here specialized bacteria in wild fish, thought to contribute with disease resistance and energetic conversion from food, seemed to diminish in hatchery fish. Besides, in farming facilities, the reared salmon is not the only bearer of microbiota. Some farms utilize recirculating aquaculture systems (RAS) to breed fishes using intricate filtration methods. Such RAS systems have been found to contain biofilms in tanks and filters containing broader microbial diversity than that found on the skin and in the digestive tract of Atlantic salmon [113]. This can be an important health factor as surrounding water affect the skin microbiota and, to a lesser extent, the intestinal microbiota of the skin [110]. Another aspect of the farming process is the choice of feed. It may impact both fish health, growth, and production cost, and the diet itself is thought to be the main contribution to variation in gut microbiota [110]. Plant-meal diets have become an alternative to marine-derived diets and represent a sustainable, cost-effective measure in the production of salmon [114]. Studies have shown increased abundance and diversity of the prokaryotic microbiota based on such carbohydrate-based meals, which are not a natural food source for Atlantic salmon, but short-term feeding has a minor impact on the microbial composition and mostly affects less prevalent bacteria [114], [115]. Diets can further be supplemented with antibiotics to prevent bacterial infections – a treatment directed at constituents of the microbiota but which can affect the community structure as a whole. One study on the intestinal effects of oxolinic acid and florfenicol found an ecological change in the microbiota, increasing the diversity and the proportion of Proteobacteria of the distal intestine [116].

1.3.3 Application of the 16S rRNA gene in amplicon analysis of Atlantic salmon

As presented, the microbiota constituting surfaces of the Atlantic salmon can be fluctuating and complex in its surrounding environment. From the skin to the various compartments of the intestine, the microbiota composition exhibits unique profiles [117]. The sum of knowledge on this topic has generally resulted from PCR methods sequencing ranges defined by primer-pairs targeting hypervariable regions of the 16S rRNA. Such as the studies of Gajardo et al. where the V1-V2 regions were used [115], [117]. The V3-V4 by Lavoie et al., the V4 by Llewellyn et al. and the V4-V5 by Wang et al. [108], [109], [112]. Other mentioned studies have either used the V4 or the V3-V4 regions [110], [111], [113], [116]. Jointly, these represent practical examples of varied, but inconsistent utilization of the 16S gene with the same purpose of prokaryotic taxonomic classification in Atlantic salmon. Obtaining knowledge of microbial compositions in different environments is still an ongoing process, and it remains unclear whether different variable regions and databases have a significant impact on the outcome. Most studies applying 16S rRNA regions, choose to cluster sequences into operational taxonomic units (OTUs) under a threshold of 97% sequence identity [118]. Testing of mock communities has proven the accuracy of the regions V2, V4, and V6-V7 to be consistent, but lacked the resolution to attain families and genera in some taxonomic cases as *Enterobacteriaceae* [119]. Nonetheless, the V1-V2 region showed poor taxonomic consistency when profiling the microbiota of activated sludge [120]. Microbial communities related to marine plankton samples constitute higher rates of rare bacterial classes and a higher proportion of *Pelagibacteraceae* using the V6-V8 region [121]. With these regions, Archaea was considerably less frequent and the phylum Euryarchaeota absent compared to the V4-V5. Arguably, the environment can reflect different compositions and correspondingly many conclusions due to inconsistent 16S rRNA usage. Preevaluation is suggested as important in determining what regions are best suitable for the given community [120]. Thus, the use of full-length 16S rRNA may alleviate the adverse situation of choosing regions. Application of full-length is less common and less understood compared to the use of variable regions and may attain more specific taxonomic classes due to more informative sequences. Anyhow, the classification of representative OTU sequences from variable regions can to a great extent, be assigned less specific taxonomic levels like kingdom and phylum. Still, this approach limits the understanding of the microbiota composition to vague trends. Following these unspecific taxonomic classes, assumptions can be made as the number of unique species and functions covered by a phylum can be considerable [122]. Most studies can achieve OTU classification as specific as the family level, or better, at the genus level. Using this classification method, soybean/wheat supplemented feed, for instance, has proven 18 times increase in lactic acid bacteria in salmon digesta compared to regular fishmeal [115]. However, a substantial proportion of OTUs may become unclassified at more specific levels in the taxonomy. In plant substituted feed, comprising a balanced carbohydrate/protein and a high carbohydrate/low protein diet, Villasante et al. presented intestinal microbiota samples having up to 50% of OTUs unclassified at the family level and close to 60% at the genus level [114]. Databases are fundamental in achieving taxonomic classifications [120], and OTUs remaining unknown may relate to missing references. This might be due to taxa not yet discovered and implemented in taxonomic reference databases. Other reasons may result from an inability to distinguish highly similar sequences and perform an accurate classification with reference sequences.

The Atlantic salmon microbiota is also known to be perturbed by bacteria, viruses, and parasites, causing a range of diseases with the subsequent loss of production [123]. In response, vaccine development has targeted some bacteria pathogens such as *Piscirickettsia salmonis*, *Yersinia rockery*, *Tenacibaculum finnmarkense*, *Flavobacterium psychrophilum*, *Vibrio anguillarum*, *Aeromonas salmonicida*, *A. salmonicida*, and *Moritella viscosa*. Furthermore, aqua culture production is susceptible to lice and lice-induced secondary bacterial infections. Studies on lice and antibiotic feeding have demonstrated elevated levels of bacteria from the *Vibrionaceae* family, holding the genera *Vibrio* and *Aliivibrio* [107], [116]. Knowledge depicting how species like these biologically interact with salmonids and what route of infection exist, can help developing treatments preventing infections. An important part of this knowledge lies in accurate and dependable techniques for identifying the presence of prokaryotes through taxonomic classification.

1.3.4 The *Aliivibrio* genus – current taxonomic standing

The current taxonomic system by NCBI position the bacteria genus *Aliivibrio* in the *Vibrionaceae* family, a member of the Vibrionales order and further situated within the Proteobacteria phylum [23]. This taxonomic lineage is built upon literature, but shares similarities with the genome-based standardized taxonomy [36]. Some radical differences for some classes do distinguish these taxonomies. For instance, the order Vibrionales is obsolete in the standardized taxonomy due to reassignments based on relative evolutionary divergence [35]. As a consequence, the order Enterobacterales act as an umbrella for 16 bacteria families including the *Vibrionaceae*. In the NCBI taxonomy, *Aliivibrio* is one of 15 genera in the *Vibrionaceae* and presently holds seven species with standing names in nomenclature [124]. These are *Aliivibrio fischeri*, *Aliivibrio finisterrensis*, *Aliivibrio sifiae*, *Aliivibrio thorii*, *Aliivibrio wodanis*, *Aliivibrio logei* and *A. salmonicida*. Due to quality requirements, like completeness, contamination, and highly similar genomes, only five species *A. fischeri*, *A. finisterrensis*, *A. sifiae*, *A. wodanis*, and *A. salmonicida* are present in release 95 of GTDB [36]. Considering the advancement in technology, genus *Aliivibrio* has seen no major update or systematic revisions being conducted since Ast et al. conducted their sampling study 2009 and with the introduction of *A. sifiae* in 2010 [125], [126].

1.3.5 A brief history – from *Photobacterium* to *Aliivibrio* and the identification of new species

The genus currently named *Aliivibrio* has seen several amendments over the course of history regarding the naming of species and their place in the bacteria lineage. In this context, *A. fischeri* represent an essential component in the genus which has lead the evolution of the genus from the time of its description by Beijerinck in 1889 [127]. *A. fischeri* is particularly studied for its colonization and bioluminescence in the light organ of bobtail squids [128]. However, the former classification of marine luminous bacteria, were broadly assigned the genus *Photobacterium* based on their phenotypic similarities, including the *Photobacterium fischeri* by Beijerinck [127]. Later conclusions published in 1955 from wider studies on morphology, antibiotic resistances, vibriostatic agents, and curvature suggested *Vibrio* as a better representative for *P. fischeri* [129]. This updated description and naming as *Vibrio fischeri* was further corroborated in 1970 by Hendrie et al. and finally approved in the 1980 list of bacteria names [130], [131]. This taxonomic standing remained until the revision and reclassification of the genus in 2007, phylogenetically justifying its unique placement as the new genus "*Aliivibrio*" - meaning "the other

Vibrio" [132]. A brief timeline of major events concerning the genus is shown in Figure 5. *Aliivibrio* remains the current name of the genus to this date.

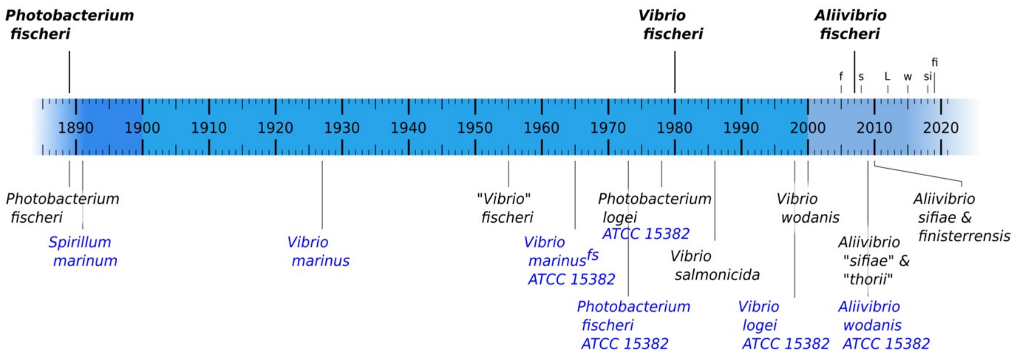


Figure 5. Timeline of *Aliivibrio* species with a leading emphasis on *A. fischeri*. Above the timeline represents the events of approved nomenclature change. Below the timeline are the publications representing introduction or change to species nomenclature. Strain ATCC 15382 and forma specialis (fs) association with *V. marinus* and later classifications are highlighted in blue. Smaller tick-marks indicate publications of fully sequenced genome of *A. fischeri* (f), *A. salmonicida* (s), *A. logei* (L), *A. wodanis* (w), *A. sifiae* (si) and *A. finisterrensis* (fi).

Also linked to the *Aliivibrio* genus, is a bacterial strain with a similarly long history as *A. fischeri* – yet with many twists and turns. Briefly, as illustrated in Figure 5, the bacteria *Spirillum marinum*, originally named and described in 1891, was later in 1927 amended to the species *Vibrio marinus* [133]. This would become the reference for strain ATCC 15382 isolated from the skin of a Pacific cod (*Gadus macrocephalus*). This particular strain was designated as forma specialis of *V. marinus* due to slightly different biological properties, e.g. higher maximum growth temperature. This classification became later termed invalidity due to inadequate descriptions of *V. marinus* and reclassified as *P. fischeri* on the basis of systematic phenotypic reviews [127]. Strain ATCC 15382 was in 1978 compared with a selection of ten marine luminous samples in which several inconsistent molecular and phenotypic properties became discovered [134]. Bang et al. delineated numerous strains formerly thought to be *P. fischeri* and, including strain ATCC 15382, suggested *Photobacterium logei* as a novel species. Recently, Manukhov et al. describe *A. logei* as a psychrophilic bacterium found to share similar genes as *A. fischeri* in its lux operon [135]. With the advent of 16S rRNA in the 1990s PCR primers targeting the V1 region aided in the classification of several bacterial samples from bobtail squid, including strain ATCC 15382 as a reference [136]. At this time, the V1 sequence size became the main evidence for keeping stain ATCC 15382 classified as *Vibrio logei*. In this same period, a study concerning the 'winter ulcer' disease afflicting farmed

Atlantic salmon in Norway identified *V. wodanis* as a causative agent [137], [138]. As a pathogen, the *A. wodanis* type strain (ATCC BAA-104^T) was later included in the phylogenetic analysis carried out by Ast et al. utilizing a multilocus sequence analysis (MLSA) of six housekeeping loci [125]. The conclusion was that strain ATCC 15382 was not *A. logei*, but *A. wodanis* or member of a new group came on the basis of MLSA sequence similarity. However, Ast et al. did not compare several strains of *A. wodanis* nor rigorously assess distances in terms of molecular evolution within and between included species. Disagreement with a former 16S rRNA study additionally questioned the true classification of strain ATCC 15382 [138].

As introduced in the previous section, the taxonomic diversity in genus *Aliivibrio* has expanded between intervals from the introduction and description of *A. fischeri* and later *A. logei* [134]. One of the driving forces in these discoveries has been in favor of the bioluminescent symbiotic behavior with fishes and squids. This continued with the publication of *Aliivibrio* sp. "thorii" which has representative samples from both seawater and the light organ of squids [125], [136]. This suggests bioluminescence in *Aliivibrio* species is not restricted to host interactions. In a similar manner, the description of *A. sifiae* puts it in the unique category within *Aliivibrio*. Sampling indicates independence of any host, but it exerts bioluminescence as free-living cells in a seawater habitat [125], [126]. Accordingly, four *Aliivibrio* species are currently expressing bioluminescent traits.

From the remaining, there are two species known to represent fish pathogens: *A. wodanis and A. salmonicida* described in 1986 and 2000 respectively [138], [139]. Both species are associated with industrial rearing of Atlantic salmon. While *A. wodanis* mainly contribute to an external ulcerative disease resulting in lower mortality rates, infections by *A. salmonicida* causes 'cold water vibrosis' and can lead to internal hemorrhages greatly impacting fish farms [138]–[140]. The final species of *Aliivibrio* is *A. finisterrensis,* which was described in 2010 with Manila clam (*Venerupis* (*Ruditapes*) *philippinarum*), and later, farmed Atlantic salmon as associated hosts [141], [142]. By description, *A. finisterrensis* is not bioluminescent and currently not considered a pathogen, but its association with siderophores for iron acquisition is a characteristics frequently found in other pathogenic bacteria [143].

1.3.6 Methods of classification and their advancements from the perspective of *Aliivibrio*

Most of the seven species currently constituting *Aliivibrio* (including *Aliivibrio* sp. "thorii") exhibit traits of interest like bioluminescence and pathogenicity. The discovery and study methods conducted on these species have also evolved over the course of time. The early years in taxonomic classification were largely limited to observable and measurable traits. Thus, bacterial cells like those of *P. fischeri* became classified in this way and in combination with several biochemical tests [129]. Advancing on these techniques, Bang et al. later delineated *P. logei* using a detailed description of morphology, optimal growth conditions, and compound utilization as a means of differentiation based on molecular and phenotypic properties [134]. Nonetheless, the usefulness and cost of assessing these traits may not be optimal in the long run. This can be viewed in context with the statistical limitation of using the mentioned descriptors in order to obtain phylogenetic resolution. By applying a combination of descriptors Reichelt et al., produced a phylogram in 1973 based on the numerical analysis of 147 carbon compounds as well as physiological properties [127]. Utilization of this statistical approach managed to split *Photobacteria* into three distinct clades – one of them representing *P. fischeri*. The diagram's resolving power, however, was limited compared with contemporary 16S rRNA gene analysis. In the context of Aliivibrio, 16S rRNA was first used to phylogenetically describe *V. wodanis* in the year 2000 [138]. Thus, *V. wodanis* became the first *Aliivibrio* classified with the assistance of the revolutionary 16S rRNA sequencing technology. Nonetheless, thorough physiological and biochemical tests ensured *V. wodanis* adhered to the modern concept of polyphasic taxonomy [144]. This briefly states that all available genotypic (DNA and RNA related methods), phenotypic, and phylogenetic data should form a consensus in which to conform the taxonomic conclusion. This principle of working with taxonomy has remained the chief approach in formally describing species. Notwithstanding, the material analyzed for inferring phylogenetic relations, the 16S rRNA gene and its regions, have become more flexible with the addition of other conserved genes [32].

Studies have indeed shown the 16S rRNA gene to provide little discriminatory power at the taxonomic level of genera and species making up the *Vibrionaceae* and *Vibrio* groups [145]–[147]. Attempts using alternative genes, like the ferric uptake regulation gene (*fur*), has proven useful for classifying species of *Vibrionaceae* [148]. These individual markers have the potential to advance the accuracy of classification in amplicon studies. Yet, they may fall short compared to the resolution of concatenated gene-sets in phylogenetic studies. Additional sequence data has the potential to strengthen inter-species resolution by providing a sufficient level of sequence variance. The optimal selection of genes for this purpose are universal and unique for all target strains, have conserved sites for PCR primers, and avoid sequences susceptible to horizontal transfer between bacteria [149]. In the 2007 study delineating *Aliivibrio, Vibrio,* and *Photobacterium,* the addition of four genes; *gyrB*, *rpoA*, *recA*, *pyrH* became used as a means to extend the level of resolution from simply applying the 16S rRNA gene alone [132]. Additional biochemical tests supported the resulting MLSA in distinguishing the three genera. These tests were omitted in the later phylogenetic study of Ast et al. As a consequence, the work leading to the identification and proposal for *Aliivibrio* sp. “thorii” and *Aliivibrio* sp. “sifiae” as novel species was unable to formally describe them [125].

Compared to the delineation by Urbanczyk et al. the gene constituents used in the analysis by Ast et al. included the additional genes *gapA* and *luxABE.* The latter represent three segments of the lux operon involved in the bioluminescent trait of marine bacteria like *A. logei* [135]. However, the use of trait-specific genes is not customary. Beaz-Hidalgo et al. constructed a similar MLSA design as Urbanczyk et al., but opted to use *atpA* instead of *gyrB* in their design [141]. Furthermore, the MLSA contributing to the official description of *A. sifiae* comprised the same core genes as included by Ast et al. but omitted the *luxABE* [126]. A summary of genes and their use in phylogenetic studies of Aliivibrio is shown in Table 3. Most MLSA studies of *Aliivibrio* have evaluated the consensus phylogeny of included gene sequences, but have provided less knowledge of how individual genes perform. In their description and delineation of *A. sifiae* with other *Aliivibrio* species, Yoshizawa et al. indeed evaluated both the consensus gene set (having a concatenated length of 4195 bp) and its individual gene constituents [126]. All genes except *gapA* managed to differentiate *A. sifiae*. However, by not considering *Aliivibrio* sp. “thorii” a full inference of *Aliivibrio* still remains to be seen. The same study lacked sequence material from *A. finisterrensis* for the phylogenies inferred from *gyrB* and *gapA*. Although Yoshizawa et al. provided the most complete phylogenetic study of *Aliivibrio* to date, the overall delineation of species is inconsistent in terms of gene-sets used. Several of the genes applied have been linked with poor resolution, like *gapA*, *recA*, *gyrB*, and *pyrH*, based on how well they inferred monophyly in *Vibrio* species [147]. Monophyly can be indicative of useful genes, but do not directly relate to phylogenetic topology inferred from a concatenated MLSA. As introduced, eight different genes have been in use. The 16S rRNA gene, *rpoA*, *recA* and *pyrH,* which has persisted as the core set while *gyrB*, *atpA*, *gapA*, and *luxABE* are infrequent. Yet, estimates of topological congruency between individual genes and a robust MLSA for *Aliivibrio* remain unaccounted. In addition, no composed minimal set of genes achieve optimal resolution close to that of an MLSA. This may prove valuable for classification purposes of *Aliivibrio* species as well as to reduce the time and cost of amplifying poorly performing genes.

Table 3. Genes and their protein products are included in phylogenetic studies of genus *Aliivibrio* (Urbanzcyk et al. [133], Ast et al. [126], Yoshizawa et al. [127] and Beaz-Hidalgo et al. [142]).

Gene [1]	Full name / protein product [1]	Cell function [1]	Used in publication			
			[133]	[126]	[127]	[142]
16S rRNA	16S ribosomal RNA	Translation	X	X	X	X
gyrB	DNA gyrase subunit B	Isomerase	X	X	X	
rpoA	DNA-directed RNA polymerase subunit alpha	RNA synthesis	X	X	X	X
pyrH	Uridylate kinase	Phosphotra nsferase	X	X	X	X
recA	recA bacterial DNA recombination protein	DNA repair	X	X	X	X
gapA	Glyceraldehyde-3-phosphate dehydrogenase A	Glycolysis		X	X	
luxABE	Bacterial luciferase, Long-chain-fatty-acid--luciferin-component ligase	Quorum sensing		X		
atpA	ATP synthase subunit alpha	ATP synthase				X

[1]Information based on UniProtKB, Swiss-Prot-listed data [191] for *A. fischeri*.

Data from individual type strains, as means of species reference, have been common ground for more recent studies [125], [126], [132], [138], [141]. As a name bearer of nomenclature, the type strains contribute greatly to comparisons [150]. A type strain is usually designated the first strain of a newly discovered species and work as a placeholder for its formal description and characteristics. It further tends to be used as a biological reference in culture collections. However, this designation of type strains does not emphasize the genomic diversity that can exist within species and contribute to their adaptations [151]. For that reason, a type strain by itself may not be an ideal representative for a given species due to its inability to represent interspecies variability [147]. Instead, the species concept coined ‘phylo-phenetic’ is advocating monophyletic and genomically coherent clusters of organisms, sharing a high degree of similarity [150]. This approach of circumscribing independent species based on sequence similarity in multiple strains remains unaccounted for in the *Aliivibrio* phylogeny. While studies have measured the intra-species sequence distance between species type strains using the 16S rRNA [126], [132], inter-species distances, separating strains of the same species, continue to be omitted regardless of the sequence material analyzed. Utilization of multiple strains can therefore aid in the holistic knowledge regarding a species or taxonomic class of interest. A forerunner to species circumscription can be exemplified by the standardized taxonomy, a systematic use of pairwise ANI measurements between genomes [36], [152]. The few genomes currently representing *Aliivibrio* hampers the possibility of ANI methods. Performing MLSA represent a cost-effective intermediate between phylogeny based on 16S rRNA and genome-wide ANI [149].

2 Aims of the study

The primary goal of this work is to develop, maintain and evaluate targeted databases for marine prokaryotic genomics. Current-day bioinformatic repositories lack functionality, contextual detail, and quality awareness for providing robust marine reference data. Therefore, contributing with open marine prokaryotic databases for the scientific community demands a substantial curation effort and the commitment to gold standards and FAIR principles for data sharing. From the end-user perspective, the databases should be freely accessible as an online service. From here provide firm and validated reference data with relevant attributes for samples collected in defined marine environments.

In parallel, the secondary goal is to evaluate the taxonomic classification of prokaryotes in mucosal surfaces of Atlantic salmon and review the current phylogeny of the bacteria genus *Aliivibrio* – harboring pathogens of salmon species. Central in these analyses is the use of 16S rRNA. As a conserved gene sequence it is disputed for lacking accuracy. This work will look at methods to extend knowledge and accuracy in classification and phylogenetic resolution. Correspondingly, these studies will be considered as potential use cases in terms of database application (primary goal) and taxonomic awareness in regards to phylogenetic fidelity.

3 Included papers

Detailed in this section are the papers considered by this thesis and the involvement I have contributed with in the wake of their finalization.

3.1 Paper 1

Title	The MAR databases: development and implementation of databases specific for marine metagenomics
Authors	Terje Klemetsen, Inge A. Raknes, Juan Fu, Alexander Agafonov, Sudhagar V. Balasundaram, Giacomo Tartari, Espen M. Robertsen and Nils P. Willassen
Description	This paper detail the development of the marine prokaryotic databases MarRef, MarDB, and MarCat. The web service hosting the databases, the Marine Metagenomics Portal, is additionally introduced and described in terms of functionality.
Contribution	The database design, including definitions and which attributes describing marine samples was part of my contribution. I also took part in the contextual data curation and writing of the manuscript.
Date of publication	2. November 2017 (Online)
Publication status	Published in Nucleic Acids Research
Citation	[153]

3.2 Paper 2

Title	The MAR databases: A manually curated resource for marine microbial genomics and metagenomics
Authors	Terje Klemetsen, Juan Fu, Alexander Agafonov, Sudhagar V. Balasundaram, Espen M. Robertsen and Nils P. Willassen
Description	This paper is a report on the MAR database project as of summer 2020. It details the database size expansions, workflow, functional novelties, implementation of analysis tools, infrastructure improvements, and the new salmon and fungal databases SalDB and MarFun.
Contribution	I participated with the design of SalDB and MarFun and took part in the data curation. I also performed the comparative analysis of the taxonomic component and participated in writing the manuscript.
Date of publication	-
Publication status	Manuscript for Nucleic Acids Research
Citation	-

3.3 Paper 3

Title	A substantial quality assessment of prokaryotic genomes in the MAR databases reveals an urgent need for submission quality control
Authors	Terje Klemetsen, Espen M. Robertsen and Nils P. Willassen
Description	The Paper reviews genome quality of entries listed in the MarRef and MarDB databases. It further partition genomes according to recovery method and examine how these conform with public classification schemes.
Contribution	My contribution involved the interpretation of analysis result and the writing of the manuscript.
Date of publication	-
Publication status	Manuscript for Bioinformatics
Citation	-

3.4 Paper 4

Title	Full−length 16S rRNA gene classification of Atlantic salmon bacteria and effects of using different 16S variable regions on community structure analysis
Authors	Terje Klemetsen, Christian R. Karlsen and Nils P. Willassen
Description	This paper examines the technical aspect of employing the full 16S gene for taxonomic classification of microbiota in Atlantic salmon. In this setting, two commonly used regions of the gene are compared against the full length equivalent.
Contribution	I performed the bioinformatic analysis and structured the resulting output from the amplicon sequence data. I also interpreted parts of the results and took part in writing the manuscript.
Date of publication	4. July 2019
Publication status	Published in MicrobiologyOpen
Citation	[154]

3.5 Paper 5

Title	Phylogenetic Revision of the Genus *Aliivibrio*: Intra- and Inter-Species Variance Among Clusters Suggest a Wider Diversity of Species
Authors	Terje Klemetsen, Christian R. Karlsen and Nils P. Willassen
Description	This paper examines the phylogeny of the taxonomic genus *Aliivibrio*. In particular, it attempts to clarify prior inconsistencies, place and describe current species using contemporary available sequence data.
Contribution	I contributed by collecting sequence data, perform the bioinformatic analysis, interpreted the results and writing of the manuscript.
Date of publication	18. February 2021
Publication status	Published in Frontiers
Citation	[155]

4 Results and Discussion

This work describes the inception and development of the MAR databases as a genomic resource of marine microbiota. Since the spring of 2016, a team of dedicated people has provided unique expertise and contributed to the design process, content collection and functionality of the MAR databases[1]. The development process has taken place at UiT – The Arctic University of Norway, Center for Bioinformatics.

The content of this work is split into three main sections. The initial phase of this project is described in 4.1 and considers the database design process and the curation of the first version. In 4.2, the advancement, including new data and functionality, is discussed along with the evaluation of quality in sequence data held by the MAR databases. The final section 4.3 discusses the two parallel studies and how these illustrate taxonomic applications and challenges for the MAR databases. This includes how the databases may be used for classification purposes in amplicon settings with Atlantic salmon, and how the phylogenetic revision of the bacteria genus *Aliivibrio* can impact and improve taxonomic accuracy.

[1] Nils P. Willassen, Terje Klemetsen, Juan Fu, Mayeul Marcadella, Igor A. Molchanov, Sudhagar V. Balasundaram, Espen M. Robertsen, Alexander Agafonov, Inge A. Raknes, Giacomo Tartari

4.1 Marine genomic databases

The limited possibility to specifically select marine genomic data from primary databases spurred the concept culminating in a targeted resource for prokaryotic marine genomics. Contrasting all-encompassing primary databases of the INSDC [3], the MAR databases were intended to provide curated public data limited to samples of marine origin. Primary databases are ineffective in selective filtering of biomes, like marine samples, based on descriptive contextual data. This lack of functionality became a motivation for the database development and curation effort required to determine marine affiliation. Consequently, the purpose of developing the MAR databases was to promote marine sciences while providing tools, frameworks, and standards specific to the field. The development took place and was supported by the European life-sciences Infrastructure for biological Information (ELIXIR) at elixir-europe.org. The MAR database project was enrolled in the EU project EXCELERATE under work package six for marine metagenomics. It contributes to the ELIXIR appeal to build biological competence and infrastructure in Europe [156]. As a resource for projects and researchers in the marine domain, the MAR databases provides marine selective prokaryotic genomic data in Europe. Since its launch in 2018, the Marine Metagenomics Portal (MMP) (https://mmp.sfb.uit.no) has been the website hosting the MAR database services. The service has a goal of receiving updates with new features and database entries on a bi-annual basis. Updates include, among others, the use and adoption of standards and guidelines like the GSC MIxS [157] and FAIR data principles [39] to solidified the databases as dependable marine resources. To enumerate, since its inception until the end of September 2021 the MMP have had a total of 16,424 unique users visiting the web service. Interactions from these have amounted to 28,069 sessions with 121,486 page views. The effort and quality, particularly regarding MarRef, has additionally been recognized by external services. Repositories like the WoRMS database [25] as well as the ENA Biosamples [57] provide refferences to the MMP for concurring data entries.

4.1.1 Designing the MAR databases

Some of the motivations behind the marine databases are best explained through the lack of functionality in large resource databases, particularly concerning genomics. These are repositories connected by the INSDC [3], as the ENA [4], the NCBI [5] and the DDBJ [6]. For genomic data, there is no feasible way to choose all samples of marine origin, or any other origins for that matter. Still, services provided in primary databases include published study data from various project submitters. These tend to hastily become published in public repositories and subsequently incorporated in database services like BLAST [64]. By favoring flexibility in the submission process, scientists have been mostly free to describe a sample as they see fit for their individual projects. Parts of the submitted contextual data relevant to the given genome are recorded in the BioSamples database systems [57]. Contextual data of the BioSamples often represent some principal information concerning the sampling event, as location, biome, habitat, and host species. Although accessible, the contextual data provided by submitters are not structured sufficiently, as will be discussed later. Guidelines, exemplification, and standardization during submission have been lacking for years, leading to various interpretations by whom is submitting data – further inducing inconsistencies. Ultimately, this obstructs pragmatic filtering of genomes from their sampled habitats in broad, primary databases. We overcame this by manual undertaking a curation effort to assess publications and validate entries based on their contextual information, ensuring that the MAR databases was designed to provide a prefiltered selection of marine prokaryotic genomes.

4.1.2 Requirements for entry implementation

The initial steps in the database design involved defining the principal concepts shaping their purposes. Two main requirements defining a sample entry in the MAR databases became as follows; (1) the presence of an assembled genome or metagenome and (2) its contextual evidence of marine origin. While the first requirement depends on accessible genomic data in primary database sources, the second requirement rest on how the term *marine* is defined.

In context with requirement (1), a sequenced genome can receive varying levels of attention depending on its author(s). Which in terms will impact the final genome representation and depends on the study or project purpose. For example, there is a significant effort to fully sequence and close a genome assembly, and provide evidence of plasmids, which normally requires specific experimental equipment [158]. However, closed genomes are highly representative and advantageous for accurate comparative genomics and transcriptomic studies. This can be costly compared to reference-free assembly of draft genomes. Draft representations of genomes are, on the contrary, sequenced and assembled de novo into contigs. These cover a variable fraction of the genome while often constituting unordered contigs, much due to repetitive elements in the genome [159]. While drafts are on the lower scale of genome quality and attention, programs facilitating improvement through scaffolding and reference mapping have enhanced their representation [81]. Considering the possible genome representations, in **Paper 1**, we introduced the database MarRef to hold closed genomes while MarDB became the database for complementary, non-complete genomes. Henceforth, MarRef came to serve the purpose of holding the highest quality genomes – those having the status as ‘complete’ or ‘finished’ represented in terms of genomes with closed chromosomes and plasmids. This ensured MarRef as a robust reference database. Given these points, MarRef is consequently less comprehensive and shares similarities with databases like RefSeq [160], having defined requirements for the acceptance of genomes. Other marine genomes were assigned the MarDB database holding entries of lower quality, complementing MarRef in the coordination of all verified marine genomes. In **Paper 1**, metagenome samples of marine origin were additionally considered. The database housing these was named MarCat and became the third database hosted at the MMP. Its purpose became the management of a gene catalog derived from metagenomic samples. Combined, the three databases MarRef, MarDB, and MarCat were intended to provide selective resources of sequenced data from genomes and metagenomes that are open and accessible to the marine-focused research community. However, the identification and retrieval of marine genomes into the databases are dependent on the second point in the requirements defining the marine realm.

What constitutes a marine habitat tends to be loosely considered. Still, it remain fundamental in defining specific attributes to describe each recorded genome entry. Defining seawater samples from discrete, enclosed sea and ocean boundaries justify the bare bones to be considered. Simply disregarding coastal areas would streamline the filtration of marine samples, but as a consequence, ignore relevant sampling locations like sandy beaches and estuaries. Enforcing a strict definition could be suitable to individual projects but does not apply to a wider community interested in marine-related prokaryotes. To emphasize, the coastal border between marine and terrestrial areas is highly varied and is the outcome of natural mechanisms like geodynamic processes, the changing sea level, coastal erosion and deposition, marine modification and terrestrial inheritance [161]. The definite coastal zone boundary, as defined by Ray et al., has a biogeographical and a jurisdictional boundary [162]:

1 "The terrestrial boundary is defined by (a) the inland extent of astronomical tidal influences, or (b) the inland limit of penetration of marine aerosols with the atmospheric boundary layer and including both salts and suspended liquids, whichever is greater."

2 "The seaward limit is defined by (a) the outer extent of the continental shelf (approximately 200 m depth), or (b) the limit of territorial waters, whichever is greater."

As given by 1), the biogeographical definition, the coastal zone can to a great extent be influenced by the marine. Therefore, it supports the inclusion of coastal areas covered by the marine microbial biome as defined in **Paper 1**. This involves transition zones like estuaries and mangroves where the level of salinity is perturbed by seawater. Equally important became specific sampling materials not fitting neither the explicit marine nor the coastal definitions. A rather challenging sampling site for the definition includes those originating from the drilling of deep subseafloor sediments. Important to realize that such deep-seated samples may harbor diverse ranges of extremophilic microorganisms that are different to known surface sediments [163]. Another consideration was the hypersaline environments of coastal solar salterns. Yet, halophilic prokaryotes and fungi are known to inhabit these extreme environments [164]. It can be argued whether these sites provide explicit marine samples or represent unique niches. Moreover, artificial or man-made constructs taking the shape of aquarium systems can maintain habitats mimicking seawater, estuarine, or other relevant biomes targeted by the aquarium design. These artificial enclosures may exist anywhere suitable on land. In conclusion, the definition of marine finally incorporated all samples from the mentioned transition zones and environments exposed to or mimicking marine salt water. Samples with contextual information agreeing with the definitions are thereby eligible for the relevant database depending on its representation as previously detailed.

4.1.3 Attribute design

Another central part of the database development involved the description of genome entries through clearly defined attributes. Information provided to describe a marine sample is naturally not the same as those taken from, for example, the lung of a hospital patient. Yet, some information can be covered by the same attributes, like the host species and location name, while attributes describing water depth and salinity are specific descriptors of the aquatic environment. In all, 106 attributes were implemented in the initial version of MarRef and MarDB as detailed in **Paper 1**. Following the FAIR principles, attributes became derived from the published MIxS [43] water checklist. Additionally, a consistent identifier to enhance entry findability and accessibility, the ‘MMP ID’, became designated for each unique genome entry. Further in favor of consistency, several attributes were revised to hold limited types of information or were given choices by CVs or ontologies. For example, the comprehensive environment ontology system (ENVO) [165] was implemented with 11, 59, and 25 terms for the environment biome, feature, and material attributes, respectively. As will be further exemplified later, limiting the free-text attributes of genomic entries enabled the filtration and search functionality unseen in primary databases like ENA [4] and NCBI [5]. The number of attributes has since been expanded as described in section 4.2.1.

4.1.4 Curation of genomic entries

Critical for the inclusion of entries into the MAR databases was biocuration. This involved the sourcing of contextual information from a number of auxiliary databases and publications, as described in **Paper 1.** The Pathosystems Resource Integration Center (PATRIC) [166] became the initial source to obtain data on sequenced genomes. Like the MAR databases, it only admit sequenced genomes. However, the PATRIC database does not constitute an INSDC partnership, but provided a consistent flat-format data structure partly derived from NCBI BioProject and BioSamples information. This became helpful and time-saving for the filtration of marine and non-marine samples (Figure 6) as well as in the curation process. Despite being a practical resource, depending on a secondary database like PATRIC may not be optimal for the future development. Like other services, secondary databases depend on funding and continuous upkeep to remain active.

Curators of the MAR databases associated the verified genomes and metagenome samples with their respective databases. However, for many genome entries metadata or sample description was inconclusive and hampered precise classification (marine or non-marine). Some samples could be indicative of marine origin, due to vague or incomplete metadata, and required careful examination of relevant publications for verification. For instance, if no marine indicator was given a sample from e.g. a host with the scientific name *Elysia rufescens*, it will require a curator to evaluate if the host species is marine by consulting a source like WoRMS [25]. When the entry lack detailed contextual data from PATRIC, the MAR databases are additionally benefited by accessing relevant publications to extract spatio-temporal sampling information and phenotiypic descriptions to enrich the data entry. The work of identifying relevant papers, collecting and transferring information is the main task of biocurators [60] – also those working on the MarRef and MarDB databases. Completing the contextual filtration of marine genomes gave 612 and 3726 prokaryotic entries in MarRef and MarDB databases, respectively. With the size difference and genome representation in mind, MarRef became feasible for exhaustive curation. MarDB, on the other hand, would serve as a semi-manual curated database through refinement of contextual data.

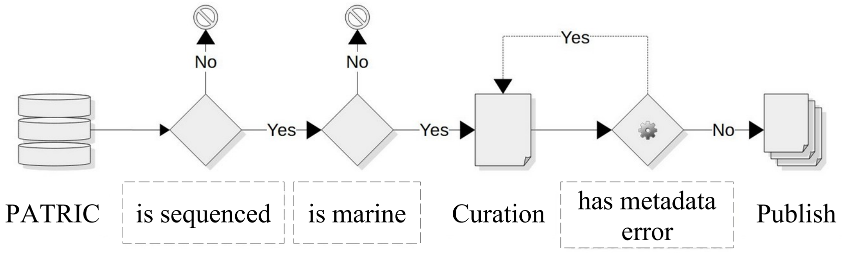


Figure 6. Simplified entry integration flow. Given that a prokaryotic entry has been genome sequenced and is verified as marine it is forwarded for curation. A flawless entry is then applicable for the forthcoming update. However, absolute verification of marine origin may happen during the curation.

During the curation, contextual data from BioProject, BioSamples, and documented information from publication sources needed consistent storage to enable advanced functionality. A caveat of primary databases, like NCBI, is attributed to flexibility enabling free-text to be entered. Taking the attribute 'depth' as an example, its application remains limited due to the acceptance of free text values. As an illustration, one study may choose to record the depth attribute as the text 'surface' while another may record the number value '0'. Thus, it becomes impossible to perform filtration on data given as both text and numbers from a computational standpoint. For primary databases, attributes are none-modifiable due to the proprietary right of the submitter [67]. On the contrary, the MAR databases were intended to enable functional filtering. For the depth attribute, describing the sampling depth below sea surface, this meant all values had to be numerical and using the same unit. The curator would therefore convert to the value to meter as a positive number describing the sampling depth. This enabled the selection of ranges while browsing data at the MMP. In combination with the curated ENVO terms for biome, feature, and material, the filtration system permits any user of the MMP service to choose genome data of any particular interest.

Hence, the MAR databases provide a registry of prokaryotic, marine-derived genomes, and a comprehensive, consistent data system capable of advanced contextual exploration. In relation to other marine databases mentioned in the introduction, the MarRef and MarDB fill the gap constituting prokaryotic genomics.

4.2 Advancement and evaluation of the MAR databases

4.2.1 Functional improvement of the MAR databases

The MAR databases are ongoing developments that include updating, adapting and improving both the front- and back-end services of the MMP. Finding and filtering new genome entries, acquiring sequence data, curation and data analysis are planned operations repeated bi-annually. **Paper 2** describe in detail the release of MarRef and MarDB databases version 5. All in all, from initial launch to the summer of 2020 these two databases grew by 158% and 355% entries, respectively. The notable increase in MarDB came as a consequence of multiple publications introducing MAGs. Several large-scale metagenomic sampling projects, like those of *Tara* Ocean, combine and perform binning to generate thousands of MAGs [92], [167]–[169]. One reason for including MAGs was to increase the diversity of listed marine prokaryotic species. Since uncultivable species can reside in metagenomic samples, MAGs have a potential to represent these and fill the gaps of hitherto unfamiliar Bacteria and Archaea. The organisms constituting these uncultivable species has been generally named biology's dark matter [170]. Along with the SAG recovery method, the MAG and SAG entries represent valuable and substantial resources for the MAR databases. Consequently, a repertoire of uncultivable marine prokaryotes may in this way have been implemented and serve as complementary genomes to the cultivated WGS entries. By broadening the assortment of genomic content in the MAR databases, the subsequent coverage of marine prokaryote species increases. This wider diversity may have improved the potential for MarRef and MarDB for classification purposes of marine samples. In terms of functionality, the introduction of genomes as either WGS, MAGs, or SAGs became an integral part of the MAR databases under the attribute 'Analysis project type'. This attribute informs the database user about the recovery method behind any given genome and, as detailed in section 4.2.2, is particularly useful considering the expected genome quality.

The prokaryotic MAR databases consist of a contextual part that subsequently corresponds and dictates the sequence data held within. While the contextual data for each entry in MarRef and MarDB depend on curation and available information from auxiliary sources, the sequence data lies promptly available for analysis. Considering the time and knowledge needed to run programs and produce results from our sequence data, we opted to implement and provide convenient analysis output directly accessible on the MMP. For instance, the antiSMASH toolkit [171] became included in the data workflow and predicts secondary metabolites for listed entries. This implementation enables users of the MMP to circumvent a separate antiSMASH analysis, which can take hours. Instead, the output information is included in the MMP framework, enabling the search for specific secondary metabolites by types and clusters. Such secondary metabolites in bacteria provides the cell with advantages towards its environment, and the scientific community finds interest in these molecules for their potential medical use [172]. Thus, providing a marine genomics resource embedded with these natural products may aid in the study and discovery of useful molecules. From a user perspective, the output can be instantly accessed when browsing entries at the MMP site, and also downloaded for further comparison between genomes.

A second tool implemented to make use of genomic data for taxonomic purposes was the GTDB-Tk [173]. As introduced in section 1.1.3 there are two contrasting taxonomic systems for classification of prokaryotes; the literature-dependent taxonomies spearheaded by NCBI [23], Greengenes [29], SILVA [28], RDP [30]; and the standardized bacterial taxonomy constituting the GTDB [35]. Providing both the NCBI and the standardized taxonomy enables users of the MAR databases to choose whichever is suitable or preferred according to their advantages and disadvantages. For example, complications like inconsistent taxon labeling and lack of conformity with phylogenetic studies have been demonstrated for the literature dependent taxonomies [33],

[34], [174]. However, available genomes of decent quality are relatively limited in the standardized taxonomy compared to the abundant 16S rRNA representing organisms in the INSDC. Nevertheless, early reviews of the taxonomic diversity in **Paper 2**, particularly in MarDB, revealed a considerable lack of taxonomic specificity for numerous entries. For instance, more than half of entries composing the *Euryarchaeota* phylum lack more accurate taxon classifications or are designated candidatus or candidate divisions due to uncultivable strains. Utilizing the GTDB-Tk tool on MarRef and MarDB re-evaluated the taxonomic composition according to the standardized taxonomy. Consequently, 91.61% of entries MarDB achieved classification at the genus level. This is a considerable improvement compared to the 52.76% of entries having a classification at this level by the NCBI taxonomy. However, the standardized taxonomy provides only placeholder names when no reference exists [36] and can appear less informative than published species definitions with standing names in nomenclature. The GTDB-Tk approach, nevertheless, avoids human error by automating the classification of genomes. Conclusively, the introduction of tools discussed here required the database attributes to be revamped. For example, the antiSMASH and GTDB-Tk tools outputted values explaining secondary-metabolite gene-clusters and taxonomic lineages useful for the database end-users. These amendments resulted in 124 total attributes describing the entries.

From the period of version 1, we sought to implement databases with specific goals. Just as SalmoBase [18] is a genomic sequence resource for the Atlantic salmon (*Salmo salar*) and Rainbow trout (*Oncorhynchus mykiss*), the complementary microbiota related to these fishes remain uncharted. This became the motivation behind the development of SalDB, a microbiota database for the *Salmonidae* family of fishes. We also introduced the marine-specific database MarFun for holding genomes of eukaryotic fungi. As detailed in **Paper 2**, due to implications of freshwater conditions with anadromous fishes the SalDB became incompatible with the original MAR databases. Yet, SalDB is in terms of attribute design identical to MarRef and MarDB. MarFun, being an eukaryotic database, required descriptive attributes fitting the eukaryotic cell – making it incompatible with the former MAR databases. A significant change to the repertoire of MMP was the temporary decommission of MarCat, the gene catalog based on marine metagenomic samples. Its dependency on the analytic tool META-pipe [175], which still is in development, resulted in its temporary discontinuation.

4.2.2 Genomic quality assessment of the MAR databases

After two years of updates the MarRef and MarDB reached version 5, providing refined attributes, improved functionality, and constituted 970 and 13237 entries, respectively. Rich contextual data describing content, including metrics related to genomic quality, became incentives for designing interactive statistics for the MMP. This would provide users the ability to question the content of the MAR databases like MarRef and MarDB. Three attributes were derived from the CheckM [86] tool for assessing completeness, contamination, and strain heterogeneity in genomes. However, evaluation of their distributions indicated quality issues in the sequence data derived from INSDC sources. Since information on genome condition is not consistently provided by primary databases, these quality issues became the main topic of **Paper 3**. In this publication the genomic characteristics as completeness, contamination, strain heterogeneity, and quality score proved to be highly varied, in particular for entries listed in MarDB. Therefore, we sought to report on the quality of entries in the MarRef and MarDB on the basis of GSC quality classifications [44]. This implied the listed entries and their genome representation, recovery method (WGS, MAG or SAG), and reported contig numbers, rRNA, tRNA, and CheckM-metrics mentioned above.

The quality difference of entries between the two databases was not unexpected, as detailed in **Paper 3**, with MarRef outperforming MarDB. Having only four genomes scoring less than 80% for completeness, the curation effort behind **Paper 1 and 2** satisfyingly ensured the most representative reference genomes for the MarRef database. Notably however, finished genomes did not guarantee a full sett of rRNA (5S, 16S and 23S) and unique tRNAs – questioning the GSC High-quality category [44]. Nevertheless, the entries of MarDB exemplified general trends of completeness and contamination both at a holistic level and dependent on recovery methods (WGS, MAG, or SAG), as illustrated in Figure 7. Unfortunate, yet anticipated, highly spurious genome assemblies was found within the MarDB not fitting any GSC classes. More than 200 genome assemblies qualified worse than the Low-quality draft by exceeding 10% contamination. Consequently, we introduced the class Very Low-quality draft as a means to avoid similar genome representations escaping classification. This came in addition to the class Near-complete for reliable genomes lacking the rRNA/tRNA requirements [92], [93]. It is worth mentioning that Parks et al. introduced a classification system supporting any possible combination of completeness and contamination [86].

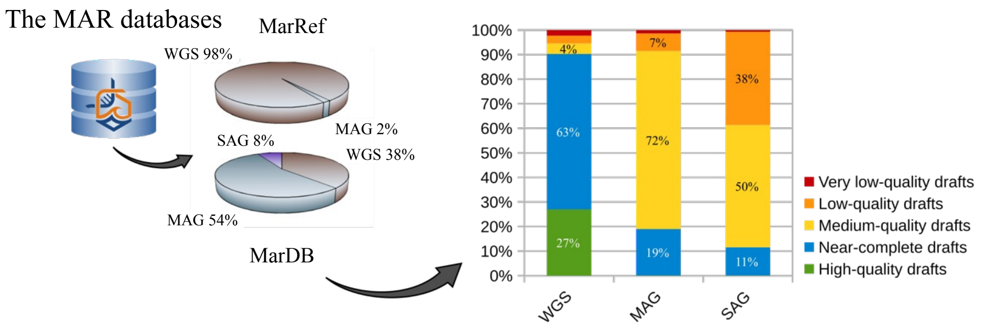


Figure 7. Contents of the MAR databases MarRef and MarDB version 5: Highlighting the proportion of entries within based on recovery method (WGS, MAG, SAG). The stacked bar chart further illustrates the distribution of quality in genome assemblies of MarDB according to the classification scheme presented in Paper 3.

Yet, the GSC classification system is wider recognized and has, for instance, been utilized to set quality cutoffs for 317 MAG genomes derived from 58 deep-sea metagenomes [176]. Even if very low-quality genomes are unfortunate, the classification systems should be inclusive and avoid creating loopholes. During the MAR database development, the attribute ‘genome status’ became implemented to serve as a means of awareness of genome representation and is a product of the six quality classes presented in **Paper 3**. The distribution of genomes into these classes further revealed that WGS type genomes are most prominent in the near-complete class, being deficient in the r/tRNA requirement. MAG and SAG types are most prominent in the medium quality class, with a substantial part of SAG genomes also in the low-quality class. Therefore, the choice of genomes used as reference material must be carefully considered given the recovery methods of either MAG or SAG. RefSeq, on the other hand, resolutely omit genomes recovered as MAG [160].

Since the MAR databases only consider marine derived samples, there are likely genomes originating from other environments that share similar quality issues in public databases. We conclusively argue in **Paper 3** that primary databases of the INSDC should do more to resolve the issue of low-quality drafts in their repositories. Restrictions on publishing low-quality genomes are suggested [44], but not fully enhanced since they are not removed nor flagged. Similar to providing standard genome metrics in the NCBI assembly [177], like size, contigs, etc., quality stats from software as CheckM should be considered to provide details for labeling genome assemblies if the quality is deemed adverse. We further suggest a limitation for reference genomes to comprise a completeness score > 90%, a contamination score < 5%, and a quality score (completeness x 5 contamination) > 65. Avoiding genomes of low contamination by following these minimum standards can furnish the MAR databases with a robust, marine-selective reference collection. Low levels of contamination will additionally benefit k-mer based methods for classification, like Kraken [178]. Such methods have proven problematic in some situations using the NCBI RefSeq database due to plausible contamination, data fluidity, and a taxonomy conflicting with phylogeny [179]. In their publication, Nasko et al. advocated the idea of a hierarchical system derived from stored sequence data, as the GTDB of the standardized taxonomy [35], [36], rather than literature-based taxonomy to correct for unreliable reference data. Hence, in the latest update of the MarRef and MarDB databases, the NCBI taxonomy [23] and the standardized taxonomy [35] contribute with their taxonomic lineages for every entry. In effect, the utilization of the MAR databases can be extended to taxonomic assignment of genomes and amplicon data of **Paper 4**. However, MAG genomes pose a limitation in RNA content due to the metagenomic binning process [180] and subsequently prevent their usefulness related to 16S rRNA classification.

4.3 Classification and phylogeny: potential use cases and impact on prokaryotic database management

4.3.1 Prokaryotic amplicon data from Atlantic salmon: a response to lack of taxonomic coverage

Bioinformatic databases can have a multitude of functionalities. A common case for sequences and related contextual data is its application when combined with external toolkits for analytic tasks needing reference data. The methodical idea behind **Paper 4** investigated the taxonomic usefulness of full-length 16S rRNA from PacBio amplicon sequencing. It was unique in an aqua culture setting targeting samples from the skin and gut mucosa (distal intestine) of Atlantic salmon. We sought to explore the practical application of the full-length compared to the commonly targeted regions for amplicon sequencing. Because primer pairs tend to be inconsistently applied across studies, 16S rRNA gene regions in use varies and result in outcomes not directly comparable. To demonstrate, the level of conservation in the 16S rRNA sequence varies over its gene length, thereby resolving taxonomy unequally and consequently may result in biased classifications [118]. For example, one study using primer systems corresponding with the V3 and V4 regions indicated the intestinal microbiota community of Atlantic salmon could be altered towards lactic acid bacteria through dietary supplements [116]. Applying the V6 region, another study on diets found that predominant intestinal variations were attributed *to Lactobacillales* in Atlantic salmon [181].

On the contrary to these studies, we choose an exhaustive approach with taxonomic classification circumventing the clustering OTUs. Instead, the full-length of the 16S rRNA gene were classified with the LCAClassifier [182] and SilvaMod database. The SilvaMod is based on the SILVA database, an all-encompassing ribosomal RNA database comprising prokaryote and eukaryote sequences [28]. As described in **Paper 4**, our samples from skin and gut constituted two populations of Atlantic salmon given two different diets based on fish and krill meals. Given these populations, we compared the variance between hypervariable regions spanning V3 to V4, V5 to V6, and the full length of the 16S rRNA. Results however, indicated databases as a factor for obtaining high-resolution taxonomic classification.

Improved resolving power for taxa as well as higher accuracy seemed to be the benefits of applying the full-length 16S rRNA gene (**Paper 4**). The significant variations in successfully classified sequences split the datasets based on the sampled areas (gut and skin). This was expected, since the two environments are exposed to widely different external factors. Also expected, the gut microbiota was most affected by changing the diet. The skin could potentially be influenced by the diet in subtle ways as from feces in the rearing water. Nevertheless, considerably fewer genera were identified in the intestinal samples compared to the skin samples. These findings corroborate the reported low microbiome complexity in the distal colon of Atlantic salmon [183]. Few bacteria taxa tend to dominate the Atlantic salmon gut, but some can show sporadic dominance in analysis. Genus *Mycoplasma* is one of such example and has been linked to both farmed and wild fish [109], [184], [185]. Still, the lack of microbiome complexity in **Paper 4** might result from limited accuracy of the 16S rRNA sequence or inadequate taxonomic references in the SilvaMod database. It also points toward uncultivable bacteria non-existent in current-day public databases. The widely studied gut microbiota in humans is an interesting example of how reference data can improve. For instance, analysis of human fecal samples from a study in 1999 could only classify 24% of the total microbial community [186]. Years of further advancement on the subject, including improvements in technology, database development, and the extraction of MAGs have pushed the classification potential to about 70%, as reported by a recent study [93]. In a similar way, the gut of Atlantic salmon may still host species in its microbiota that are not well known or discovered due to cultivation challenges. This can be substantiated in **Paper 4**, where most sample sequences from intestinal-derived microbiota attained less than 50% classification with known prokaryotic families and genera. In skin mucosa, between 50-80% of sequences were classified at the family and genus level. Thus, it can point to a greater proportion of uncultivable bacteria in the intestine compared to the skin surface, and subsequent limitations in taxonomic databases for these particular environment niches. In terms of improved resolving power for taxonomic classification, the full-length 16S rRNA sequences may thus be favorable to environments with decent reference coverage.

This deficiency in references for Atlantic salmon microbiota became one of the inspirations for developing the SalDB and SilvaMar resources for the MMP (detailed in **Paper 2**). This can be regarded as a step forward in improving the knowledge of salmon related microbiota and might be further used for classification purposes. The SalDB (introduced in section 4.2.1) is a specialized database targeting prokaryotic genomes sampled from the microbiota of *Salmonidae* family fishes. The complete microbial diversity on fish surfaces is expected to exceed the 348 entries currently comprising SalDB. And like the MAR databases, SalDB is dependent on published genomes. Yet, the current number of entries can illustrate difficulties in cultivation and insufficient sampling from various niche habitats related to species of *Salmonidae*. Contrarily, the SilvaMar was designed as a resource for ribosomal 16S rRNA classification specific to marine habitats. It is based on a sequence subset from the SILVA database [28] corresponding to entries in the MarRef and MarDB databases. A specialized resource like SilvaMar is less constricted to size. This can enable faster taxonomic classification with tool-engines like BLAST [64] and Qiime2 [187] while still utilizing the entire database, avoiding size reducing steps like clustering similar sequences at given thresholds. For example, the nonredundant reference dataset provided by SILVA cluster sequences at 99% identity to dispose redundant sequences and reduce database size [28]. The threshold usage for 16S rRNA sequences has been debated and may not delineate well at the species level, but conserveness makes 16S rRNA a reliable sequence for family, genus, and to some extent, species and strain classifications [12].

The resulting species level taxonomic classification obtained in **Paper 4** could only be attributed to the skin samples and only for the two species *Flavobacterium frigidarium* and *Chryseobacterium marinum*. While the full-length sequences were found to provide this resolution of classification the variable regions resolved genus level at best. However, significant variations were observed between the datasets that could potentially impact study conclusions. A tendency of likely false positives was found in relation to the taxa *Clostridiales*, *Flavobacterium*, and *Psychrobacter* for the regions. This bias in taxonomic assignments was further emphasized by the taxon *Carnobacterium.* While the V3-V4 region indicated false positives,

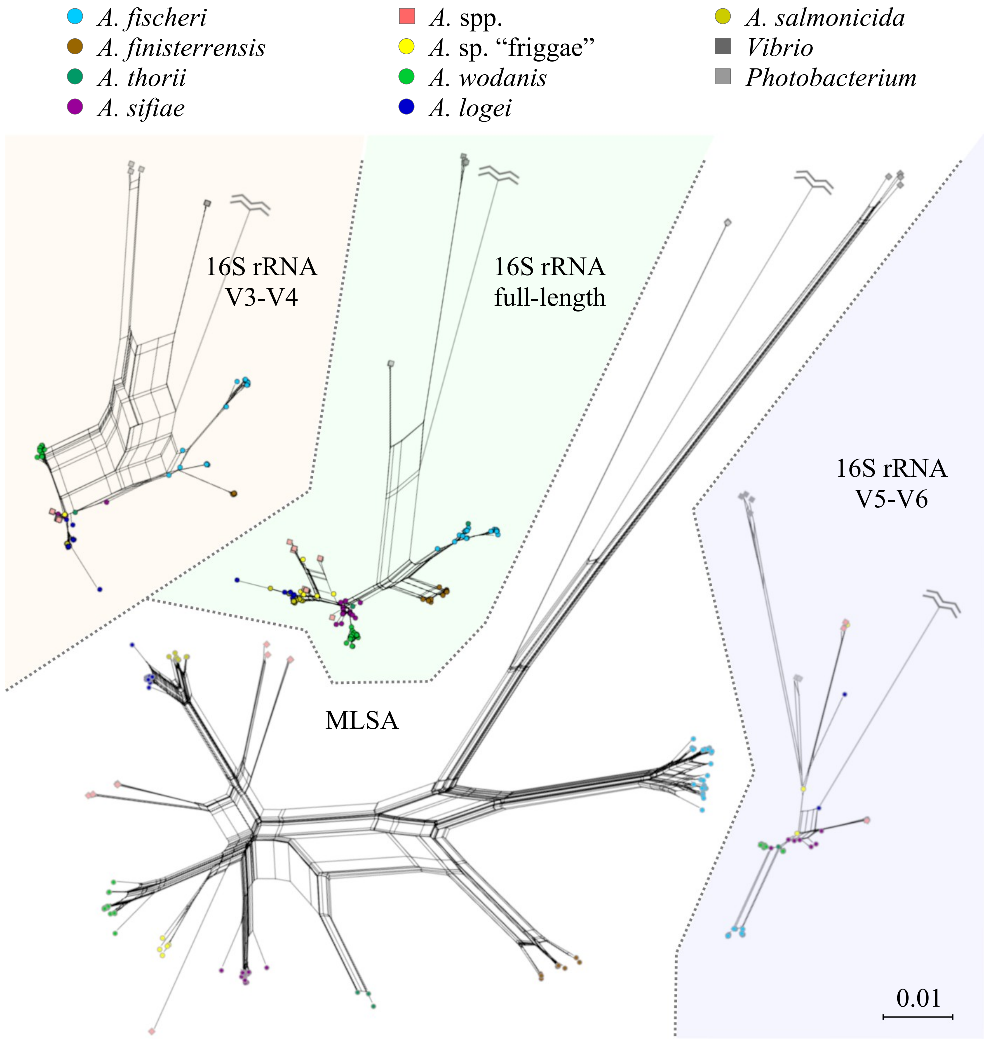


Figure 8. Exemplification of applied 16S rRNA regions. Here illustrated with the dataset used in **Paper 5** and following the variable region selection described in **Paper 4**. It shows the relative resolving power obtained by a six-gene MLSA, full-length 16S rRNA, 16S rRNA regions V3-V4 and V5-V6. The top legend shows coloration of species based on the MLSA analysis. The scale bar representing nucleotide substitutions per site and is representative for all four network diagrams. The outgroup constituting *Photorhabdus luminescens* TT01^{T} has in this illustration been pruned to reduce tree sizes.

Carnobacterium was practically absent from classification results of the V5-V6 region. These discrepancies in classification can thus be attributed to the 16S rRNA regional conserveness and limited resolution. Figure 8 demonstrates the resolution obtained by the 16S rRNA and the regions discussed while using the dataset from **Paper 5**. Conclusively, using the full-length does not guarantee species level classification for Atlantic salmon samples. However, it provides increased accuracy at the genus level compared with the hypervariable regions. To achieve greater depth of resolution, the 16S rRNA gene sequence might be complemented by housekeeping genes or full genome sequencing [32].

4.3.2 Phylogeny of genus *Aliivibrio*: updating taxonomy based on the delineation of highly similar species

As previously discussed for marker gene analysis, the 16S rRNA gene sequence is not sufficient for accurate delineation of prokaryotic species. The lack of resolution by individual genes is well known in the *Vibrionaceae* family of bacteria and concatenated gene sets has been demonstrated to improve the phylogenetic resolution [146], [148], [188]. Genus *Aliivibrio* is one of 13 currently known genera in the NCBI taxonomy related to this family. *Aliivibrio* comprise marine symbionts, like *A. fischeri*, and pathogens as *A. salmonicida* and *A. wodanis* [132], [140], [189]. However, the phylogenetic picture of *Aliivibrio* remains outdated with no further studies since 2010; a time when *A. sifiae* became described as a novel species to the genus [126]. Since then, no extensive examination has considered *Aliivibrio* and the additional strains currently available in INSDC databases. **Paper 5**, therefore, utilized all available sequence data for the phylogenetic inference of *Aliivibrio*.

Additional data from local sequencing initiatives at the UiT – The Arctic University of Norway also contributed to the phylogenetic analysis. Our goal was to update the phylogeny of *Aliivibrio* and aimed at providing an accurate taxonomic description of the genus. This would raise awareness of species classifications in the genus and taxonomic listings in databases. Given that former studies on *Aliivibrio* phylogeny largely have omitted the use of multiple strains this analysis would emphasize and detail the species boundaries based on inter-species distances.

Therefore, the analysis presented in **Paper 5** did not use the 16S rRNA as a singular marker to infer the *Aliivibrio* phylogeny. Instead, the marker became accompanied by a gene set previously shown capable of firmly delineating *Vibrionaceae* [146]. We applied the MLSA method to incorporate genes of conserved proteins to heighten the resolving power of inferred phylogenies [32]. The set of genes included are 16S rRNA, *gapA*, *gyrB*, *pyrH*, *recA*, and *rpoA*. These are illustrated in Figure 9 by order and relative gene lengths that were concatenated. The concatenation protocol was specifically designed with reproducibility in mind, as this particular step tends to be poorly documented by studies. The gene sequence data originated from various protocols, like targeted restriction cutting using different primer pairs or genome sequencing. This lead to some incomplete sequences lacking parts of the gene, often in the flanking regions. By introducing a 5% cutoff for allowed gaps in the flanking regions of an alignment, the process of trimming alignments became semi-automated. The genes included in the alignment design became specifically chosen to include all formerly known and proposed species of *Aliivibrio*.

Extending the design by applying other genes or using fully sequenced genomes would disregard some species and important strains. Consequently, we detailed in **Paper 5** all *Aliivibrio* strains currently known to have the required set of genes in public and local repositories. Thus, with the given dataset, a comprehensive illustration of genus *Aliivibrio* could be generated by the phylogenetic approach.

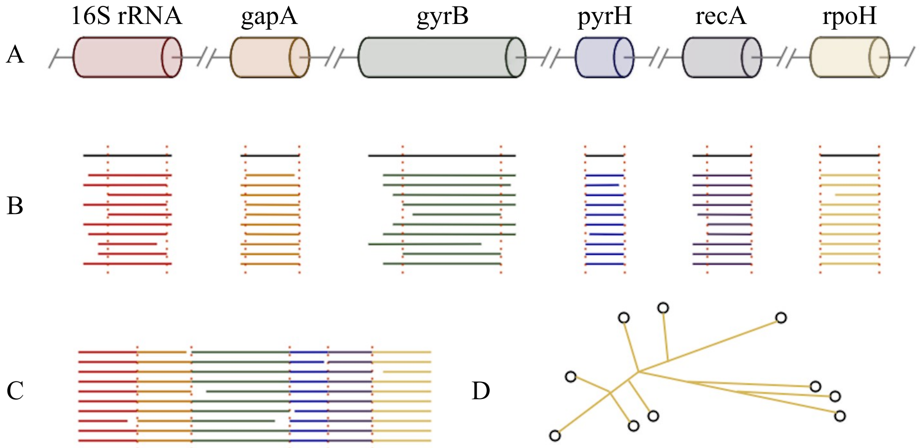


Figure 9. The MLSA construction. Gene loci from in-house genome sequencing and INSDC repositories were collected **A**) and aligned **B**) before being concatenated **C**). The concatenated dataset resulted in the phylogenetic network and trees **D**).

In total, we could assemble a set of 143 bacterial strains for the *Aliivibrio* genus. This set also included comparable samples of neighboring *Photobacteria* for the phylogenetic analysis. A relatively large dataset, as in this study, can cause problems in terms of processing and visualization [190]. First, the testing needed to obtain statistical support values for tree branches can lead to increased time consumption with long MLSA sequences. Secondly, the tree output as a cladogram can get excessively detailed and unfitted for printed publications. To overcome these obstacles we applied a conjunction of neighbor-joining and SplitsTree4 [191] with a splits-network model representing the *Aliivibrio* phylogeny in **Paper 5**. The benefit of this model lies in its ability to closely fit highly similar strain nodes and provide a clear-cut visualization of the resulting species arrangement. Caveats however, can be attributed to the projection of support values for edges connecting the various species nodes. Consequently, we supplemented the lacking support values by providing a regular neighbor-joining tree as a cladogram in the appendix section to provide valuable data for detailed phylogenetic robustness.

In general, our results substantiated some of the latest phylogenetic studies targeting genus *Aliivibrio* [125], [126], but also found support for taxa in the genome-based taxonomy [36]. By showing inclusiveness for available strains we delineated another five potentially novel species as detailed in **Paper 5**. Applying a multitude of strains rather than type strains aided in the species delineation. In particular, strains formerly assigned *A. wodanis* became allocated a new cluster we named *Aliivibrio* sp. "friggae". Furthermore, several strains were corrected from having dubious taxonomic classifications. Some species like *A. sifiae* and *A. finisterrensis* were attributed far more strains than previous studies have shown. Regardless of the current data size, the inferred phylogeny still leaves considerable gaps between species clusters. Hence, we consider these spacing as indicators for incomplete species abundance in genus *Aliivibrio*. Future sampling and isolation of genomes, cultivated or SAG/MAG derived, may lead to further amendments in the *Aliivibrio* phylogeny. This again points to a lack of genomic reference coverage in diverse marine environments.

Based on the phylogeny we further detailed the intra-species sequence variance and the inter-species distance separating species. The utilization of multiple strains were fundamental in these estimates as they gave an improved understanding of the species variance – within and between species and genera. *A. salmonicida* and *A. logei* had the shortest intra-species distance. These circumscription radii became assistive in describing species, but should be regarded as contemporary as future strains assigned any of the clusters might impact its size. A similar concept circumscribing species is based on genome-wide ANI values between sequenced genomes and is used by the standardized taxonomy [36]. Still, the coverage of *Aliivibrio* genomes is lacking in the standardized taxonomy, but may in future studies resolve *Aliivibrio* at greater details than what could be presented in **Paper 5**. This will likely require additional genome sequenced bacteria in genus *Aliivibrio*.

Because species differentiation within *Vibrionaceae* is impractical with 16S rRNA [147], [192], and the sequencing of multiple genes for MLSA studies may prove demanding, we performed comparisons between dual gene combinations to assess topological congruence with the inferred MLSA phylogeny. By providing recommendations for a minimal marker gene combination we could advance the accuracy of classifying *Aliivibrio*-like bacteria with less effort. In **Paper 5**, the most significant topological similarity was found for the concatenated *gyrB-recA* gene alignment, which shared 80.46% with the MLSA tree. Applying these genes individually did not favor species delineation comparable with the MLSA. This corroborated earlier statements of unreliable typologies from using *gyrB* and *gapA* genes in conjunction with the *Vibrio* genus [147]. Therefore, the close similarity between the *gyrB-recA* configuration and the MLSA may arise from a combinatory effect when concatenating the sequence data. Effectively producing a merged tree topology. This is illustrated by the strain *Aliivibrio* sp. appey-12 being miss-placed in the *gapA* tree, but largely influenced and repositioned in the *gyrB-recA* tree. Given that the *gyrB* and *gapA* genes have similar resolving power in the wider bacteria, future studies may benefit from developing methods for targeting them in amplicon sequencing as in **Paper 4**. Ideally, this gene combination might lead to improved species delineation, particularly for *Aliivibrio*, *Vibrio*, and *Photobacteria*. In the event that the conserved 16S rRNA were to be replaced by an alternative gene candidate, it would require reference databases in order to effectively classify sample data. Here, the construction of a *gyrB-recA* database and protocols will necessitate correct classification in a similar manner as the SILVA [28] and SILVA MAR (**Paper 2**). With current technology, the use of single genes is simpler in terms of making restriction cuts for amplicon sequencing. In **Paper 5** only six genes were considered and did not include the *coaE* gene. The *coaE* is a favored candidate for 16S replacement due to its near correlation with genome wide AAI [192]. With this in mind further potential exist for single genes to delineate *Aliivibrio* species. Access to sequenced genomes and advantageous gene sequences, thus, is a necessity.

Only a marginal proportion of the available and included strains had status as fully sequenced genomes. This hampered the possibility to evaluate alternative gene sets and perform genome-wide data comparisons like AAI without losing important strains. Fully sequenced genomes of reasonable quality are classifiable by the standardized taxonomy using GTDB-Tk [173]. Some of the strains proposed as novel species in **Paper 5** were genome sequenced. Those available to the public and passing the quality requirements are listed in the GTDB [35]. Although our study applied a set of six gene MLSA, the results corroborated the genome level taxonomy listed in the GTDB database. For instance, the genome sequenced strains EL58, 1S128, 1S165, and 1S175 forming the novel species *Aliivibrio* sp. “friggae” and “vili” are unassigned known species in GTDB.

4.4 Future perspectives

The web service upheld by the MMP, including the searchable MAR databases and included sequence resources, has remained operational since the start of 2018 and is intended to continue the advancement in forthcoming years. As mentioned in **Paper 2**, the MAR databases have received updated entry lists following a biannual schedule. Additional and specialized databases targeting genomes of marine fungi (MarFun) and prokaryotic microbiota of Atlantic salmon (SalDB) have later been introduced with improved functionalities. The expansion of data and implementation of state-of-the-art standards and principles are continual processes in the database development. Outlooks for the MMP are to improve and adapt to the needs of the marine genomic research community and deliver updated services. Bringing together high-quality standards for marine genomic sequence data (**Paper 3**), and relying on accurate taxonomy (**Paper 5**) is essential to provide the most up-to-date reference data for classification purposes (**Paper 4**). However, accomplishing this faces several challenges.

Sequence quality and database size (number of listed entries) are current concerns as pointed out in **Paper 2 and 3**, particularly for MarDB. For instance, the large number of entries induces difficulties when processing contextual data as well as delays when querying the MMP portal. This, consequently, impact both curators and end-users. In respect to the corresponding sequence data, the process of downloading, implementation, tool analysis, and preparation of BLAST databases demand a considerable effort to materialize. However, most tasks, including contextual preparation, are manual or semi-manual in each update cycle. With the deluge of genomes entering primary databases, the implementation of data into the MAR databases should be adjusted for scalability to handle several thousand genomes. Tasks not explicitly requiring human intervention are beneficial if automated to reduce counterproductive time consumption. Actions have been taken to reform the systematics surrounding the update flow of the MAR databases and put to use in a pilot database of a planned marine viral database.

During the first months of the SARS-COV-2 pandemic outbreak in spring 2020, we designed the SARS-COV-2 database (covid19.sfb.uit.no) and started to collect publicly available contextual and sequence data on the virus – similarly as for the MAR databases. The frequent sequence update and simple attribute design provided an opportunity to use the SARS-COV-2 database as a pilot project to optimize the data workflows. First, it gave the means to automate contextual data import, contextual data checks, and performing the deployment of new database versions in the SARS-COV-2 data portal through GitLab (gitlab.com/uit-sfb/sarscovid19db). Secondly, it provided a stronger foundation for documenting contextual origin through the ECO ontology [63], which has only partial integration in the MAR databases (**Paper 2**). Improving data evidence gives a database credibility. This can be reflected in the database ability to document data provenance and the level of effort put into the biocuration of its data [58]. This can be exemplified by the Swiss-Prot and

TrEMBL databases that provide evidence codes as a descriptive connection between presented data and its source of origin [193]. Likewise, the SARS-COV-2 database has utilized ECO evidence connecting contextual data from sources like NCBI virus [194], NGDC [7], as well as specific virus documentation and analytic tools as Pangolin (github.com/cov-lineages/pangolin). The SARS-COV-2 database development has become a framework for contextual data storage, codification of contextual sources, corresponding sequence storage, and its workflow with the portal endpoint. The database is now linked from the COVID-19 Data Portal Norway (covid19dataportal.no) as an infrastructure initiative in Europe [8]. One strategic goal is to port the SARS-COV-2 framework to fit the MAR databases and the planned marine virus database. Once in place, attention can be focused elsewhere, like further automation, devotion to the curation effort, and the implementation of text mining for marine contextual data.

It is worth noting that due to the emergence of the pandemic and the development of the SARS-COV-2 database, the planned MMP web interface was not finalized on time for the publication deadline of **Paper 2**. Nonetheless, the upcoming pilot implementation of the MMP resources forces **Paper 2** to be redrafted based on the very latest amendments regarding the operation and updates of the databases. This includes the latest provisional MMP portal version at https://mmp2.sfb.uit.no.

In spite of the full-length 16S rRNA having greater comparability between amplicon studies than regions, the biological coverage represented by reference data can still be improved. To alleviate lacking reference sets for classification purposes, as discussed in **Paper 4**, future studies may contribute with rigorous sampling, cultivation, and genome sequencing of niche communities like the mucosal surfaces of Atlantic salmon. Publication of such material can broaden the genomic data from these microbial niches. The low cultivability in such samples can be supplemented by non-cultivable species through single-cell approaches or metagenomic sampling with the construction of MAGs. In fact, efforts in later years have facilitated the linking of 16S rRNA with corresponding MAGs to a greater extent [180], [195], [196] and, thus, permits a broadening of data in amplicon reference sets. Therefore, with the MarRef, MarDB, and SalDB we aim to continue building the databases both in terms of knowledge and reference set using genome recovery methods as MAG and SAG. Version 4 of the MAR databases became an example of how reference sets based on sequence data could be built for classification purposes (**Paper2**). Given wider reference sets and firm taxonomic annotations, the use of full-length 16S rRNA will be open for reevaluation. Analysis however, can be compared with reference data from RefSeq [160], SILVA [28], Greengenes [29], and RDP [30] and could be prepared and engaged through programs like Qiime2 [187] and LCAClassifier [182].

Ideal for classification purposes is a coherent taxonomy supported by robust phylogenetic dimensions. The ongoing process of discovering novel prokaryotic species and fitting them in phylogenetic relationships is a dynamic aspect of advancing taxonomies. Therefore, the current phylogeny of *Aliivibrio* may be thought of as a contemporary picture and likely to expand and change with time. Still, for the MAR database cases, there is a continual strive towards utilizing the latest state-of-the-art taxonomic systems. In **Paper 5**, twelve species of various environmental or host specializations are presented. This presentation does not elucidate the genomic inner workings contributing to the observed diversity in the genus. For this to be clarified, comparative studies may potentially bring about a deeper understanding of genotypic differences. *A. salmonicida* and *A. logei* are examples of exceptionally close species. Pan-genome analysis in this case, and in the case of *Aliivibrio* sp. «friggae» and neighboring *A. wodanis,* can help understand variations in gene composition, gene synteny and biological capabilities like pathogenicity and chemical signaling through quorum sensing.

Studies on microbial communities, like those encountered in the aquaculture, are benefited by robust taxonomies and well-described phylogenetic differentiation. Particularly concerning the differentiation of pathogen and commensal taxa – both constituting the *Aliivibrio*. The knowledge about commensal bacteria and their functions, in Atlantic salmon for instance, are lacking. Additional sequencing, amplicon, and genomic studies will be needed in order to understanding their contribution in the microbiota complex. The familiarity of microbiota in aquaculture systems remains to be elucidated, and the various compositions presented lack the desired standardization for meta-analysis across projects.

5 References

[1] K. R. Kumar, M. J. Cowley, and R. L. Davis, “Next-Generation Sequencing and Emerging Technologies,” *Semin. Thromb. Hemost.*, vol. 45, no. 7, pp. 661–673, May 2019, doi: 10.1055/s-0039-1688446.

[2] H. Müller and F. Naumann, “Data Quality in Genome Databases.,” pp. 269–284, 2003, doi: http://dx.doi.org/10.18452/9205.

[3] M. Arita *et al.*, “The International Nucleotide Sequence Database Collaboration,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D121–D124, Jan. 2021, doi: 10.1093/NAR/GKAA967.

[4] P. W. Harrison *et al.*, “The European Nucleotide Archive in 2020.,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D82–D85, Jan. 2021, doi: 10.1093/nar/gkaa1028.

[5] E. W. Sayers *et al.*, “Database resources of the National Center for Biotechnology Information,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D10–D17, Jan. 2021, doi: 10.1093/nar/gkaa892.

[6] A. Fukuda, Y. Kodama, J. Mashima, T. Fujisawa, and O. Ogasawara, “DDBJ update: Streamlining submission and access of human data,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D71–D75, Jan. 2021, doi: 10.1093/nar/gkaa982.

[7] Y. Xue *et al.*, “Database resources of the national genomics data center, china national center for bioinformation in 2021,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D18–D28, Jan. 2021, doi: 10.1093/nar/gkaa1022.

[8] G. Cantelli *et al.*, “The European Bioinformatics Institute: Empowering cooperation in response to a global health crisis,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D29–D37, Jan. 2021, doi: 10.1093/nar/gkaa1077.

[9] B. M. Forde and P. W. O’Toole, “Next-generation sequencing technologies and their impact on microbial genomics,” *Brief. Funct. Genomics*, vol. 12, no. 5, pp. 440–453, Sep. 2013, doi: 10.1093/bfgp/els062.

[10] J. D. Ferreira, B. Inácio, R. M. Salek, and F. M. Couto, “Assessing Public Metabolomics Metadata, Towards Improving Quality,” *J. Integr. Bioinform.*, vol. 14, no. 4, Dec. 2017, doi: 10.1515/jib-2017-0054.

[11] J. W. Blunt, A. R. Carroll, B. R. Copp, R. A. Davis, R. A. Keyzers, and M. R. Prinsep, “Marine natural products,” *Natural Product Reports*, vol. 35, no. 1. The Royal Society of Chemistry, pp. 8–53, Jan. 25, 2018, doi: 10.1039/c7np00052a.

[12] J. S. Johnson *et al.*, “Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis,” *Nat. Commun.*, vol. 10, no. 1, pp. 1–11, Nov. 2019, doi: 10.1038/s41467-019-13036-1.

[13] A. L. Mitchell *et al.*, “EBI Metagenomics in 2017: Enriching the analysis of microbial communities, from sequence reads to assemblies,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D726–D735, Jan. 2018, doi: 10.1093/nar/gkx967.

[14] A. L. Mitchell *et al.*, “MGnify: The microbiome analysis resource in 2020,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D570–D578, Jan. 2020, doi: 10.1093/nar/gkz1035.

[15] A. Zanzi and J. T. Martinsohn, “FishTrace: a genetic catalogue of European fishes,” *Database (Oxford).*, vol. 2017, p. 75, Jan. 2017, doi: 10.1093/database/bax075.

[16] Y. J. Liew, M. Aranda, and C. R. Voolstra, "Reefgenomics.Org - A repository for marine genomics data," *Database*, vol. 2016, Jan. 2016, doi: 10.1093/database/baw152.

[17] S. Sagar, M. Kaur, A. Radovanovic, and V. B. Bajic, "Dragon exploration system on marine sponge compounds interactions," *J. Cheminform.*, vol. 5, no. 2, Feb. 2013, doi: 10.1186/1758-2946-5-11.

[18] J. K. A. Samy *et al.*, "SalmoBase: An integrated molecular data resource for *Salmonid* species," *BMC Genomics*, vol. 18, no. 1, Jun. 2017, doi: 10.1186/s12864-017-3877-1.

[19] C. J. Hyde, Q. P. Fitzgibbon, A. Elizur, G. G. Smith, and T. Ventura, "CrustyBase: An interactive online database for crustacean transcriptomes," *BMC Genomics*, vol. 21, no. 1, Sep. 2020, doi: 10.1186/s12864-020-07063-2.

[20] X. Li *et al.*, "ConoMode, a database for conopeptide binding modes," *Database*, vol. 2020, p. 58, 2020, doi: 10.1093/database/baaa058.

[21] M. Tangherlini *et al.*, "GLOSSary: The GLobal Ocean 16S subunit web accessible resource," *BMC Bioinformatics*, vol. 19, no. S15, p. 443, Nov. 2018, doi: 10.1186/s12859-018-2423-8.

[22] Q. Carradec *et al.*, "A global ocean atlas of eukaryotic genes," *Nat. Commun.*, vol. 9, no. 1, Dec. 2018, doi: 10.1038/s41467-017-02342-1.

[23] C. L. Schoch *et al.*, "NCBI Taxonomy: A comprehensive update on curation, resources and tools," *Database*, vol. 2020. Oxford University Press, 2020, doi: 10.1093/database/baaa062.

[24] A. C. Parte, "LPSN - List of prokaryotic names with standing in nomenclature (Bacterio.net), 20 years on," *International Journal of Systematic and Evolutionary Microbiology*, vol. 68, no. 6. Microbiology Society, pp. 1825–1829, Jun. 01, 2018, doi: 10.1099/ijsem.0.002786.

[25] M. J. Costello *et al.*, "Global Coordination and Standardisation in Marine Biodiversity through the World Register of Marine Species (WoRMS) and Related Databases," *PLoS One*, vol. 8, no. 1, p. e51629, Jan. 2013, doi: 10.1371/journal.pone.0051629.

[26] S. Federhen, "The NCBI Taxonomy database," *Nucleic Acids Res.*, vol. 40, no. D1, p. D136, Jan. 2012, doi: 10.1093/nar/gkr1178.

[27] B. Yang, Y. Wang, and P.-Y. Y. Qian, "Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis," *BMC Bioinformatics*, vol. 17, no. 1, p. 135, Dec. 2016, doi: 10.1186/s12859-016-0992-y.

[28] C. Quast *et al.*, "The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools," *Nucleic Acids Res.*, vol. 41, no. D1, p. D590, Jan. 2013, doi: 10.1093/nar/gks1219.

[29] T. Z. DeSantis *et al.*, "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB," *Appl. Environ. Microbiol.*, vol. 72, no. 7, pp. 5069–5072, Jul. 2006, doi: 10.1128/AEM.03006-05.

[30] J. R. Cole *et al.*, "Ribosomal Database Project: Data and tools for high throughput rRNA analysis," *Nucleic Acids Res.*, vol. 42, no. D1, Jan. 2014, doi: 10.1093/nar/gkt1244.

[31] F. O. Glöckner *et al.*, "25 years of serving the community with ribosomal RNA gene reference databases and tools," *Journal of Biotechnology*, vol. 261. Elsevier, pp. 169–176, Nov. 10, 2017, doi: 10.1016/j.jbiotec.2017.06.1198.

[32] S. P. Glaeser and P. Kämpfer, “Multilocus sequence analysis (MLSA) in prokaryotic taxonomy,” *Systematic and Applied Microbiology*, vol. 38, no. 4. Elsevier GmbH, pp. 237–245, Jun. 01, 2015, doi: 10.1016/j.syapm.2015.03.007.

[33] R. Edgar, “Taxonomy annotation and guide tree errors in 16S rRNA databases,” *PeerJ*, vol. 2018, no. 6, 2018, doi: 10.7717/peerj.5030.

[34] K. A. Lydon and E. K. Lipp, “Taxonomic annotation errors incorrectly assign the family *Pseudoalteromonadaceae* to the order Vibrionales in Greengenes: Implications for microbial community assessments,” *PeerJ*, vol. 2018, no. 7, 2018, doi: 10.7717/peerj.5248.

[35] D. H. Parks *et al.*, “A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life,” *Nat. Biotechnol.*, vol. 36, no. 10, pp. 996–1004, Nov. 2018, doi: 10.1038/nbt.4229.

[36] D. H. Parks, M. Chuvochina, P. A. Chaumeil, C. Rinke, A. J. Mussig, and P. Hugenholtz, “A complete domain-to-species taxonomy for Bacteria and Archaea,” *Nat. Biotechnol.*, vol. 38, no. 9, pp. 1079–1086, Apr. 2020, doi: 10.1038/s41587-020-0501-8.

[37] F. Shaw *et al.*, “COPO: a metadata platform for brokering FAIR data in the life sciences,” *F1000Research*, vol. 9, p. 495, Jun. 2020, doi: 10.12688/f1000research.23889.1.

[38] S. J. S. Khalsa, “Data and Metadata Brokering – Theory and Practice from the BCube Projec,” *Data Sci. J.*, vol. 16, pp. 1–8, Jan. 2017, doi: https://doi.org/10.5334/dsj-2017-001.

[39] M. D. Wilkinson *et al.*, “Comment: The FAIR Guiding Principles for scientific data management and stewardship,” *Sci. Data*, vol. 3, no. 1, pp. 1–9, Mar. 2016, doi: 10.1038/sdata.2016.18.

[40] M. Boeckhout, G. A. Zielhuis, and A. L. Bredenoord, “The FAIR guiding principles for data stewardship: Fair enough?,” *European Journal of Human Genetics*, vol. 26, no. 7. Nature Publishing Group, pp. 931–936, Jul. 01, 2018, doi: 10.1038/s41431-018-0160-0.

[41] S. Brunak, “Nucleotide Sequence Database Policies,” *Science (80-.).*, vol. 298, no. 5597, pp. 1333b – 1333, Nov. 2002, doi: 10.1126/science.298.5597.1333b.

[42] T. Barrett *et al.*, “BioProject and BioSample databases at NCBI: Facilitating capture and organization of metadata,” *Nucleic Acids Res.*, vol. 40, no. D1, p. D57, Jan. 2012, doi: 10.1093/nar/gkr1163.

[43] P. Yilmaz *et al.*, “Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications,” *Nature Biotechnology*, vol. 29, no. 5. Nature Publishing Group, pp. 415–420, May 06, 2011, doi: 10.1038/nbt.1823.

[44] R. M. Bowers *et al.*, Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea, vol. 35, no. 8. Nature Publishing Group, 2017.

[45] S. Roux *et al.*, “Minimum Information about an Uncultivated Virus Genome (MIUViG).,” *Nat. Biotechnol.*, vol. 37, no. 1, pp. 29–37, Jan. 2019, doi: 10.1038/nbt.4306.

[46] R. L. Walls *et al.*, “Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies,” *PLoS One*, vol. 9, no. 3, Mar. 2014, doi: 10.1371/journal.pone.0089606.

[47] G. Mayer *et al.*, “Controlled vocabularies and ontologies in proteomics: Overview, principles and practice,” *Biochim. Biophys. Acta - Proteins Proteomics*, vol. 1844, no. 1 PART A, pp. 98–107, 2014, doi: 10.1016/j.bbapap.2013.02.017.

[48] B. Smith *et al.*, “The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration,” *Nature Biotechnology*, vol. 25, no. 11. NIH Public Access, pp. 1251–1255, Nov. 2007, doi: 10.1038/nbt1346.

[49] M. Ashburner *et al.*, “Gene ontology: Tool for the unification of biology,” *Nature Genetics*, vol. 25, no. 1. NIH Public Access, pp. 25–29, May 2000, doi: 10.1038/75556.

[50] A. Tomczak *et al.*, “Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations,” *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, Mar. 2018, doi: 10.1038/s41598-018-23395-2.

[51] J. Malone *et al.*, “Modeling sample variables with an Experimental Factor Ontology,” *Bioinformatics*, vol. 26, no. 8, pp. 1112–1118, Mar. 2010, doi: 10.1093/bioinformatics/btq099.

[52] R. R. Brinkman *et al.*, “Modeling biomedical experimental processes with OBI,” *J. Biomed. Semantics*, vol. 1, no. 1, Jun. 2010, doi: 10.1186/2041-1480-1-S1-S7.

[53] P. L. Buttigieg *et al.*, “The environment ontology: Contextualising biological and biomedical entities,” *J. Biomed. Semantics*, vol. 4, no. 1, pp. 1–9, Dec. 2013, doi: 10.1186/2041-1480-4-43.

[54] G. V Gkoutos, P. N. Schofield, and R. Hoehndorf, “The anatomy of phenotype ontologies: principles, properties and applications,” *Brief. Bioinform.*, vol. 19, no. 5, pp. 1008–1021, Sep. 2018, doi: 10.1093/BIB/BBX035.

[55] K. Degtyarenko *et al.*, “ChEBI: A database and ontology for chemical entities of biological interest,” *Nucleic Acids Res.*, vol. 36, no. SUPPL. 1, Jan. 2008, doi: 10.1093/nar/gkm791.

[56] M. Martínez-Romero *et al.*, “Fast and Accurate Metadata Authoring Using Ontology-Based Recommendations.,” *AMIA Annu. Symp. Proc.*, vol. 2017, pp. 1272–1281, 2017, Accessed: Aug. 18, 2021. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29854196.

[57] M. Courtot *et al.*, “BioSamples database: an updated sample metadata hub,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1172–D1178, Jan. 2019, doi: 10.1093/nar/gky1061.

[58] A. Holinski, M. L. Burke, S. L. Morgan, P. McQuilton, and P. M. Palagi, “Biocuration - mapping resources and needs,” *F1000Research*, vol. 9, p. 1094, Dec. 2020, doi: 10.12688/f1000research.25413.2.

[59] M. B. Duhaime, R. Kottmann, D. Field, and F. O. Glöckner, “Enriching public descriptions of marine phages using the genomic standards consortium MIGS standard,” *Stand. Genomic Sci.*, vol. 4, no. 2, pp. 271–285, 2011, doi: 10.4056/sigs.621069.

[60] A. Venkatesan, N. Karamanis, M. Ide-Smith, J. Hickford, and J. McEntyre, “Understanding life sciences data curation practices via user research,” *F1000Research*, vol. 8, p. 1622, Sep. 2019, doi: 10.12688/f1000research.19427.1.

[61] E. C. Dimmer *et al.*, “The UniProt-GO Annotation database in 2011,” *Nucleic Acids Res.*, vol. 40, no. D1, pp. D565–D570, Jan. 2012, doi: 10.1093/nar/gkr1048.

[62] M. C. Chibucos *et al.*, “Standardized description of scientific evidence using the Evidence Ontology (ECO),” *Database*, vol. 2014, no. 0, Jul. 2014, doi: 10.1093/database/bau075.

[63] M. Giglio *et al.*, “Eco, the evidence & conclusion ontology: Community standard for evidence information,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1186–D1194, Jan. 2019, doi: 10.1093/nar/gky1036.

[64] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, 1990, doi: 10.1016/S0022-2836(05)80360-2.

[65] M. C. Chibucos, D. A. Siegele, J. C. Hu, and M. Giglio, “The evidence and conclusion ontology (ECO): Supporting GO annotations,” in *Methods in Molecular Biology*, vol. 1446, Humana Press Inc., 2017, pp. 245–259.

[66] F. P. Breitwieser, J. Lu, and S. L. Salzberg, “A review of methods and databases for metagenomic classification and assembly,” *Brief. Bioinform.*, vol. 20, no. 4, pp. 1125–1139, Mar. 2018, doi: 10.1093/bib/bbx120.

[67] J. Bengtsson-Palme *et al.*, “Strategies to improve usability and preserve accuracy in biological sequence databases,” *Proteomics*, vol. 16, no. 18, pp. 2454–2460, Sep. 2016, doi: 10.1002/pmic.201600034.

[68] M. R. Bouadjenek, K. Verspoor, and J. Zobel, “Literature consistency of bioinformatics sequence databases is effective for assessing record quality,” *Database*, vol. 2017, no. 1, p. 21, Jan. 2017, doi: 10.1093/database/bax021.

[69] M. Eisenstein, “Closing in on a complete human genome,” *Nature*, vol. 590, no. 7847, pp. 679–681, Feb. 2021, doi: 10.1038/d41586-021-00462-9.

[70] Q. Chen, J. Zobel, and K. Verspoor, “Duplicates, redundancies and inconsistencies in the primary nucleotide databases: A descriptive study,” *Database*, vol. 2017, no. 1, p. baw163, Jan. 2017, doi: 10.1093/database/baw163.

[71] W. Li, L. Jaroszewski, and A. Godzik, “Clustering of highly homologous sequences to reduce the size of large protein databases,” *Bioinformatics*, vol. 17, no. 3. Oxford Academic, pp. 282–283, Mar. 01, 2001, doi: 10.1093/bioinformatics/17.3.282.

[72] M. Albertsen, P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson, and P. H. Nielsen, “Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes,” *Nat. Biotechnol.*, vol. 31, no. 6, pp. 533–538, May 2013, doi: 10.1038/nbt.2579.

[73] S. J. Salter *et al.*, “Reagent and laboratory contamination can critically impact sequence-based microbiome analyses,” *BMC Biol.*, vol. 12, no. 1, pp. 1–12, Nov. 2014, doi: 10.1186/s12915-014-0087-z.

[74] J. Caswell *et al.*, “Defending our public biological databases as a global critical infrastructure,” *Front. Bioeng. Biotechnol.*, vol. 7, no. APR, p. 58, Apr. 2019, doi: 10.3389/fbioe.2019.00058.

[75] E. Kopylova, L. Noé, and H. Touzet, “SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data,” *Bioinformatics*, vol. 28, no. 24, pp. 3211–3217, Dec. 2012, doi: 10.1093/bioinformatics/bts611.

[76] B. Langmead and S. L. Salzberg, “Fast gapped-read alignment with Bowtie 2,” vol. 9, no. 4, pp. 357–359, 2012, doi: 10.1038/nmeth.1923.

[77] C. A. Marotz, J. G. Sanders, C. Zuniga, L. S. Zaramela, R. Knight, and K. Zengler, “Improving saliva shotgun metagenomics by chemical host DNA depletion,” *Microbiome*, vol. 6, no. 1, 2018, doi: 10.1186/s40168-018-0426-3.

[78] F. P. Breitwieser, M. Pertea, A. V. Zimin, and S. L. Salzberg, “Human contamination in bacterial genomes has created thousands of spurious proteins,” *Genome Res.*, vol. 29, no. 6, pp. 954–960, 2019, doi: 10.1101/gr.245373.118.

[79] C. M. Fraser, J. A. Eisen, K. E. Nelson, I. T. Paulsen, and S. L. Salzberg, “The value of complete microbial genome sequencing (you get what you pay for),”

Journal of Bacteriology, vol. 184, no. 23. pp. 6403–6405, Dec. 2002, doi: 10.1128/JB.184.23.6403-6405.2002.

[80] E. Mardis, J. McPherson, R. Martienssen, R. K. Wilson, and W. R. McCombie, “What is Finished, and Why Does it Matter,” *Genome Res.*, vol. 12, no. 5, pp. 669–671, May 2002, doi: 10.1101/GR.032102.

[81] F. S. Kremer, A. J. A. McBride, and L. da S. Pinto, “Approaches for in silico finishing of microbial genomesequences,” *Genet. Mol. Biol.*, vol. 40, no. 3, p. 553, 2017, doi: 10.1590/1678-4685-GMB-2016-0230.

[82] M. Land *et al.*, “Insights from 20 years of bacterial genome sequencing,” *Functional and Integrative Genomics*, vol. 15, no. 2. Springer, pp. 141–161, Mar. 01, 2015, doi: 10.1007/s10142-015-0433-4.

[83] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, “IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth,” *Bioinformatics*, vol. 28, no. 11, pp. 1420–1428, Jun. 2012, doi: 10.1093/bioinformatics/bts174.

[84] M. J. Hubisz, M. F. Lin, M. Kellis, and A. Siepel, “Error and error mitigation in low-coverage genome assemblies,” *PLoS One*, vol. 6, no. 2, 2011, doi: 10.1371/journal.pone.0017034.

[85] A. Thrash, F. Hoffmann, and A. Perkins, “Toward a more holistic method of genome assembly assessment,” *BMC Bioinformatics*, vol. 21, no. Suppl 4. BioMed Central, Jul. 06, 2020, doi: 10.1186/s12859-020-3382-4.

[86] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, “CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes,” *Genome Res.*, vol. 25, no. 7, pp. 1043–1055, Jul. 2015, doi: 10.1101/gr.186072.114.

[87] B. Segerman, “The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories.,” *Front. Cell. Infect. Microbiol.*, vol. 2, p. 116, 2012, doi: 10.3389/fcimb.2012.00116.

[88] C. J. Creevey, T. Doerks, D. A. Fitzpatrick, J. Raes, and P. Bork, “Universally distributed single-copy genes indicate a constant rate of horizontal transfer,” *PLoS One*, vol. 6, no. 8, p. e22099, 2011, doi: 10.1371/journal.pone.0022099.

[89] J. Alneberg *et al.*, “Genomes from uncultivated prokaryotes: A comparison of metagenome-assembled and single-amplified genomes 06 Biological Sciences 0604 Genetics,” *Microbiome*, vol. 6, no. 1, pp. 1–14, Sep. 2018, doi: 10.1186/s40168-018-0550-0.

[90] A. M. Eren *et al.*, “Anvi’o: An advanced analysis and visualization platformfor ’omics data,” *PeerJ*, vol. 2015, no. 10, 2015, doi: 10.7717/peerj.1319.

[91] K. Tennessen *et al.*, “ProDeGe: A computational protocol for fully automated decontamination of genomes,” *ISME J.*, vol. 10, no. 1, pp. 269–272, Jan. 2016, doi: 10.1038/ismej.2015.100.

[92] D. H. Parks *et al.*, “Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life,” *Nat. Microbiol.*, vol. 2, no. 11, pp. 1533–1542, Nov. 2017, doi: 10.1038/s41564-017-0012-7.

[93] A. Almeida *et al.*, “A new genomic blueprint of the human gut microbiota.,” *Nature*, vol. 568, no. 7753, pp. 499–504, Apr. 2019, doi: 10.1038/s41586-019-0965-1.

[94] P. Kämpfer and S. P. Glaeser, “Prokaryotic taxonomy in the sequencing era - the polyphasic approach revisited,” *Environmental Microbiology*, vol. 14, no. 2. John

Wiley & Sons, Ltd, pp. 291–317, Feb. 01, 2012, doi: 10.1111/j.1462-2920.2011.02615.x.

[95] G. M. Garrity, “A new genomics-driven taxonomy of bacteria and archaea: Are we there yet?,” *J. Clin. Microbiol.*, vol. 54, no. 8, pp. 1956–1963, Aug. 2016, doi: 10.1128/JCM.00200-16.

[96] G. Muyzer, E. C. De Waal, and A. G. Uitterlinden, “Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA,” *Appl. Environ. Microbiol.*, vol. 59, no. 3, pp. 695–700, 1993, doi: 10.1128/aem.59.3.695-700.1993.

[97] F. Ju and T. Zhang, “16S rRNA gene high-throughput sequencing data mining of microbial diversity and interactions,” *Applied Microbiology and Biotechnology*, vol. 99, no. 10. Springer, pp. 4119–4129, Mar. 27, 2015, doi: 10.1007/s00253-015-6536-y.

[98] P. Yarza *et al.*, “Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences,” *Nat. Rev. Microbiol.*, vol. 12, no. 9, pp. 635–645, 2014, doi: 10.1038/nrmicro3330.

[99] Y. Van de Peer, S. Chapelle, and R. De Wachter, “A quantitative map of nucleotide substitution rates in bacterial rRNA,” *Nucleic Acids Res.*, vol. 24, no. 17, pp. 3381–3391, 1996, doi: 10.1093/nar/24.17.3381.

[100] S. Chakravorty, D. Helb, M. Burday, N. Connell, and D. Alland, “A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria,” *J. Microbiol. Methods*, vol. 69, no. 2, pp. 330–339, May 2007, doi: 10.1016/j.mimet.2007.02.005.

[101] J. Wagner, P. Coupland, H. P. Browne, T. D. Lawley, S. C. Francis, and J. Parkhill, “Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification,” *BMC Microbiol.*, vol. 16, no. 1, pp. 1–17, Nov. 2016, doi: 10.1186/s12866-016-0891-4.

[102] P. D. Schloss, M. L. Jenior, C. C. Koumpouras, S. L. Westcott, and S. K. Highlander, “Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system,” *PeerJ*, vol. 2016, no. 3, 2016, doi: 10.7717/peerj.1869.

[103] A. Rhoads and K. F. Au, “PacBio Sequencing and Its Applications,” *Genomics, Proteomics and Bioinformatics*, vol. 13, no. 5. Elsevier, pp. 278–289, 2015, doi: 10.1016/j.gpb.2015.08.002.

[104] A. Iversen, F. Asche, Ø. Hermansen, and R. Nystøyl, “Production cost and competitiveness in major salmon farming countries 2003–2018,” *Aquaculture*, vol. 522, p. 735089, May 2020, doi: 10.1016/j.aquaculture.2020.735089.

[105] Fiskeridirektoratet (organization), “Atlantic salmon, rainbow trout and trout - Grow out production.” Accessed: Apr. 08, 2021. [Online]. Available: https://www.fiskeridir.no/Akvakultur/Tall-og-analyse/Akvakulturstatistikk-tidsserier/Laks-regnbueoerret-og-oerret/Matfiskproduksjon.

[106] R. B. M. Pincinato, F. Asche, H. Bleie, A. Skrudland, and M. Stormoen, “Factors influencing production loss in salmonid farming,” *Aquaculture*, vol. 532, p. 736034, Feb. 2021, doi: 10.1016/j.aquaculture.2020.736034.

[107] M. S. Llewellyn *et al.*, “Parasitism perturbs the mucosal microbiome of Atlantic Salmon,” *Sci. Rep.*, vol. 7, Mar. 2017, doi: 10.1038/srep43465.

[108] C. Wang, G. Sun, S. Li, X. Li, and Y. Liu, “Intestinal microbiota of healthy and unhealthy Atlantic salmon *Salmo salar* L. in a recirculating aquaculture system,” *J. Oceanol. Limnol.*, vol. 36, no. 2, pp. 414–426, Mar. 2018, doi: 10.1007/s00343-017-6203-5.

[109] M. S. Llewellyn *et al.*, “The biogeography of the atlantic salmon (*Salmo salar*) gut microbiome,” *ISME J.*, vol. 10, no. 5, pp. 1280–1284, May 2016, doi: 10.1038/ismej.2015.189.

[110] T. M. Uren Webster, S. Consuegra, M. Hitchings, and C. Garcia de Leaniz, “Interpopulation Variation in the Atlantic Salmon Microbiome Reflects Environmental and Genetic Diversity,” *Appl. Environ. Microbiol.*, vol. 84, no. 16, Aug. 2018, doi: 10.1128/aem.00691-18.

[111] J. Lokesh and V. Kiron, “Transition from freshwater to seawater reshapes the skin-associated microbiota of Atlantic salmon,” *Sci. Rep.*, vol. 6, Jan. 2016, doi: 10.1038/srep19707.

[112] C. Lavoie, M. Courcelle, B. Redivo, and N. Derome, “Structural and compositional mismatch between captive and wild Atlantic salmon (*Salmo salar*) parrs’ gut microbiota highlights the relevance of integrating molecular ecology for management and conservation methods,” *Evol. Appl.*, vol. 11, no. 9, pp. 1671–1685, Oct. 2018, doi: 10.1111/eva.12658.

[113] J. J. Minich *et al.*, “Microbial ecology of atlantic salmon (*Salmo salar*) hatcheries: Impacts of the built environment on fish mucosal microbiota,” *Appl. Environ. Microbiol.*, vol. 86, no. 12, Jun. 2020, doi: 10.1128/AEM.00411-20.

[114] A. Villasante, C. Ramírez, N. Catalán, R. Opazo, P. Dantagnan, and J. Romero, “Effect of dietary carbohydrate-to-protein ratio on gut microbiota in atlantic salmon (*Salmo salar*),” *Animals*, vol. 9, no. 3, Mar. 2019, doi: 10.3390/ani9030089.

[115] K. Gajardo *et al.*, “Alternative protein sources in the diet modulate microbiota and functionality in the distal intestine of Atlantic salmon (*Salmo salar*),” *Appl. Environ. Microbiol.*, vol. 83, no. 5, 2017, doi: 10.1128/AEM.02615-16.

[116] S. Gupta, J. Fernandes, and V. Kiron, “Antibiotic-induced perturbations are manifested in the dominant intestinal bacterial phyla of Atlantic salmon,” *Microorganisms*, vol. 7, no. 8, Aug. 2019, doi: 10.3390/microorganisms7080233.

[117] K. Gajardo *et al.*, “A high-resolution map of the gut microbiota in Atlantic salmon (*Salmo salar*): A basis for comparative gut microbial research,” *Sci. Rep.*, vol. 6, Aug. 2016, doi: 10.1038/srep30893.

[118] R. C. Edgar, “Updating the 97% identity threshold for 16S ribosomal RNA OTUs,” *bioRxiv*, p. 192211, Sep. 2017, doi: 10.1101/192211.

[119] J. J. Barb *et al.*, “Development of an Analysis Pipeline Characterizing Multiple Hypervariable Regions of 16S rRNA Using Mock Samples,” *PLoS One*, vol. 11, no. 2, p. 148047, Feb. 2016, doi: 10.1371/journal.pone.0148047.

[120] F. Guo, F. Ju, L. Cai, and T. Zhang, “Taxonomic Precision of Different Hypervariable Regions of 16S rRNA Gene and Annotation Methods for Functional Bacterial Groups in Biological Wastewater Treatment,” *PLoS One*, vol. 8, no. 10, p. 76185, Oct. 2013, doi: 10.1371/journal.pone.0076185.

[121] C. Willis, D. Desai, and J. Laroche, “Influence of 16S rRNA variable region on perceived diversity of marine microbial communities of the Northern North Atlantic,” *FEMS Microbiol. Lett.*, vol. 366, no. 13, p. 152, Jul. 2019, doi: 10.1093/femsle/fnz152.

[122] E. L. Johnson, S. L. Heaver, W. A. Walters, and R. E. Ley, “Microbiome and metabolic disease: revisiting the bacterial phylum Bacteroidetes,” *Journal of Molecular Medicine*, vol. 95, no. 1. Springer, pp. 1–8, Nov. 29, 2017, doi: 10.1007/s00109-016-1492-2.

[123] A. Miccoli, M. Manni, S. Picchietti, and G. Scapigliati, “State-of-the-art vaccine research for aquaculture use: The case of three economically relevant fish species,”

Vaccines, vol. 9, no. 2. Multidisciplinary Digital Publishing Institute (MDPI), pp. 1–29, Feb. 01, 2021, doi: 10.3390/vaccines9020140.

[124] C. T. Parker, B. J. Tindall, and G. M. Garrity, "International Code of Nomenclature of Prokaryotes," *Int. J. Syst. Evol. Microbiol.*, vol. 69, no. 1, p. S1, Jan. 2019, doi: 10.1099/ijsem.0.000778.

[125] J. C. Ast, H. Urbanczyk, and P. V. Dunlap, "Multi-gene analysis reveals previously unrecognized phylogenetic diversity in *Aliivibrio*," *Syst. Appl. Microbiol.*, vol. 32, no. 6, pp. 379–386, Sep. 2009, doi: 10.1016/j.syapm.2009.04.005.

[126] S. Yoshizawa, H. Karatani, M. Wada, A. Yokota, and K. Kogure, "*Aliivibrio sifiae* sp. nov., luminous marine bacteria isolated from seawater," *J. Gen. Appl. Microbiol.*, vol. 56, no. 6, pp. 509–518, 2010, doi: 10.2323/jgam.56.509.

[127] J. L. Reichelt and P. Baumann, "Taxonomy of the marine, luminous bacteria," *Arch. Mikrobiol.*, vol. 94, no. 4, pp. 283–330, Dec. 1973, doi: 10.1007/BF00769027.

[128] C. Bongrand and E. G. Ruby, "The impact of *Vibrio fischeri* strain variation on host colonization," *Current Opinion in Microbiology*, vol. 50. Elsevier Ltd, pp. 15–19, Aug. 01, 2019, doi: 10.1016/j.mib.2019.09.002.

[129] R. Spencer, "The Taxonomy of certain Luminous Bacteria," *J. Gen. Microbiol.*, vol. 13, no. 1, pp. 111–118, Aug. 1955, doi: 10.1099/00221287-13-1-111.

[130] M. S. Hendrie, W. Hodgkiss, and J. M. Shewan, "The Identification, Taxonomy and Classification of Luminous Bacteria," *J. Gen. Microbiol.*, vol. 64, no. 2, pp. 151–169, Dec. 1970, doi: 10.1099/00221287-64-2-151.

[131] V. B. D. Skerman, V. McGowan, and P. H. A. Sneath, "Approved lists of bacterial names," *Int. J. Syst. Bacteriol.*, vol. 30, no. 1, pp. 225–420, Jan. 1980, doi: 10.1099/00207713-30-1-225.

[132] H. Urbanczyk, J. C. Ast, M. J. Higgins, J. Carson, and P. V. Dunlap, "Reclassification of *Vibrio fischeri*, *Vibrio logei*, *Vibrio salmonicida* and *Vibrio wodanis* as *Aliivibrio fischeri* gen. nov., comb. nov., *Aliivibrio logei* comb. nov., *Aliivibrio salmonicida* comb. nov. and *Aliivibrio wodanis* comb. nov.," *Int. J. Syst. Evol. Microbiol.*, vol. 57, no. 12, pp. 2823–2829, Dec. 2007, doi: 10.1099/ijs.0.65081-0.

[133] R. R. Colwell, "Proposal of a neotype, ATCC 15381, for *Vibrio marinus* (Russell 1891) Ford 1927 and request for an opinion," *Int. Bull. Bacteriol. Nomencl. Taxon.*, vol. 15, no. 3, pp. 165–176, Jul. 1965, doi: 10.1099/00207713-15-3-165.

[134] S. S. Bang, P. Baumann, and K. H. Nealson, "Phenotypic characterization of *Photobacterium logei* (sp. nov.), a species related to *P. fischeri*," *Curr. Microbiol.*, vol. 1, no. 5, pp. 285–288, Sep. 1978, doi: 10.1007/BF02601683.

[135] I. V Manukhov, S. A. Khrul'nova, A. Baranova, and G. B. Zavilgelsky, "Comparative analysis of the lux operons in *Aliivibrio logei* KCh1 (a Kamchatka Isolate) and *Aliivibrio salmonicida*," *Journal of Bacteriology*, vol. 193, no. 15. American Society for Microbiology Journals, pp. 3998–4001, Aug. 01, 2011, doi: 10.1128/JB.05320-11.

[136] P. M. Fidopiastis, S. von Boletzky, and E. G. Ruby, "A new niche for *Vibrio logei*, the predominant light organ symbiont of squids in the genus *Sepiola*.," *J. Bacteriol.*, vol. 180, no. 1, pp. 59–64, Jan. 1998, doi: 10.1128/JB.180.1.59-64.1998.

[137] T. Lunder, O. Evensen, G. Holstad, and T. Hastein, "Winter ulcer' in the Atlantic salmon *Salmo salar*. Pathological and bacteriological investigations and transmission experiments," *Dis. Aquat. Organ.*, vol. 23, no. 1, pp. 39–49, Sep. 1995, doi: 10.3354/dao023039.

[138] T. Lunder, H. Sorum, G. Holstad, A. G. Steigerwalt, P. Mowinckel, and D. J. Brenner, "Phenotypic and genotypic characterization of *Vibrio viscosus* sp. nov. and *Vibrio wodanis* sp. nov. isolated from Atlantic salmon (*Salmo salar*) with 'winter ulcer,'" *Int. J. Syst. Evol. Microbiol.*, vol. 50, no. 2, pp. 427–450, Mar. 2000, doi: 10.1099/00207713-50-2-427.

[139] E. Egidius, R. Wiik, K. Andersen, K. A. Hoff, and B. Hjeltnes, "*Vibrio salmonicida* sp. nov., a New Fish Pathogen," *Int. J. Syst. Bacteriol.*, vol. 36, no. 4, pp. 518–520, Oct. 1986, doi: 10.1099/00207713-36-4-518.

[140] A. Kashulin, N. Seredkina, and H. Sørum, "Cold-water vibriosis. The current status of knowledge," *Journal of Fish Diseases*, vol. 40, no. 1. John Wiley & Sons, Ltd (10.1111), pp. 119–126, Jan. 01, 2017, doi: 10.1111/jfd.12465.

[141] R. Beaz-Hidalgo, A. Doce, S. Balboa, J. L. Barja, and J. L. Romalde, "*Aliivibrio finisterrensis* sp. nov., isolated from Manila clam, *Ruditapes philippinarum* and emended description of the genus *Aliivibrio*," *Int. J. Syst. Evol. Microbiol.*, vol. 60, no. 1, pp. 223–228, Jan. 2010, doi: 10.1099/ijs.0.010710-0.

[142] E. Hatje, C. Neuman, H. Stevenson, J. P. Bowman, and M. Katouli, "Population Dynamics of *Vibrio* and *Pseudomonas* Species Isolated from Farmed Tasmanian Atlantic Salmon (*Salmo salar* L.): A Seasonal Study," *Microb. Ecol.*, vol. 68, no. 4, pp. 679–687, 2014, doi: 10.1007/s00248-014-0462-x.

[143] H. Sugita, H. Mizuki, and S. Itoi, "Diversity of siderophore-producing bacteria isolated from the intestinal tracts of fish along the Japanese coast," *Aquac. Res.*, vol. 43, no. 4, pp. 481–488, Mar. 2012, doi: 10.1111/j.1365-2109.2011.02851.x.

[144] P. Vandamme *et al.*, *Polyphasic taxonomy, a consensus approach to bacterial systematics*, vol. 60, no. 2. American Society for Microbiology, 1996, pp. 407–438.

[145] T. Sawabe, K. Kita-Tsukamoto, and F. L. Thompson, "Inferring the Evolutionary History of Vibrios by Means of Multilocus Sequence Analysis," *J. Bacteriol.*, vol. 189, no. 21, pp. 7932–7936, Nov. 2007, doi: 10.1128/JB.00693-07.

[146] T. Sawabe *et al.*, "Updating the Vibrio clades defined by multilocus sequence phylogeny: proposal of eight new clades, and the description of *Vibrio tritonius* sp. nov.," *Front. Microbiol.*, vol. 4, p. 414, Dec. 2013, doi: 10.3389/fmicb.2013.00414.

[147] J. Pascual, M. C. Macián, D. R. Arahal, E. Garay, and M. J. Pujalte, "Multilocus sequence analysis of the central clade of the genus *Vibrio* by using the 16S rRNA, *recA, pyrH, rpoD, gyrB, rctB* and *toxR* genes," *Int. J. Syst. Evol. Microbiol.*, vol. 60, no. 1, pp. 154–165, Jan. 2010, doi: 10.1099/ijs.0.010702-0.

[148] H. Machado and L. Gram, "The *fur* gene as a new phylogenetic marker for *Vibrionaceae* species identification," *Appl. Environ. Microbiol.*, vol. 81, no. 8, pp. 2745–2752, 2015, doi: 10.1128/AEM.00058-15.

[149] J. R. Cole, K. Konstantinidis, R. J. Farris, and J. Tiedje, "Microbial diversity and phylogeny: Extending from rRNAs to genomes," *Environ. Mol. Microbiol.*, pp. 1–19, 2010, [Online]. Available: https://www.researchgate.net/publication/285237823_Microbial_diversity_and_phylogeny_Extending_from_rRNAs_to_genomes.

[150] R. Rosselló-Mora and R. Amann, "The species concept for prokaryotes," *FEMS Microbiol. Rev.*, vol. 25, no. 1, pp. 39–67, Jan. 2001, doi: 10.1111/j.1574-6976.2001.tb00571.x.

[151] A. Mira, A. B. Martín-Cuadrado, G. D'Auria, and F. Rodríguez-Valera, "The bacterial pan-genome: A new paradigm in microbiology," *Int. Microbiol.*, vol. 13, no. 2, pp. 45–57, Sep. 2010, doi: 10.2436/20.1501.01.110.

[152] M. Richter and R. Rosselló-Móra, “Shifting the genomic gold standard for the prokaryotic species definition,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 45, pp. 19126–19131, Nov. 2009, doi: 10.1073/pnas.0906412106.

[153] T. Klemetsen *et al.*, “The MAR databases: development and implementation of databases specific for marine metagenomics,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D692–D699, Jan. 2018, doi: 10.1093/nar/gkx1036.

[154] T. Klemetsen, N. P. Willassen, and C. Karlsen, “Full-length 16S rRNA gene classification of Atlantic salmon bacteria and effects of using different 16S variable regions on community structure analysis,” *Microbiologyopen*, 2019, doi: 10.1002/mbo3.898.

[155] T. Klemetsen, C. R. Karlsen, and N. P. Willassen, “Phylogenetic Revision of the Genus *Aliivibrio*: Intra- and Inter-Species Variance Among Clusters Suggest a Wider Diversity of Species,” *Front. Microbiol.*, vol. 12, p. 272, Feb. 2021, doi: 10.3389/fmicb.2021.626759.

[156] L. C. Crosswell and J. M. Thornton, “ELIXIR: a distributed infrastructure for European biological data,” *Trends Biotechnol.*, vol. 30, no. 5, pp. 241–242, May 2012, doi: 10.1016/j.tibtech.2012.02.002.

[157] D. Field *et al.*, “The minimum information about a genome sequence (MIGS) specification,” *Nature Biotechnology*, vol. 26, no. 5. Nature Publishing Group, pp. 541–547, May 08, 2008, doi: 10.1038/nbt1360.

[158] E. Fadeev, F. De Pascale, A. Vezzi, S. Hübner, D. Aharonovich, and D. Sher, “Why close a bacterial genome? The plasmid of *Alteromonas macleodii* HOT1A3 is a vector for inter-specific transfer of a flexible genomic Island,” *Front. Microbiol.*, vol. 7, no. MAR, p. 248, Mar. 2016, doi: 10.3389/fmicb.2016.00248.

[159] M. Hunt, C. Newbold, M. Berriman, and T. D. Otto, “A comprehensive evaluation of assembly scaffolding tools,” *Genome Biol.*, vol. 15, no. 3, p. R42, Mar. 2014, doi: 10.1186/gb-2014-15-3-r42.

[160] D. H. Haft *et al.*, “RefSeq: An update on prokaryotic genome annotation and curation,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D851–D860, Jan. 2018, doi: 10.1093/nar/gkx1068.

[161] Finkl C.W., “Coastal Classification: Systematic Approaches to Consider in the Development of a Comprehensive Scheme,” *J. Coast. Res.*, vol. 20(1), pp. 166–213, 2004, doi: 10.2112/1551-5036(2004)20[166:CCSATC]2.0.CO;2.

[162] G. C. Ray *et al.*, “Interim guidelines for identification and selection of coastal biosphere reserves,” 1981. [Online]. Available: http://npshistory.com/publications/mab/us-mab-report/6.pdf.

[163] H. Sass and R. J. Parkes, “Sub-seafloor Sediments: An Extreme but Globally Significant Prokaryotic Habitat (Taxonomy, Diversity, Ecology),” in *Extremophiles Handbook*, Springer, Tokyo, 2011, pp. 1015–1041.

[164] N. Gunde-Cimerman, J. Ramos, and A. Plemenitaš, “Halotolerant and halophilic fungi,” *Mycological Research*, vol. 113, no. 11. Elsevier, pp. 1231–1241, Nov. 01, 2009, doi: 10.1016/j.mycres.2009.09.002.

[165] P. L. Buttigieg, E. Pafilis, S. E. Lewis, M. P. Schildhauer, R. L. Walls, and C. J. Mungall, “The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation,” *J. Biomed. Semantics*, vol. 7, no. 1, p. 57, Dec. 2016, doi: 10.1186/s13326-016-0097-6.

[166] E. E. Snyder *et al.*, “PATRIC: The VBI PathoSystems Resource Integration Center,” *Nucleic Acids Res.*, vol. 35, no. SUPPL. 1, Jan. 2007, doi: 10.1093/nar/gkl858.

[167] E. Karsenti *et al.*, "A holistic approach to marine Eco-systems biology," *PLoS Biol.*, vol. 9, no. 10, p. e1001177, Oct. 2011, doi: 10.1371/journal.pbio.1001177.

[168] B. J. Tully, E. D. Graham, and J. F. Heidelberg, "The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans," *Sci. Data*, vol. 5, Jan. 2018, doi: 10.1038/sdata.2017.203.

[169] L. W. Hugerth *et al.*, "Metagenome-assembled genomes uncover a global brackish microbiome," *Genome Biol.*, vol. 16, no. 1, Dec. 2015, doi: 10.1186/s13059-015-0834-7.

[170] Y. Marcy *et al.*, "Dissecting biological 'dark matter' with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 29, pp. 11889–11894, Jul. 2007, doi: 10.1073/pnas.0704662104.

[171] K. Blin *et al.*, "AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W81–W87, Jul. 2019, doi: 10.1093/nar/gkz310.

[172] D. Giordano *et al.*, "Marine Microbial Secondary Metabolites: Pathways, Evolution and Physiological Roles," in *Advances in Microbial Physiology*, vol. 66, Academic Press, 2015, pp. 357–428.

[173] P. A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks, "GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database," *Bioinformatics*, vol. 36, no. 6, pp. 1925–1927, Mar. 2020, doi: 10.1093/bioinformatics/btz848.

[174] A. M. Kozlov, J. Zhang, P. Yilmaz, F. O. Glöckner, and A. Stamatakis, "Phylogeny-aware identification and correction of taxonomically mislabeled sequences," *Nucleic Acids Res.*, vol. 44, no. 11, pp. 5022–5033, Jun. 2016, doi: 10.1093/nar/gkw396.

[175] E. M. Robertsen *et al.*, "META-pipe - Pipeline Annotation, Analysis and Visualization of Marine Metagenomic Sequence Data," Apr. 2016, Accessed: Aug. 21, 2021. [Online]. Available: https://arxiv.org/abs/1604.04103v1.

[176] S. G. Acinas *et al.*, "Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities," *Commun. Biol.*, vol. 4, no. 1, Dec. 2021, doi: 10.1038/s42003-021-02112-2.

[177] P. A. Kitts *et al.*, "Assembly: A resource for assembled genomes at NCBI," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D73–D80, 2016, doi: 10.1093/nar/gkv1226.

[178] D. E. Wood and S. L. Salzberg, "Kraken: Ultrafast metagenomic sequence classification using exact alignments," *Genome Biol.*, vol. 15, no. 3, pp. 1–12, Mar. 2014, doi: 10.1186/gb-2014-15-3-r46.

[179] D. J. Nasko, S. Koren, A. M. Phillippy, and T. J. Treangen, "RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification," *Genome Biol.*, vol. 19, no. 1, pp. 1–10, Oct. 2018, doi: 10.1186/s13059-018-1554-6.

[180] H. R. Gruber-Vodicka, B. K. B. Seah, E. Pruesse, G.-V. HR, S. BKB, and P. E, "phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes," *mSystems*, vol. 5, no. 5, Oct. 2020, doi: 10.1128/MSYSTEMS.00920-20.

[181] V. Schmidt, L. Amaral-Zettler, J. Davidson, S. Summerfelt, and C. Good, "Influence of fishmeal-free diets on microbial communities in atlantic salmon (*Salmo Salar*) recirculation aquaculture systems," *Appl. Environ. Microbiol.*, vol. 82, no. 15, pp. 4470–4481, 2016, doi: 10.1128/AEM.00902-16.

[182] A. Lanzén *et al.*, “CREST - Classification Resources for Environmental Sequence Tags,” *PLoS One*, vol. 7, no. 11, Nov. 2012, doi: 10.1371/journal.pone.0049334.

[183] C. Fogarty *et al.*, “Diversity and composition of the gut microbiota of Atlantic salmon (*Salmo salar*) farmed in Irish waters,” *J. Appl. Microbiol.*, vol. 127, no. 3, pp. 648–657, Sep. 2019, doi: 10.1111/JAM.14291.

[184] W. E. Holben, P. Williams, M. Saarinen, L. K. Särkilahti, and J. H. A. Apajalahti, “Phylogenetic Analysis of Intestinal Microflora Indicates a Novel *Mycoplasma* Phylotype in Farmed and Wild Salmon,” *Microb. Ecol. 2002 442*, vol. 44, no. 2, pp. 175–185, 2002, doi: 10.1007/S00248-002-1011-6.

[185] Y. Jin, I. L. Angell, S. R. Sandve, L. G. Snipen, Y. Olsen, and K. Rudi, “Atlantic salmon raised with diets low in long-chain polyunsaturated n-3 fatty acids in freshwater have a *Mycoplasma*-dominated gut microbiota at sea,” *Aquac. Environ. Interact.*, vol. 11, pp. 31–39, Jan. 2019, doi: 10.3354/AEI00297.

[186] A. Suau *et al.*, “Direct analysis of genes encoding 16S rRNA from complex communities reveals many novel molecular species within the human gut,” *Appl. Environ. Microbiol.*, vol. 65, no. 11, pp. 4799–4807, 1999, doi: 10.1128/aem.65.11.4799-4807.1999.

[187] E. Bolyen *et al.*, “Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2,” *Nat. Biotechnol.*, vol. 37, no. 8, pp. 852–857, Aug. 2019, doi: 10.1038/s41587-019-0209-9.

[188] E. F. Boyd *et al.*, “Post-Genomic Analysis of Members of the Family *Vibrionaceae*,” *Microbiol. Spectr.*, vol. 3, no. 5, Oct. 2015, doi: 10.1128/microbiolspec.VE-0009-2014.

[189] C. Karlsen, C. Vanberg, H. Mikkelsen, and H. Sørum, “Co-infection of Atlantic salmon (*Salmo salar*), by *Moritella viscosa* and *Aliivibrio wodanis*, development of disease and host colonization.,” *Vet. Microbiol.*, vol. 171, no. 1–2, pp. 112–21, Jun. 2014, doi: 10.1016/j.vetmic.2014.03.011.

[190] F. Menardo *et al.*, “Treemmer: A tool to reduce large phylogenetic datasets with minimal loss of diversity,” *BMC Bioinformatics*, vol. 19, no. 1, May 2018, doi: 10.1186/s12859-018-2164-8.

[191] D. H. Huson and D. Bryant, “Application of phylogenetic networks in evolutionary studies,” *Molecular Biology and Evolution*, vol. 23, no. 2. pp. 254–267, Feb. 01, 2005, doi: 10.1093/molbev/msj030.

[192] Y. Lan, G. Rosen, R. Hershberg, L. Y, R. G, and H. R, “Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains,” *Microbiome*, vol. 4, no. 1, 2016, doi: 10.1186/s40168-016-0162-5.

[193] A. Bateman *et al.*, “UniProt: The universal protein knowledgebase in 2021,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D480–D489, Jan. 2021, doi: 10.1093/nar/gkaa1100.

[194] J. R. Brister, D. Ako-Adjei, Y. Bao, and O. Blinkova, “NCBI viral Genomes resource,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D571–D577, Jan. 2015, doi: 10.1093/nar/gku1207.

[195] F. Zeng, Z. Wang, Y. Wang, J. Zhou, and T. Chen, “Large-scale 16S gene assembly using metagenomics shotgun sequences,” *Bioinformatics*, vol. 33, no. 10, pp. 1447–1456, May 2017, doi: 10.1093/bioinformatics/btx018.

[196] T. R. Lesker *et al.*, “An Integrated Metagenome Catalog Reveals New Insights into the Murine Gut Microbiome,” *Cell Rep.*, vol. 30, no. 9, pp. 2909-2922.e6, Mar. 2020, doi: 10.1016/j.celrep.2020.02.036.

Part II – Scientific papers

Paper 1

T. Klemetsen et al., “The MAR databases: development and implementation of databases specific for marine metagenomics,” Nucleic Acids Res., vol. 46, no. D1, pp. D692–D699, Jan. 2018, doi: 10.1093/nar/gkx1036.

D692–D699 *Nucleic Acids Research, 2018, Vol. 46, Database issue* *Published online 2 November 2017*
doi: 10.1093/nar/gkx1036

The MAR databases: development and implementation of databases specific for marine metagenomics

Terje Klemetsen[1], Inge A. Raknes[1], Juan Fu[1], Alexander Agafonov[1], Sudhagar V. Balasundaram[1], Giacomo Tartari[1,2], Espen Robertsen[1] and Nils P. Willassen[1,*]

[1]Centre for Bioinformatics, Faculty of science and technology, UiT The Arctic University of Norway, PO Box 6050 Langnes, TromsøN-9037, Norway and [2]Department of Information Technology, UiT The Arctic University of Norway, PO Box 6050 Langnes, TromsøN-9037, Norway

Received August 14, 2017; Revised October 12, 2017; Editorial Decision October 13, 2017; Accepted October 18, 2017

ABSTRACT

We introduce the marine databases; *MarRef, MarDB and MarCat* (https://mmp.sfb.uit.no/databases/), which are publicly available resources that promote marine research and innovation. These data resources, which have been implemented in the Marine Metagenomics Portal (MMP) (https://mmp.sfb.uit.no/), are collections of richly annotated and manually curated contextual (metadata) and sequence databases representing three tiers of accuracy. While *MarRef* is a database for completely sequenced marine prokaryotic genomes, which represent a marine prokaryote reference genome database, *MarDB* includes all incomplete sequenced prokaryotic genomes regardless level of completeness. The last database, *MarCat,* represents a gene (protein) catalog of uncultivable (and cultivable) marine genes and proteins derived from marine metagenomics samples. The first versions of *MarRef* and *MarDB* contain 612 and 3726 records, respectively. Each record is built up of 106 metadata fields including attributes for sampling, sequencing, assembly and annotation in addition to the organism and taxonomic information. Currently, *MarCat* contains 1227 records with 55 metadata fields. Ontologies and controlled vocabularies are used in the contextual databases to enhance consistency. The user-friendly web interface lets the visitors browse, filter and search in the contextual databases and perform BLAST searches against the corresponding sequence databases. All contextual and sequence databases are freely accessible and downloadable from https://s1.sfb.uit.no/public/mar/.

INTRODUCTION

Microorganisms are ubiquitous in the marine environment, where they play key roles in many global and local biogeochemical processes such as nutrient recycling (1). These microorganisms and the communities they form, drive and respond to changes in the environment and alterations in the marine environment (2). With an estimated 10^4 to 10^6 cells per milliliter seawater and totally over 10^{29} bacterial cells in open sea, the marine microorganisms provide the grounds for immense genetic diversity (3).

Since the first complete bacterial genome published in 1995 (4), the number of sequenced microbial genomes has increased dramatically. Currently, more than 103 000 prokaryotic genomes are available in the National Center for Biotechnology Information (NCBI) Genome microbial database (https://www.ncbi.nlm.nih.gov/genome/microbes/). Originally sequencing efforts were prioritized to study cultured microbes. However, it is well established that the vast majority of bacterial and archaeal taxa remain uncultivated *in vitro* (5). Recently, cultivation-independent methods such as single cell genomics and genomes reconstructed from metagenomic deep sequencing, have begun to yield complete or near-complete genomes from many novel lineages (5–7). Metagenomics, the study of genetic material recovered directly from environmental samples, is a powerful tool for surveying the diversity of marine microbes, which are important for the study of marine sciences. Prominent examples of metagenomics studies in the marine field include the Sorcerer II expeditions (8), Malaspina expedition (9), Global Ocean Sampling (GOS) campaign (10) and Tara Oceans expedition (11). Most of these data as well as other marine metagenomic data are stored in publicly available metagenomic databases such as iMicrobe (https://www.imicrobe.us/), Viral Informatics Resource for Metagenome Exploration (VIROME) (12), EBI metagenomics (13), Integrated Microbial Genomes and Microbiomes (IMG/M) (14) and Metagenomics Rapid Annota-

[*]To whom correspondence should be addressed. Tel: +47 7764 4651; Email: nils-peder.willassen@uit.no

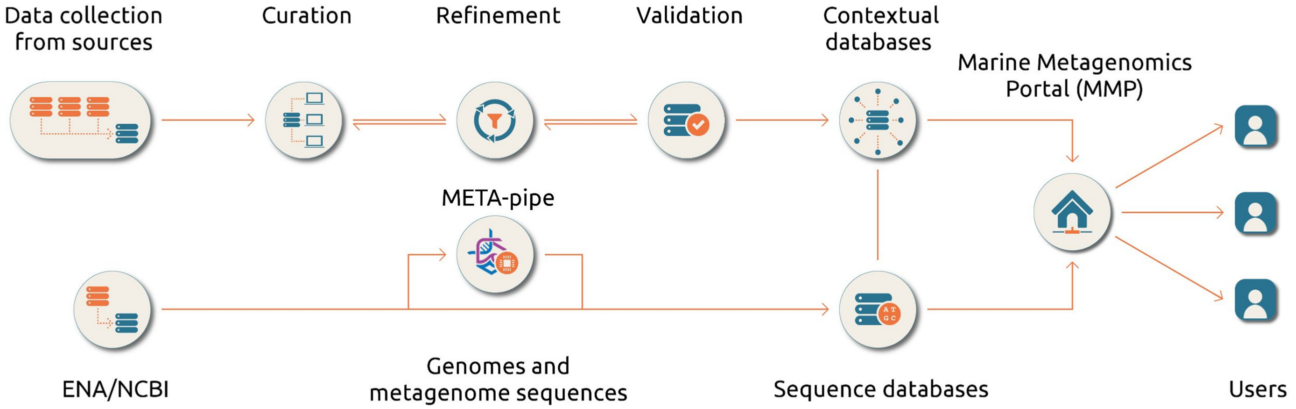

Data collection from sources
Curation
Refinement
Validation
Contextual databases
Marine Metagenomics Portal (MMP)
META-pipe
ENA/NCBI
Genomes and metagenome sequences
Sequence databases
Users

Figure 1. General and simplified procedures for construction of the MAR databases. The top part represents the flow of contextual data records from its collection to implementation on the web server. The bottom part illustrates how sequence data becomes implemented and processed. Only metagenomic sequences in relation with *MarCat* has been processed using META-pipe for the first release.

tion using Subsystem Technology (MG-RAST) (15). Reference sequence databases with comprehensive metadata are essential for analyzing and interpreting of marine metagenomic data (16,17). There are several general microbial databases e.g. Prokaryotic RefSeq Genomes (18), Genomes OnLine Database (GOLD) (19), Pathosystems Resource Integration Center (PATRIC) (20) and MicroScope (21), which contains marine microbial genomes. Even though the Microbial Ecological Genomics Database (MegDB), available at the Megx.net portal, includes marine bacterial, archaeal and phage genomes and metagenomes, it is mainly a georeference database which provides less metadata besides the geolocation information of the samples (22)**.**

Up to now, no dedicated sequence data resources exist for the marine metagenomics domain (17), which not only hamper the utilization of the vast genetic resources for biotechnology research and innovation (e.g. bioprospecting), but also impede the development of sustainable tools and resources aimed at environmental monitoring, monitoring of fish and shellfish pathogens and development of sustainable feed for marine aquaculture.

Since all research and innovation is based on comparison to existing knowledge and information, the lack of unified formats, controlled vocabularies (CV) and ontologies (formal specifications of the terms) make it difficult not only to identify records in databases but also to compare data within and/or between different databases. Therefore, sustainable and highly accurate data resources that are easy to access, browse and retrieve data from, are vital for performing high class and beyond the state of art research and innovation.

Here, we introduced the contextual and sequence MAR databases: *MarRef*, *MarDB* and *MarCat*, with manually curated metadata including attributes for sampling, sequencing, assembly and annotation in addition to the organism and taxonomic information and their corresponding nucleotide and protein sequences.

OVERVIEW OF THE RESOURCES

Definition of marine microbial biome

To define a ‘marine microbial biome’ or a ‘marine microorganism’ is not straightforward since there are many habitats, which are on the borderline between marine and terrestrial ecosystems, such as sandy shores and near river deltas. We have chosen to define a ‘*marine microbial biome*’ as ‘*An aquatic microbial biome comprises of microbial communities from open oceans, coastal and protected habitats up to the high-water mark with salinity from 0.5 ppt (parts per thousand) as in estuaries (brackish water) environments to above 100 ppt as in sea ice brine. The biome also includes marine microbial communities obtained from marine species associated with these habitats*’.

Additionally, we accept soil samples from sandy shores, intertidal zones, salt marshes (coastal salt marshes or tidal marshes), mudflats and estuaries, in addition to habitats such as seawater saltern, sea ice brines, black smokers (hydrothermal vents) where the salinity can be extremely high or low compared to the seawater. Microorganisms and microbiomes associated with marine species, as defined by the World Register of Marine Species (WoRMS) have also been defined as marine (23). This includes microorganisms associated with or causing diseases in marine animals and plants such as corals, shellfish, fish, macroalgae and seagrass.

Short description of MarRef, MarDB and MarCat

The construction of the marine contextual databases and their corresponding sequence databases (BLAST databases) are shown in Figure 1. Each genome or metagenome assigned to a ‘*marine microbial biome*’, according to our definition, is included in the databases.

The *MarRef*, *MarDB* and *MarCat* sequence databases are based on the non-redundant genome and metagenome datasets obtained from ENA (European Nucleotide Archive, http://www.ebi.ac.uk/ena) and NCBI (https://www.ncbi.nlm.nih.gov/). While *MarRef* is a database for completely sequenced marine prokaryotic genomes, *MarDB* includes all in-complete sequenced marine prokaryotic genomes regardless the level of completeness. *MarCat* represents a gene (protein) catalog of predicted marine genes derived from marine metagenomic samples. Metagenomic sequences were obtained from ENA and their corresponding gene and protein annotation unique to each sample was generated using META-pipe, a pipeline for taxonomic classification and functional annotation of metagenomic sample (arXiv:1604.04103). The corresponding contextual databases support the international community-driven standards of the Genomics Standards Consortium (http://gensc.org/) and are fully compliant with its recommendations for minimum information about any (x) sequence (MIxS) standards. These databases also include the proposed standards for provenance of analysis proposed by the ELIXIR EXCELERATE marine metagenomics community (24).

CONTEXTUAL DATABASES

Data collection

The *MarRef*, *MarDB* and *MarCat* contextual databases are built by compiling data from a number of publicly available sequence, taxonomy and literature databases in a semi-automatic fashion. Other databases or resources such as bacterial diversity and culture collections databases, web mapping services and ontology databases were used extensively for curation of metadata. Resources used in the curation of the marine databases are shown in Table 1.

Curation

For curation, imported data files were compiled, converted to tab separated value files (TAB) format and imported into base, a full-featured desktop database front end, provided by LibreOffice (https://no.libreoffice.org/).

MarRef and *MarDB* contain in total 612 and 3726 records (Figure 2), respectively, with 106 metadata fields, out of which 30 fields are represented by CV and the remaining are free text or numeric fields. These 106 metadata fields include information about sampling environment, the organism and taxonomy, phenotype, pathogenicity, secondary metabolites, assembly and annotation.

The gene (protein) catalog database derived from marine metagenomic samples, *MarCat,* contains 1227 records, including samples from the Tara Ocean expedition (248 records) and Ocean Sampling Day (150 records). Each record contains 55 metadata fields.

The use of CV and ontologies can shortly be described by the following example. There are three environmental metadata fields used for describing the sampling site of a microorganism in *MarRef* and *MarDB*; environmental *biome*, *feature* and *material* which are controlled by a total of 95 terms. The environmental *biome* metadata field contains 11 controlled Environment Ontology (ENVO) terms covering environments such as Estuarine biome (ENVO:01000020), Marginal sea biome (ENVO:01000046), Marine benthic biome (ENVO:01000024), Marine mud (ENVO:00005795), Marine pelagic biome (ENVO:01000023), Marine water body (ENVO:00001999) and Ocean biome (ENVO:01000048). The environmental *feature* and *material* metadata fields are controlled by 59 and 25 terms, respectively. The ontologies used in the environmental *biome*, *feature* and *material* fields are all well-defined and described (http://www.environmentontology.org/), allowing consistency across the datasets.

The databases link out to other publicly available resources. For example, in *MarRef* sixteen of the metadata fields have active links to the literature databases such as PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) and PMC Europe (https://europepmc.org/), ontology databases such as ENVO (https://bioportal.bioontology.org/ontologies/ENVO) and Gazetteer (GAZ) (https://bioportal.bioontology.org/ontologies/GAZ), sequence databases such as the universal protein resource (UniProt) (http://www.uniprot.org/proteomes/) and ENA, taxonomy databases such as NCBI Taxonomy (https://www.ncbi.nlm.nih.gov/taxonomy) and Silva (https://www.arb-silva.de/) and the Bacterial Diversity Metadatabase, BacDive (https://bacdive.dsmz.de/). Links to other external resources such as compound and secondary metabolites databases are provided if available. These links allow site visitors to easily access other web pages in order to obtain more information about each record.

For *MarRef*, all metadata fields have been manually curated to ensure consistency across the datasets, which allow the end user to easily search and filter records. While *MarRef* is thoroughly curated, *MarDB* and *MarCat* are only partly curated.

Records in the marine databases, *MarRef, MarDB* and *MarCat* follow the MlxS standard guidelines developed by the Genomic Standard Consortium, in addition to ontologies such as ENVO and GAZ.

Refinement and validation

OpenRefine (http://openrefine.org/) was used for refining the metadata fields by cleaning, trimming of leading and trailing whitespace, transforming data from one format into another and extending it with web services and external data. A validation tool was developed to convert the tab separated value files (TSV) to extensible markup language files (XML) and from TSV to XML to link the source TSV curation databases to the XML database. The validator defines a set of rules for the conversion–warnings and errors during conversion are reported.

SEQUENCE DATABASES

The *MarRef*, *MarDB* and *MarCat* sequence databases are based on the non-redundant genome and metagenome datasets obtained from ENA and NCBI and by manually inspection assigned as belonging to the '*marine microbial biome*' according to our definition.

MarRef and MarDB

While *MarRef* is a database for completely sequenced marine prokaryotic genomes, *MarDB* includes all remaining sequenced marine prokaryotic genomes regardless the level of completeness. Both the *MarRef* and *MarDB* databases

Table 1. Public data resources utilized for the construction of MarRef, MarDB and MarCat

Type	Database	URL
Sequence databases	ENA, European Nucleotide Archive	ebi.ac.uk/ena
	UniProt, Universal Protein Resource	uniprot.org
	NCBI, National Center for Biotechnology Information	ncbi.nlm.nih.gov
Contextual databases	PATRIC, Pathosystems Resource Integration Center	patricbrc.org
	GOLD, Genomes OnLine Database	gold.jgi.doe.gov
Taxonomic databases	SILVA, SILVA high quality ribosomal RNA database	arb-silva.de
	NCBI Taxonomy browser	ncbi.nlm.nih.gov/taxonomy
Bacterial diversity metadatabases	BacDive, Bacterial Diversity Metadatabase	bacdive.dsmz.de
Culture collection databases	DSMZ, Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH	dsmz.de
	ATCC, American Type Culture Collection	atcc.org
Marine organisms database	WoRMS, World Register of Marine Species	marinespecies.org
Web mapping service	Google maps	maps.google.com
Literature databases	Europe PMC, Europe PubMed Central	europepmc.org
	PubMed	ncbi.nlm.nih.gov/pubmed
	doi, Digital Object Identifier System	doi.org
Ontology databases	BioPortal	bioportal.bioontology.org
Standards MIGS/MIMS	GSC, Genomic Standards Consortium	gensc.org

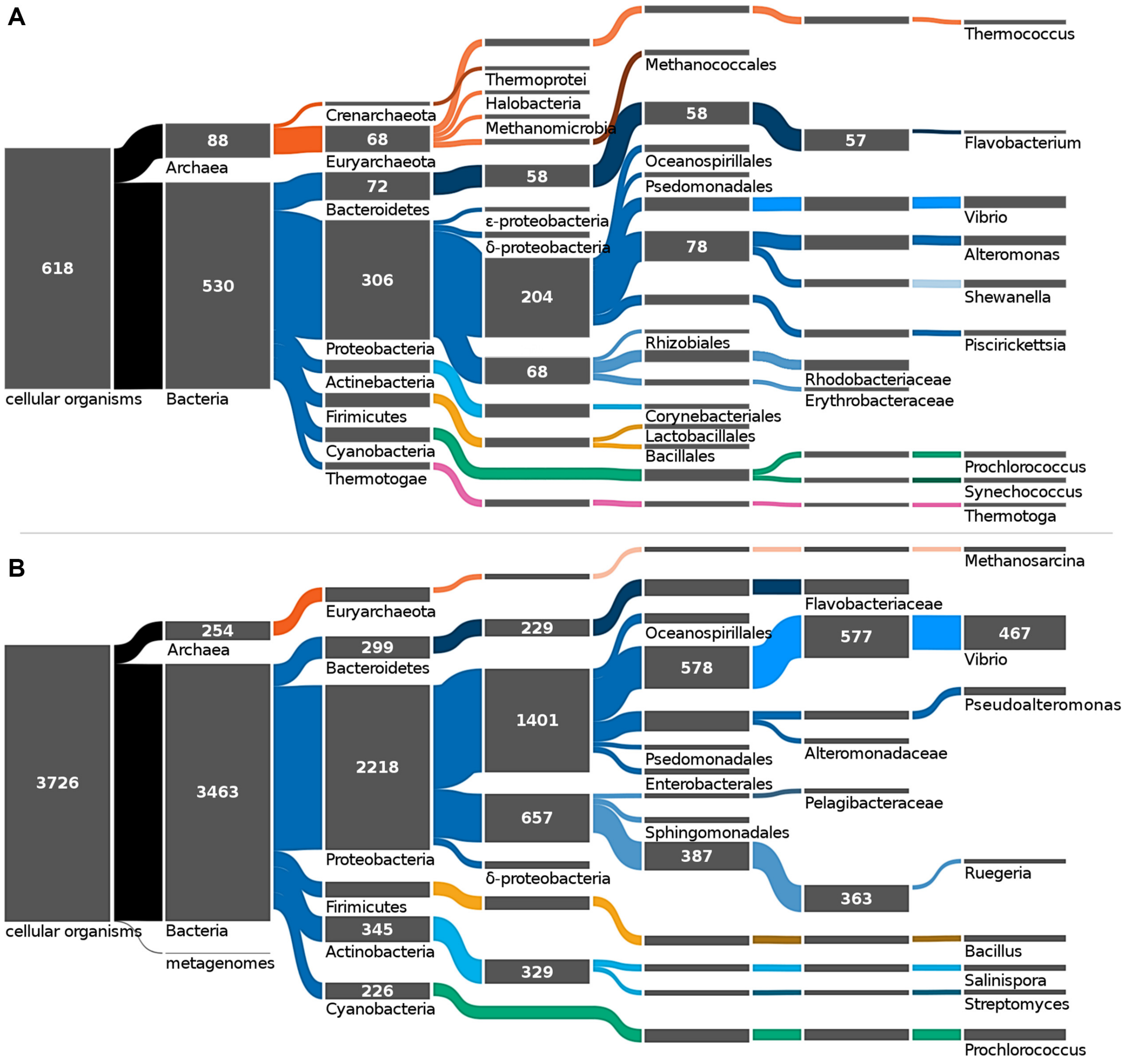


Figure 2. Most occurring marine taxa. (**A**) The reference database MarRef at its current state has 618 records of cellular organisms in the Archaea and Bacteria domains. Its complete and closed genomes are most prominent within the Proteobacteria phylum and the Alteromonadales order. (**B**) The partially curated database *MarDB* has 3726 records of sequenced genomes. Of its 287 unique genera (8 are shown) Vibrio is the most prominent with 467 records. These node-depleted Sankey diagrams were simplified to only display nodes exceeding 10 and 59 records for *MarRef* and *MarDB* respectively. An exception was made for the metagenome-derived genomes of *MarDB*.

primarily built on gene, protein and genome sequences obtained from the Prokaryotic RefSeq Genomes database (18). All archaeal and bacterial genomes in RefSeq have been annotated using the NCBI's Prokaryotic Genome Automatic Annotation Pipeline (PGAAP) (25). However, ∼20% of all records in *MarDB* did not have any RefSeq entry with PGAAP annotations. Circumventing the lack of gene and protein information of these genomes, annotation was performed on pre-assembled sequences using Prokka, a command line software tool, for annotation of prokaryotic genomes (26).

MarCat

MarCat represents a catalog of uncultivable (and cultivable) full-length genes (proteins) derived from marine metagenomic samples based on the Marine projects in EBI metagenomics (https://www.ebi.ac.uk/metagenomics/). Metagenomic sequence reads were downloaded from ENA and annotated using META-pipe (https://arxiv.org/abs/1604.04103). In short, sequencing reads were merged, filtered and assembled using MEGAHIT (27), which has been shown to be one the best assemblers for metagenomic samples in the Critical Assessment of Metagenome Interpretation (CAMI) challenge (28). From the resulting contigs, full-length CDSs were predicted using MetaGeneAnnotator (29) and functionally assigned using a compilation of results from BLAST against UniRef (30), Priam (31) and InterProScan5 (32). Using META-pipe for gene prediction and functional assignment allowed us to generate a consistent catalog across the datasets in *MarCat* (See https://f1000research.com/articles/6--70/v1 for a more detailed description of functional assignment). As a start, we used the high-coverage and high-quality sequence outputs from the Tara Oceans and Ocean Sampling Day metagenomic projects (10,11). In addition, more than 30 projects of various sizes were included based on EBI's marine projects. These were filtered in order to maintain the whole genome shotgun marine samples exclusively and also to avoid any project-interwoven freshwater samples. Some examples of these smaller projects include the Amazon continuum metagenomes (33) and western english channel diurnal study (34).

IMPLEMENTATION AND USER INTERFACE

The MAR databases have been incorporated into the Marine Metagenomics Portal (MMP) (https://mmp.sfb.uit.no/).

Contextual databases

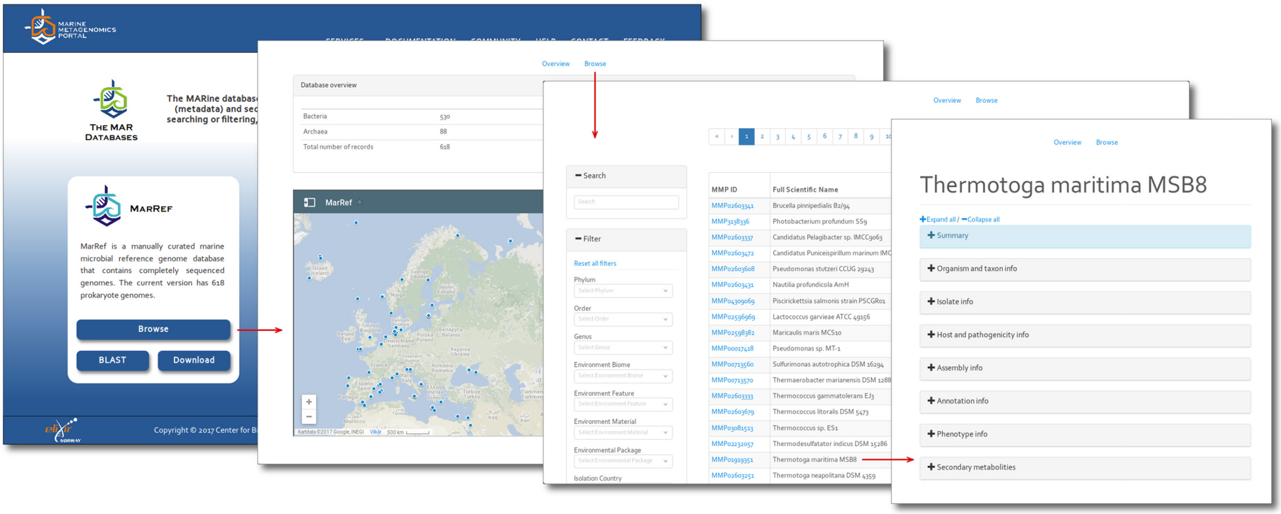
The contextual databases have been implemented using the hugo static website engine (https://gohugo.io/). The website engine reads the databases from XML files and allows the site visitor to access the information from four different layers. The first layer is the '*Database selection*' page, where the user can select the different MAR databases for browsing, BLAST sequences or downloading (Figure 3). The second layer is the specific database '*Overview*' page, which provides information about the content of the database and the geolocation of each genome/metagenome sample in the specific database. The geolocation has been embedded using google maps and each sample can be selected to display the organism/metagenome sample name and a short description of the organism/sample. The corresponding contextual information of the record can be reached by activating the MMP_ID link. The third layer is the '*Browse*' (Figure 4) page which can be reached from the 'Overview' page and allows the site visitors to:

i) *Browse* the database records of interest.
ii) *Search* across all metadata fields e.g. search for a specific organism, environmental ontologies, accession ID or any word.
iii) *Filter* records to be visible in the table based on the most important record attribute, such as taxonomy (phylum, order and genus) and environmental ontologies (biome, feature and material).
iv) *Advanced filtering* allows the site visitor to (a) add one or more filters; (b) refine current filters by adding new filters or removing already applied filters, (c) combine search and filtering and (d) remove all filters and launch a new search.

The search/filtered results will be listed in a table. Summary of the metadata will be shown when activating the 'Summary' button. The fourth layer contains the information for each record. The contextual data for a record can be viewed using the 'expand all' button. For the marine genome databases, *MarRef* and *MarDB*, the 106 metadata fields in the record is divided into seven categories; organism and taxon info, isolate info, phenotype info, secondary metabolites, host and pathogenicity info, assembly info and annotation info, in addition to Summary. For *MarCat*, the metagenome databases, the 55 metadata fields have been divided into four categories: isolate info, sampling info, host and pathogenicity info and assembly info, in addition to Summary.

BLAST

The BLAST (35) sequence databases provide similarity search against all nucleotide and protein sequences of records included in *MarRef*, *MarDB* and *MarCat*. The BLAST functionality was established using SequenceServer Version 1.09 (https://doi.org/10.1101/033142) to provide the graphical user interface for the search results. The SequenceServer allows the visitor to type, paste or drag-and-drop a FASTA file to search either a single or several databases. The interface automatically recognizes the sequence type and chooses the appropriate BLAST method and databases. Advanced parameters (command line) can be used to refine the search. The output of BLAST consists of a list of hits with the corresponding *E*-value, and a set of the traditional pairwise alignments were the target sequence can be viewed and downloaded. From the pairwise alignment the visitor can also retrieve information of the organism/metagenome sample in the MAR databases by opening the mmp button. In *MarRef* and *MarDB* information about the targets sequences can be obtained by opening



MarRef is a manually curated marine microbial reference genome database that contains completely sequenced genomes. The current version has 618 prokaryote genomes.

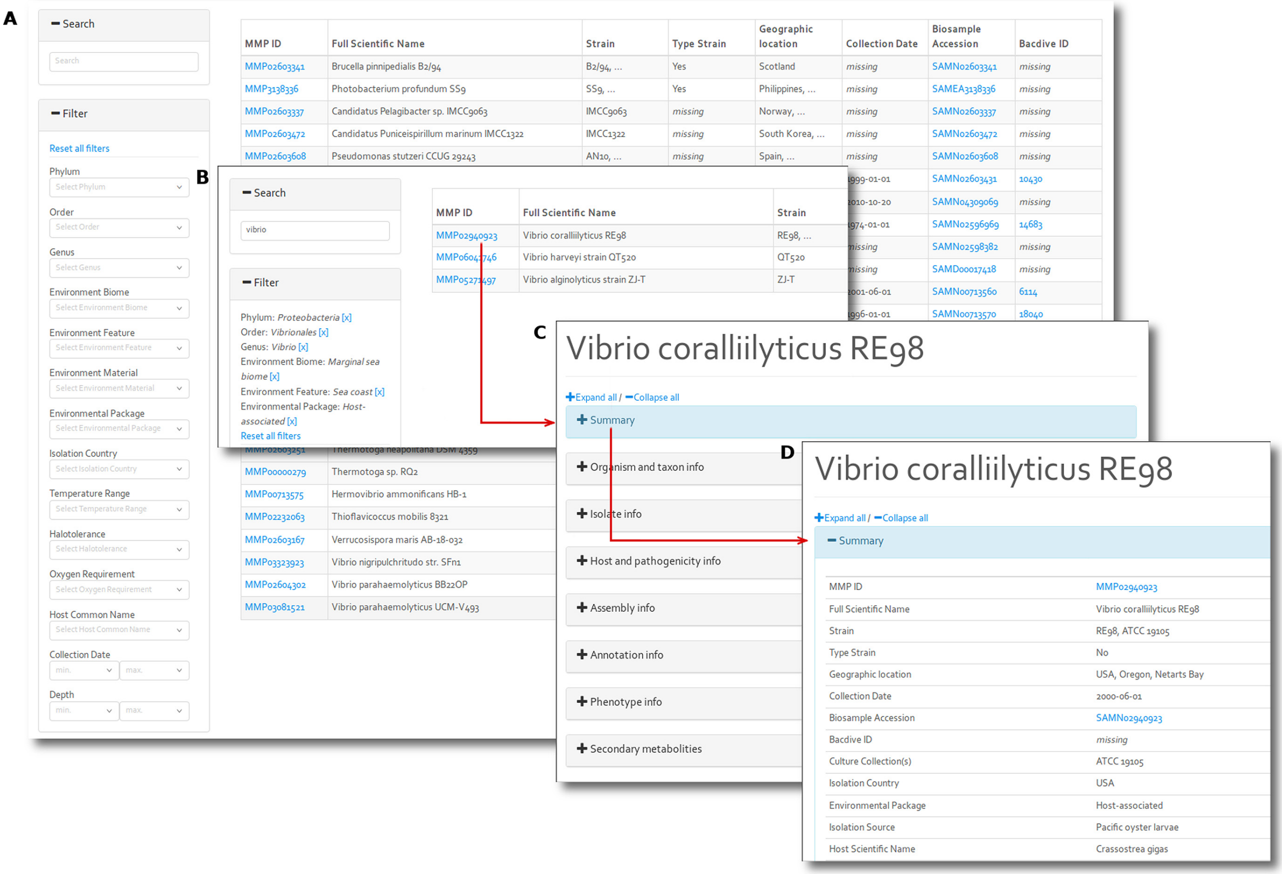
Database overview

Bacteria	530
Archaea	88
Total number of records	618

MMP ID	Full Scientific Name
MMP02603341	Brucella pinnipedialis B2/94
MMP3138336	Photobacterium profundum SS9
MMP02603337	Candidatus Pelagibacter sp. IMCC9063
MMP02603472	Candidatus Puniceispirillum marinum IMC
MMP02603608	Pseudomonas stutzeri CCUG 29243
MMP02603431	Nautilia profundicola AmH
MMP04309069	Piscirickettsia salmonis strain PSCGR01
MMP02596969	Lactococcus garvieae ATCC 49156
MMP02598382	Maricaulis maris MCS10
MMP00017418	Pseudomonas sp. MT-1
MMP00713560	Sulfurimonas autotrophica DSM 16294
MMP00713570	Thermaerobacter marianensis DSM 1288
MMP02603333	Thermococcus gammatolerans EJ3
MMP02603679	Thermococcus litoralis DSM 5473
MMP03081513	Thermococcus sp. ES1
MMP02232057	Thermodesulfatator indicus DSM 15286
MMP01919351	Thermotoga maritima MSB8
MMP02603251	Thermotoga neapolitana DSM 4359

Thermotoga maritima MSB8

Figure 3. Accessing the MAR databases and their records. From within the front page of the MMP all three metadatabases and sequence databases can be reached by following the 'Browse' or 'BLAST' buttons respectively. Browsing a metadatabase leads to the map-overview before reaching its index table. Single entries can be studied by selecting them in the map or in the table.



MMP ID	Full Scientific Name	Strain	Type Strain	Geographic location	Collection Date	Biosample Accession	Bacdive ID
MMP02603341	Brucella pinnipedialis B2/94	B2/94, ...	Yes	Scotland	*missing*	SAMN02603341	*missing*
MMP3138336	Photobacterium profundum SS9	SS9, ...	Yes	Philippines, ...	*missing*	SAMEA3138336	*missing*
MMP02603337	Candidatus Pelagibacter sp. IMCC9063	IMCC9063	*missing*	Norway, ...	*missing*	SAMN02603337	*missing*
MMP02603472	Candidatus Puniceispirillum marinum IMCC1322	IMCC1322	*missing*	South Korea, ...	*missing*	SAMN02603472	*missing*
MMP02603608	Pseudomonas stutzeri CCUG 29243	AN10, ...	*missing*	Spain, ...	*missing*	SAMN02603608	*missing*
					1999-01-01	SAMN02603431	10430
					2010-10-20	SAMN04309069	*missing*
					1974-01-01	SAMN02596969	14683
					missing	SAMN02598382	*missing*
					missing	SAMD00017418	*missing*
					2001-06-01	SAMN00713560	6114
					1996-01-01	SAMN00713570	18040
MMP02603251	Thermotoga neapolitana DSM 4359						
MMP00000279	Thermotoga sp. RQ2						
MMP00713575	Hermovibrio ammonificans HB-1						
MMP02232063	Thioflavicoccus mobilis 8321						
MMP02603167	Verrucosispora maris AB-18-032						
MMP03323923	Vibrio nigripulchritudo str. SFn1						
MMP02604302	Vibrio parahaemolyticus BB22OP						
MMP03081521	Vibrio parahaemolyticus UCM-V493						

Phylum: *Proteobacteria* [x]
Order: *Vibrionales* [x]
Genus: *Vibrio* [x]
Environment Biome: *Marginal sea biome* [x]
Environment Feature: *Sea coast* [x]
Environmental Package: *Host-associated* [x]
Reset all filters

MMP ID	Full Scientific Name	Strain
MMP02940923	Vibrio coralliilyticus RE98	RE98, ...
MMP06041746	Vibrio harveyi strain QT520	QT520
MMP05271497	Vibrio alginolyticus strain ZJ-T	ZJ-T

Vibrio coralliilyticus RE98

Vibrio coralliilyticus RE98

MMP ID	MMP02940923
Full Scientific Name	Vibrio coralliilyticus RE98
Strain	RE98, ATCC 19105
Type Strain	No
Geographic location	USA, Oregon, Netarts Bay
Collection Date	2000-06-01
Biosample Accession	SAMN02940923
Bacdive ID	*missing*
Culture Collection(s)	ATCC 19105
Isolation Country	USA
Environmental Package	Host-associated
Isolation Source	Pacific oyster larvae
Host Scientific Name	Crassostrea gigas

Figure 4. The browsing interface and filtering functionality of *MarRef* and *MarDB*. (**A**) The default view as accessed from the corresponding database overview menu. The table content is instantaneously updated when filtering and responds to search words and 14 filtering fields. (**B**) Combining search words and filters enables search criteria to narrow the listed results in a highly flexible manner. (**C**) The metadata of each record is separated in eight expandable categories, (**D**) here illustrating parts of the summary. The index of *MarCat* (not shown) is less comprehensive, thus have fever filtering options.

the NCBI button. For *MarCat*, the marine metagenomics gene catalogue, target information can be obtained from other databases such as UniProt, InterPro and Brenda. Output from the BLAST search can be downloaded in FASTA, XML files or tab-separated files (TSV) format.

Download

The download section accommodates the contextual databases, individual genome and metagenome related sequences, and BLAST databases. Contextual information for all entries/samples exists as TSV and XML files which are available for the current and prior release versions. In *MarRef* and *MarDB* sequences of individual genomes are grouped according to their names and contained in separate folders where assembly, nucleotide and protein data are accessible as FASTA files. A general feature format file is also provided for each genome. The full collection of contigs/scaffolds, nucleotide and protein sequences for the BLAST databases are accessible in the same directory tree and may also be downloaded freely. For samples in the *MarCat* database, all predicted 16 S sequences and assembled contigs in FASTA format can be downloaded. In addition, an output file from META-pipe containing all annotated contigs in the sample is also provided together with the individual predicted genes and protein sequences in FASTA format.

ONGOING DEVELOPMENTS

The ongoing activities can be classified into three broad categories: (i) acquisition of data, (ii) ontologies and CV and (iii) linked data and interoperability

Acquisition of sequence and contextual data

The collection of data from publicly available resources will continue. However, due to increasing amount of genomic and metagenomic sequence- and metadata, development of automatic and semi-automatic import tools that generate metadata for the curation database will be improved in order to build more efficient import pipelines. In this first version of the *MarRef* and *MarDB* databases, only prokaryote genomes have been included. In the future, we aim to include virus, eukaryote microbial genomes and transcriptome data. In addition, we aim to include metatranscriptomics data to enhance the quality of the *MarCat*.

Ontologies and controlled vocabularies

To enhance the curation efficiency and to provide a better reliability of the datasets, the number of metadata fields will be increased with ontologies and CV. This effort will not only streamline the manual curation, but also provide data robustness and easier aggregation and analysis. For *MarCat* we intend to include metadata fields for the provenance of analysis according to the recommendation by Hoopen *et al.* (24), which includes metagenomics analysis metadata such as filtering, assembly, taxonomy, gene prediction and functional assignment.

Linked data and interoperability

In order to expose and share the curated data, we are currently working together with EMBL-EBI to link the MAR database records to the BioSample and INSDC databases. To improve data interoperability, we intend to implement schema.org markup, so that MMP websites and services contain more structured information. This structured information will make it easier for the end user to discover, collate and analyze our data. We also aim to improve better systems for downloading single records or multiple records selected by searching or filtering of the datasets.

These improvements will be implemented in the next version of the databases scheduled for March 2018.

The functionality of the databases has been tested using different platforms and web browser, such as Safari, Firefox, Chrome and Edge, without any problems. We welcome user feedback by email to mmp@uit.no.

ACKNOWLEDGEMENTS

We thank all the experts who contributed to the MMP database annotations from their field of expertise, especially ELIXIR EXCELERATE WP6 partners. The MMP is part of the Norwegian node of the European ELIXIR project.

FUNDING

ELIXIR EXCELERATE funded by the European Commission within the Research Infrastructures programme of Horizon 2020 [676559]; Research Council of Norway [208481]; UiT The Arctic University of Norway. Funding for open access charge: UiT The Arctic University of Norway.
Conflict of interest statement. None declared.

REFERENCES

1. Arrigo,K.R. (2005) Marine microorganisms and global nutrient cycles. *Nature*, **347**, 349–455.
2. Creer,S., Deiner,K., Frey,S., Porazinska,D., Pierre Taberlet,P., Thomas,W.K., Potter,C. and Bik,H.M. (2016) The ecologist's field guide to sequence-based identification of biodiversity. *Methods Ecol. Evol.*, **7**, 1008–1018
3. Zettler,L.A., Artigas,L.A., Baross,J., Loka Bharathi,P.A., Boetius,A., Chandramohan,D., Herndl,G., Kogure,K., Neal,P., Pedrós-Alió,C. *et al.* (2010) A global census of marine microbes. In: McIntyre,A (ed). *Life in the World's Oceans: Diversity, Distribution, and Abundance*. Wiley-Blackwell, Oxford, pp. 233–245.
4. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae. *Science*, **269**, 496–512.
5. Ishoey,T., Woyke,T., Stepanauskas,R., Novotny,M. and Lasken,R.S. (2008) Genomic sequencing of single microbial cells from environmental samples. *Curr. Opin. Microbiol.*, **11**, 198–204.
6. Eloe-Fadrosh,E.A., Paez-Espino,D., Jarett,J., Dunfield,P.F., Hedlund,B.P., Dekas,A.E., Grasby,S.E., Brady,A.L., Dong,H., Briggs,B.R. *et al.* (2016) Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. *Nat. Commun.*, **7**, 10476.
7. Hedlund,B.P., Dodsworth,J.A., Murugapiran,S.K., Rinke,C. and Woyke,T. (2014) Impact of single-cell genomics and metagenomics on the emerging view of extremophile 'microbial dark matter'. *Extremophiles*, **18**, 865–875.
8. Gross,L. (2007) Untapped bounty: sampling the seas to survey microbial biodiversity. *PLoS Biol.*, **5**, e85.

9. Laursen,L. (2011) Spain's ship comes. *Nature*, **475**, 16–17.
10. Rusch,D.B., Halpern,A.L., Sutton,G., Heidelberg,K.B., Williamson,S., Yooseph,S., Wu,D., Eisen,J.A., Hoffman,J.M., Remington,K. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.*, **5**, e77.
11. Sunagawa,S., Karsenti,E., Bowler,C. and Bork,P. (2015) Computational eco-systems biology in Tara Oceans: translating data into knowledge. *Mol. Syst. Biol.*, **11**, 809.
12. Wommack,K.E., Bhavsar,J., Polson,S.W., Chen,J., Dumas,M., Srinivasiah,S., Furman,M., Jamindar,S. and Nasko,D.J. (2012) VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Stand. Genomic Sci.*, **6**, 427–439.
13. Mitchell,A., Bucchini,F., Cochrane,G., Denise,H., ten Hoopen,P., Fraser,M., Pesseat,S., Potter,S., Scheremetjew,M., Sterk,P. *et al.* (2016) EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.*, **44**, D595–D603.
14. Chen,I.A., Markowitz,V.M., Chu,K., Palaniappan,K., Szeto,E., Pillay,M., Ratner,A., Huang,J., Andersen,E., Huntemann,M. *et al.* (2017) IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.*, **45**, D507–D516.
15. Meyer,F., Paarmann,D., D'Souza,M., Olson,R., Glass,E.M., Kubal,M., Paczian,T., Rodriguez,A., Stevens,R., Wilke,A. *et al.* (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.
16. Glöckner,F.O. and Joint,I. (2010) Marine microbial genomics in Europe: current status and perspectives. *Microb. Biotechnol.*, **3**, 523–530.
17. Mineta,K. and Gojobori,T. (2016) Databases of the marine metagenomics. *Gene*, **576**, 724–728.
18. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
19. Mukherjee,S., Stamatis,D., Bertsch,J., Ovchinnikova,G., Verezemska,O., Isbandi,M., Thomas,A.D., Ali,R., Sharma,K., Kyrpides,N.C. *et al.* (2017) Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Res.*, **45**, D446–D456.
20. Wattam,A.R., Davis,J.J., Assaf,R., Boisvert,S., Brettin,T., Bun,C., Conrad,N., Dietrich,E.M., Disz,T., Gabbard,J.L. *et al.* (2017) Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res.*, **45**, D535–D542.
21. Vallenet,D., Calteau,A., Cruveiller,S., Gachet,M., Lajus,A., Josso,A., Mercier,J., Renaux,A., Rollin,J., Rouy,Z. *et al.* (2017) MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. *Nucleic Acids Res.*, **45**, D517–D528.
22. Kottmann,R., Kostadinov,I., Duhaime,M.B., Buttigieg,P.L., Yilmaz,P., Hankeln,W., Waldmann,J. and Glöckner,F.O. (2010) Megx.net: integrated database resource for marine ecological genomics. *Nucleic Acids Res.*, **38**, D391–D395.
23. Costello,M.J., Bouchet,P., Boxshall,G., Fauchald,K., Gordon,D., Bert,W., Hoeksema,B.W., Poore,G.C.B., van Soest,R.W.M., Stohr,S. *et al.* (2013) Global coordination and standardisation in marine biodiversity through the World Register of Marine Species (WoRMS) and related databases. *PLoS One*, **8**, e51629.
24. Hoopen,P.T., 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.
25. Angiuoli,S.V., Gussman,A., Klimke,W., Cochrane,G., Field,D., Garrity,G., Kodira,C.D., Kyrpides,N., Madupu,R., Markowitz,V. *et al.* (2008) Toward an online repository of standard operating procedures (SOPs) for (meta)genomic annotation. *OMICS*, **12**, 137–141.
26. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
27. Li,D., Liu,C.M., Luo,R., Sadakane,K. and Lam,T.W. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, **31**, 1674–1676.
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 metagemome interpretation—a benchmark of metagenomics software. *Nat. Methods*, doi:10.1038/nmeth.4458.
29. Noguchi,H., Taniguchi,T. and Itoh,T. (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes, *DNA Res.*, **15**, 387–396.
30. Suzek,B.E., Wang,Y., Yuqi, Huang,H., McGarvey,P.B., Wu,C.H. and the UniProt Consortium (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932.
31. Claudel-Renard,C., Chevalet,C., Faraut,T. and Kahn,D. (2008) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.*, **31**, 6633–6639.
32. Jones,P., Binns,D., Chang,H-Y, Frase,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
33. Satinsky,B.M., Zielinski,B.L., Doherty,M., Smith,C.B., Sharma,S., Paul,J.H., Crump,B.C. and Moran,M.A. (2014) The Amazon continuum dataset: quantitative metagenomic and metatranscriptomic inventories of the Amazon River plume, June 2010. *Microbiome*. **2**, 17.
34. Gilbert,J.A., Meyer,F., Schriml,L., Joint,I.R., Mühling,M. and Field,D. (2010) Metagenomes and metatranscriptomes from the L4 long-term coastal monitoring station in the western english channel. *Stand. Genomic Sci.*, **3**, 183–193.
35. Johnson,M., Zaretskaya,I., Raytselis,Y., Merezhuk,Y., McGinnis,S. and Madden,T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res.*, **36**, W5–W9.

Paper 2

T. Klemetsen *et al.*, "The MAR databases: A manually curated resource for marine microbial genomics and metagenomics," Manuscript for: *Nucleic Acids Res.*

The MAR databases: A manually curated resource for marine microbial genomics and metagenomics

Terje Klemetsen[1], Juan Fu[1], Alexander Agafonov[2], Sudhagar Veerabadran Balasundaram[1], Espen Robertsen[1], Nils Peder Willassen[1]*

[1] Center for Bioinformatics, Faculty of science and technology, UiT The Arctic University of Norway, PO Box 6050 Langnes, Tromsø N-9037, Norway
[2] Department of Information Technology, UiT The Arctic University of Norway, PO Box 6050 Langnes, Tromsø N-9037, Norway

* To whom correspondence should be addressed. Tel: +47 77644651; Email: nils-peder.willassen@uit.no

ABSTRACT

The MAR databases (https://mmp.sfb.uit.no/databases/) are collections of marine microbial contextual (metadata) and sequence databases which are richly annotated and manually curated. The MAR databases have been continuously updated since the first release in 2017. In addition to MarRef and MarDB, two new databases MarFun, a database devoted to marine fungi genomes, and SalDB, a database for salmonid-associated bacteria, have been released. The number of metadata fields (attributes) has been increased to meet the community needs and includes attributes for predicted secondary metabolite biosynthetic gene clusters and assessment of genome quality. The MAR databases is a part of the Marine Metagenomics Portal (https://mmp.sfb.uit.no/) provides a user-friendly web interface that lets the visitors browse, filter, and search in the contextual databases and perform BLAST searches against the corresponding sequence databases. All contextual and sequence databases are freely accessible and downloadable from https://mmp.sfb.uit.no/downloads/.

INTRODUCTION

Microorganisms contribute significantly to the balance and resilience of marine ecosystems via their roles in the biogeochemical cycling of elements e.g. cycling of carbon, nitrogen, phosphorus, and trace elements (1). Microorganisms inhabit a variety of marine environments, ranging from the saltwater of seas and oceans to brackish waters of the coastal estuary, and from hydrothermal vents to sea-ice brines and can appear as floating or free-swimming planktonic cells or in multi-species biofilms on organic or inorganic surfaces. They may also enter internal tissues of marine plants and animals; only a small portion is pathogenic, but certain bacteria can establish highly symbiotic relationships with their specific host organisms (2). The various habitats and the extensive amount of microorganisms (estimated to 6.6× 1029cells in the sea) results in an astronomical genetic resource (3). To access the marine genetic resource, Whole Genome Shotgun (WGS) sequencing of cultured microorganisms has been the technology of choice. However, since the majority of microorganisms remain uncultured due to lack of ability to supply required growth conditions, the major part of the genetic resources have been unexplored (4). Nonetheless, recent technological developments have opened the way for exploring the unculturable or “the microbial dark matter” and can now be assessed using metagenomic or single-cell sequencing technologies (5, 6). Metagenome Assembled Genomes (MAGs), reconstructed from metagenomes and Single-cell Amplification Genomes (SAGs) approaches have not only proven particularly useful for identifying novel taxa and phylogenetic groups but increased the collection of available marine microbial genomes for research and innovation (7).

The number of prokaryotic (archaea and bacteria) genomes have doubled in the last three years from 103 000 in 2017 to nearly 260 000 in 2020, in addition, approximately 12,000 eukaryotic microbial (fungi and protists) genomes are now listed at the National Center for Biotechnology Information (NCBI) Genome databases (https://www.ncbi.nlm.nih.gov/genome) (8). In addition to the NCBI genome databases, there are a number of general microbial databases e.g. Prokaryotic RefSeq Genomes(9), Genomes OnLine Database (GOLD) (10), Pathosystems Resource Integration Center (PATRIC) (11), Integrated Microbial Genomes and Microbiomes (IMG/M) (12), which contain marine microbial genomes. However, due to the lack of curated metadata, the use of controlled vocabularies (CV) and ontologies makes it not only difficult to identify marine records of interest but also to compare contextual and sequence data within and/or between different databases.

When the first version of the MAR databases (https://mmp.sfb.uit.no/databases/): MarRef and MarDB were released in 2017 (13), the aim was to provide manually curated, sustainable, and highly accurate data resources for the marine microbial community which followed the FAIR (Findable, Accessible, Interoperable and Reusable) principles and community standards (14). The MAR databases have been substantially updated with new records and developed further by the inclusion of new attributes e.g. attributes for genome quality assessment, predicted secondary metabolite biosynthetic gene clusters, the inclusion of Evidence and Conclusion Ontology (ECO) (15), and two new databases – SalDB and MarFun have been launched. The update and development will be presented in the following sections.

OVERVIEW OF THE RESOURCES

Data collection and processing

The workflow for generation of the MAR contextual and sequence (BLAST) databases: MarRef, MarDB, SalDB, and MarFun are shown in Figure 1. The MAR sequence databases are based on the non-redundant genome and metagenome datasets obtained from the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/) (16) and European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena) (17).

While MarRef is a reference database for finished and complete marine prokaryotic genomes, MarDB includes all in-complete sequenced marine prokaryotic genomes regardless of the level of completeness. The new database SalDB, which was established from a request from the aquaculture industry, represents a database of all bacteria known to be associated with the salmonid fish species. The last database, MarFun is devoted to marine fungi genomes and their corresponding metadata. The MAR contextual databases are built by compiling data from a number of publicly available primary data resources including sequence, taxonomy, and literature databases in a semi-automatic

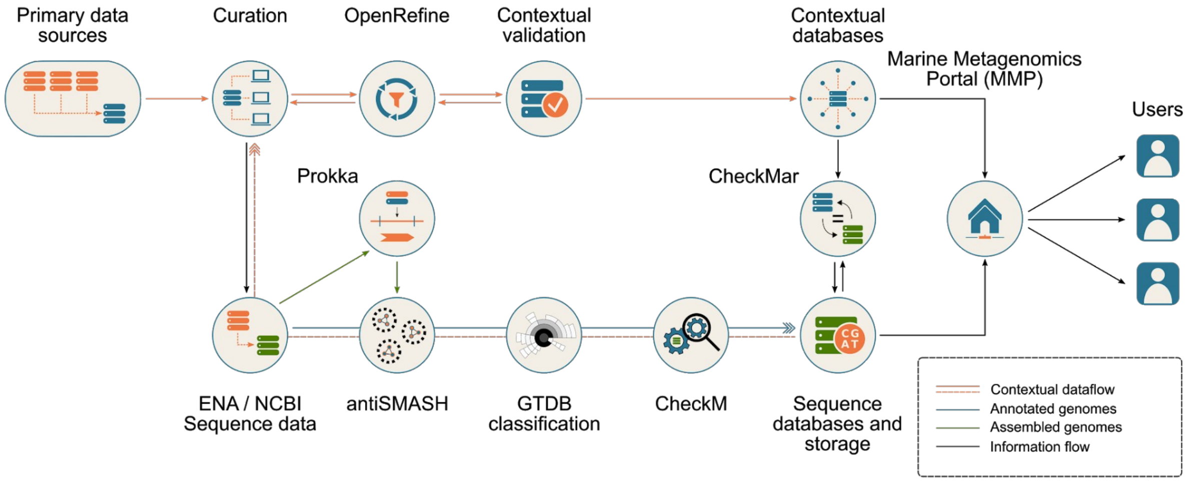


Figure 1. A brief overview describing steps in the curation and tool workflows for the implementation of new entries in the MAR databases.

fashion. Other databases or resources such as bacterial diversity and culture collections databases, web mapping services, and ontology databases were used extensively for the curation of metadata.

The MAR genome databases are bi-annually updated and the current versions of MarRef, MarDB, SalDB, and MarFun and contain 970, 13237, 348, and 28 records, respectively. While MarRef has a moderate growth from the first version (from 612 to 970 records), MarDB has grown substantially - more than threefold, from 3726 to more than 13000 records. WGS sequenced genomes dominate the records in MarRef, while the main growth in MarDB derives mainly from MAGs and SAGs as shown in Table 1. While WGS genomes were dominating the first version of MarDB, the MAGs are now the technology of choice for the generation of marine prokaryotic genomes. For MarFun, which so far contains 28 incomplete genomes, the WGSs is the preferred technology. SalDB contains nearly 350 genomes of bacteria associated with salmonid fish species.

With the latest update of MarRef, MarDB, and SalDB we implemented the GTDB taxonomy (18) as an addition to the NCBI taxonomy. The MAR databases use the latter taxonomy as is when admitting new entries, but generates the GTDB taxonomy based on the sequenced data constituting the entry. When compared, as shown in Figure 2, the taxonomies illustrate a considerable difference in our largest database MarDB. The number of entries with classification at the genus level is currently 6984 (52.76%) with the NCBI taxonomy and 12126 (91.61%) when based on the GTDB taxonomy. Unique taxonomic diversity in the MarDB as resulting from GTDB mounts to 1018 taxonomically named at genus level and additionally 1347 with names from reference genomes. As an example of the latter case, the reference genomes MGIIb-O5 and MGIIb-O3, both Archaeal genomes, obtained the most assignments with 59 and 58 entries respectively. With MarRef we registered 341 named NCBI taxa at genus level which constitutes the whole database. GTDB gave 379 uniquely named at genus level and 35 with names from reference genomes – inflating the diversity of GTDB with 73 additional unique taxa compared to the NCBI taxonomy at the same level.

CONTEXTUAL DATABASES

Contextual information related to records in NCBI and ENA is highly important and determines if the data can be linked to the marine environment and included in the MAR databases. As previously described, the MAR definition of a “marine microbial biome” is broad and includes not only open-ocean and unprotected coastal habitats, as defined in the ENVO “marine biome” but also tidal affected sampling sites like brackish waters, rocky shorelines, and sandy beaches (13). The collection of publicly available contextual data is therefore constantly under evaluation and development to ensure sufficiently and descriptive information is collected.

At the point of publication, the MAR genome databases support 124 attributes in the MarRef, MarDB, and SalDB databases and 127 in MarFun. In the current version, 23 new attributes have been introduced - 10 assembly attributes such as binning, binning version, estimated completeness, and contamination, stain heterogeneity, QS (quality score), mapping, mapping version, quality assessment, and quality assessment version and 8 secondary metabolite attributes such as predicted antiSMASH types, antiSMASH clusters, CHEBI ID, CHEBI name, compound name, Uniprot ID, Uniprot description. An “Analysis project type” attribute has been included to ease the access to MAGs, SAGs, and WGSs. Additionally, three attributes from the GTDB-Tk classifier have been implemented to hold the taxonomic lineages and related information regarding the closest reference genome and average nucleotide identity. For MarFun, three fungi specific attributes were included: ploidy, propagation, and ITsoneDB_ID.

Table 1. The total number of WGS, MAG, and SAG entries in the MAR databases.

Recovery type	MarRef v5	MarDB v5	SalDB v2	MarFun v2
WGS	953	5034	348	27
MAG	17	7180	0	1
SAG	0	1023	0	0
Total	970	13237	348	28

The MAR databases are intended to provide, among others, contextual data related to sampling, sequencing, assembly, annotation and taxonomy. A varying amount of such information is submitted alongside the genomic sequence data by authors/owners of given projects and stored in various formats in the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org/) (19). The collection of data from these sites have become mostly automated by accessing and appending information to the MAR databases. This includes also BioSample and BioProject contextual data, such as sequencing equipment and methods, sequencing coverage, assembly programs and versions, complete taxonomy lineages, identifiers, and accession numbers. As RefSeq annotated genomes are our primary source of sequence data, the annotation statistics and provenance data are stored along with numeric gene/RNA data (9).

As in the initial version of the MAR databases, we continue to utilize Prokka (20) to perform fast genome annotations as a temporary measure in cases where genome sequence data has no RefSeq annotations. Provenance data and statistics from Prokka annotations are subsequently stored as

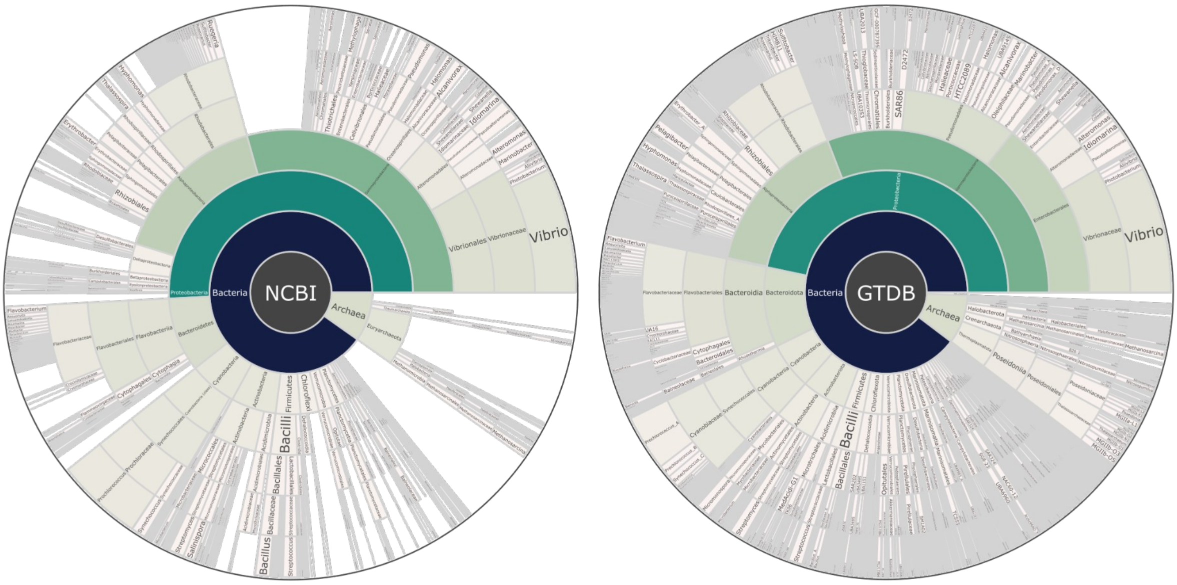


Figure 2. Taxonomic diversity of MarDB as illustrated by sunburst plots from inherent NCBI classification (left) and generated GTDB classification (right). Ring sectors from the center represent each of the taxonomic levels from the kingdom, phylum, class, order, family, and genus. Individual sectors represent the given taxonomic node and reflect its count in the MarDB database. White/unfilled areas represent missing or unclassified taxa, candidatus and candidate division classifications.

contextual data post-analysis. All accepted genomes in the MAR genome databases are scanned for secondary metabolites clusters using antiSMASH v.4 (21) and their quality is assessed by CheckM (22) to estimate genome completeness and contamination. Both antiSMASH types and clusters are appended to the entries' contextual data structure, thus making these additional searchable features. Equally, the completion, contamination, strain heterogeneity, and calculated quality score (QS) are included from each genome assessed by CheckM.

Curation of the contextual data

For MarRef, 43 attributes related to sample isolation, host/pathogenicity, phenotype have been manually curated by adding/correcting metadata from multiple resources such as literature, culture collection databases as described in the initial paper (13). Multi-source metadata has to be unified in order to resolve inconsistencies and erroneous information. In case of conflicts among the sources of metadata, literature is prioritized to culture collections, followed by secondary/specialized databases and BioSample. While MarRef is thoroughly curated, MarDB is only partly curated, mainly focused on the isolation information: environmental and geographic location ontologies and host names. Ontology terms are assigned based on experimental data from the literature, such as Environmental ontologies (ENVO) (23), geographic location (Gazetteer) (https://www.ebi.ac.uk/ols/ontologies/gaz). Since automatic mapping is prone to error (24), the environmental ontology terms are determined manually by the curators based on the isolation source, geographic location, or other information in the literature. Mapping of GAZ terms on the attributes geo_loc_name_GAZ and geo_loc_name_GAZ_ENVO are based on the most accurate location as described or inferred from the sample authors. The attribute “analysis project type” in MarRef, MarDB, SalDB, and MarFun has been manually assessed to determine MAGs, SAGs, and WGSs based upon information from ENA and literature.

In the new version of the MAR genome databases, eight Evidence and Conclusion Ontology (ECO) terms which comprise structured controlled vocabularies to enable the description of experimental, computational, and other evidence types to support the assertion captured by databases have been included.

MAR SEQUENCE AND OTHER DATA RESOURCES

Continuous procurement and expansion of the MAR databases consequently provide a unique resource in terms of sequence data. To leverage this resource to its fullest, we not only store the genomic sequences but continue to develop and maintain various implementations of tool-specific databases to support marine genomic and metagenomic research.

BLAST

Online BLAST services (https://mmp.sfb.uit.no/blast/) are available for all MAR databases (MarRef, MarDB, SalDB, and MarFun). Additionally, as of version of MarRef/MarDB and version 1 of SalDB, the different databases have also been split into “project analysis type”: MAG, SAG and WGS, to distinguish between the inherent quality differences, coverage profiles and other characteristic properties of these individual recovery methods.

Other sequence resources

Sequence data pertaining to MarRef, MarDB, and SalDB has been integrated with the taxonomic classification tool Kaiju, commonly used to assign taxonomy to metagenomic samples (25). This integration is part of the official Kaiju repository and is maintained by the MMP-team (https://github.com/bioinformatics-centre/kaiju). When installing and building Kaiju, it is possible to choose marine sequences from either MarRef, MarDB, and SalDB or a combination of these, which is automatically downloaded, formatted, and ready to use with Kaiju. We also offer preformatted databases for the taxonomic k-mer based classification tool Kraken (26). We have also created a database from predicted 16S rRNA CDS annotations of the records in MarRef, MarDB named SILVA MAR, which is readily available to use with the rRNA classification software MapSeq (27). All sequence databases and resources can be downloaded from the MAR Download page (https://mmp.sfb.uit.no/downloads/).

Interactive visualization of metadata

To acquire a comprehensive and summarized overview of the different quality aspects of the Mar databases, an interactive visualization application has been developed and embedded in the Marine Metagenomics Portal (https://mmp.sfb.uit.no/metadata/). Here, the numeric metadata of all genomes in MarRef, MarDB, and SalDB can be visualized and filtered according to available numeric metadata, including quality score, completeness, contamination, genome length, and number of contigs. Users are able to refine the default view using specific database selections, sequencing types, quality categories, and other numerical filters, effectively simplifying subsetting of database entries based on user-specified criteria.

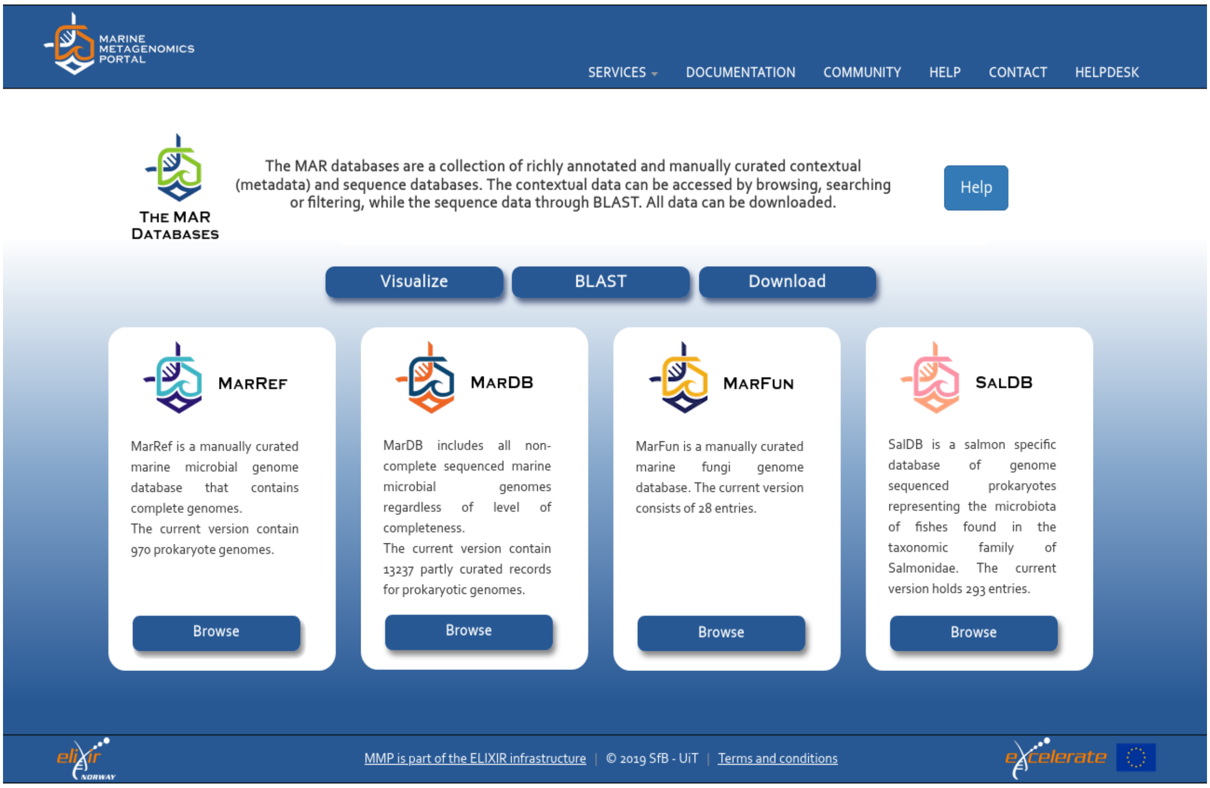
Validation of implemented data

In the final steps leading up to the publication of an upcoming database version, several measures have been implemented to ensure consistency in the MAR databases. A validation step is present during the contextual data transition into JSON/XML format and certifies the requirements for numerous attributes providing compatibility with the MMP website functions. The contextual data is the primary content definition of the MAR databases and determines what sequence data it holds. To manage discrepancies between the contextual data, sequence data, and analysis data like antiSMASH, we implemented a reporting module, CheckMar, to verify the consistency between these data types. The CheckMar main functions involve the identification of duplicates, excess or missing sequence and analysis data, and the presence of individual files within each entry. CheckMar iterates on our backend servers offered through the MMP portal on a daily basis. Reports are written to Google sheets automatically, making it a relatively simple task for any curator to manually keep track of any discrepancies and updates.

INFRASTRUCTURE IMPROVEMENTS

The MAR databases are incorporated into the Marine Metagenomics Portal (MMP) (https://mmp.sfb.uit.no/) (Figure 3). Central to the functionality of the databases are browsing metadata, BLAST searching, and auxiliary applications adapted to the database content. Since the initial release, the MarRef and MarDB databases have been updated bi-annually with the latest published data (Table 1). This increase of metadata entries and sequence data has in terms lead to a demand for improved infrastructure, in particular for MarDB. To reduce lengthy load times during website use and search filtering, the MMP site currently processes JSON, replacing the XML format initially implemented as a data storage format for metadata. While querying data the user accesses the JSON content as a compressed file, significantly reducing time consumption previously experienced by the XML format. Subsequent testing of the infrastructure improvements revealed loading times on site were cut by more than half, but were still dependent on the web browser in use and network connection.

We have continuously analyzed all collected marine genomes for secondary metabolites using antiSMASH v.4 (21). The analysis output is stored in HTML format, as specified by the antiSMASH developers, easily accessible via a link from the entry summary page of the MAR databases. All main analysis results are also accessible and searchable in the metadata, this includes antiSMASH types and clusters to be queried while browsing genomes. Additional filtering functionality has been implemented for the attributes antiSMASH types and ChEBI Name to better facilitate finding metadata of interest.



The MAR databases are a collection of richly annotated and manually curated contextual (metadata) and sequence databases. The contextual data can be accessed by browsing, searching or filtering, while the sequence data through BLAST. All data can be downloaded.

MarRef is a manually curated marine microbial genome database that contains complete genomes. The current version contain 970 prokaryote genomes.

MarDB includes all non-complete sequenced marine microbial genomes regardless of level of completeness. The current version contain 13237 partly curated records for prokaryotic genomes.

MarFun is a manually curated marine fungi genome database. The current version consists of 28 entries.

SalDB is a salmon specific database of genome sequenced prokaryotes representing the microbiota of fishes found in the taxonomic family of Salmonidae. The current version holds 293 entries.

MMP is part of the ELIXIR infrastructure | © 2019 SfB - UiT | Terms and conditions

Figure 3. The MAR database web-page at the Marine Metagenomics Portal (MMP) site hosts all 4 databases currently operational. MarFun represents the latest implementation of marine fungal genomes, the first database at MMP holding Eukaryotic genome data.

Since spring 2018 MAGs verified to originate from marine metagenomic samples have been introduced to the MAR databases. Due to the contemporary advancement of MAGs, there are different opinions about their validity as representative genomes of prokaryotes, thus we have enabled MAGs to be ignored when browsing the databases by applying the new filtering attribute Analysis project type. This enables the user to filter based on the status origin of the sample as MAG, WGS, or SAG.

The MAR databases and their metadata contain multiple links to sites like ENA (http://www.ebi.ac.uk/ena) (17), EMBL-EBI (28), Pubmed (https://www.ncbi.nlm.nih.gov/pubmed/), INSDC (http://www.insdc.org/) (19), RefSeq (9), BacDive (29) and Silva (30). To ensure these links are upheld and active we implemented URI resolution identifiers from these sites that are registered at Identifiers.org. Sites providing links to MMP include ENA and The World Register of Marine Species (WoRMS) (31).

ACKNOWLEDGMENTS

We thank all the experts who contributed to the Marine Metagenomics Portal database annotations from their field of expertise, especially ELIXIR EXCELERATE partners. The MMP and the MAR databases are a part of the ELIXIR infrastructure.

FUNDING

This work was supported by the ELIXIR EXCELERATE project funded by the European Commission within the Research Infrastructures programme of Horizon 2020 (grant agreement number 676559), the ELIXIR2 project funded by the Research Council of Norway (grant agreement number 208481) and UiT The Arctic University of Norway.

REFERENCES

1. York,A. (2018) Environmental microbiology: Marine biogeochemical cycles in a changing world. *Nat. Rev. Microbiol.*, **16**, 259.

2. Bolhuis,H. and Cretoiu,M.S. (2016) What is so special about marine microorganisms? Introduction to the marine microbiome-from diversity to biotechnological potential. In *The Marine Microbiome: An Untapped Source of Biodiversity and Biotechnological Potential*. Springer International Publishing, pp. 3–20.

3. Overmann,J. and Lepleux,C. (2016) Marine bacteria and archaea: Diversity, adaptations, and culturability. In *The Marine Microbiome: An Untapped Source of Biodiversity and Biotechnological Potential*. Springer International Publishing, pp. 21–55.

4. Garza,D.R. and Dutilh,B.E. (2015) From cultured to uncultured genome sequences: Metagenomics and modeling microbial ecosystems. *Cell. Mol. Life Sci.*, **72**, 4287–4308.

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. Rinke,C., Schwientek,P., Sczyrba,A., Ivanova,N.N., Anderson,I.J., Cheng,J.F., Darling,A., Malfatti,S., Swan,B.K., Gies,E.A., *et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, **499**, 431–437.

7. Alneberg,J., Karlsson,C.M.G., Divne,A.M., Bergin,C., Homa,F., Lindh,M. V., Hugerth,L.W., Ettema,T.J.G., Bertilsson,S., Andersson,A.F., *et al.* (2018) Genomes from uncultivated prokaryotes: A comparison of metagenome-assembled and single-amplified genomes 06 Biological Sciences 0604 Genetics. *Microbiome*, **6**, 1–14.

8. Sayers,E.W., Agarwala,R., Bolton,E.E., Brister,J.R., Canese,K., Clark,K., Connor,R., Fiorini,N., Funk,K., Hefferon,T., *et al.* (2019) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **47**, D23–D28.

9. 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.

10. Mukherjee,S., Stamatis,D., Bertsch,J., Ovchinnikova,G., Katta,H.Y., Mojica,A., Chen,I.M.A., Kyrpides,N.C. and Reddy,T.B.K. (2019) Genomes OnLine database (GOLD) v.7: Updates and new features. *Nucleic Acids Res.*, **47**, D649–D659.

11. Wattam,A.R., Davis,J.J., Assaf,R., Boisvert,S., Brettin,T., Bun,C., Conrad,N., Dietrich,E.M., Disz,T., Gabbard,J.L., *et al.* (2017) Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.*, **45**, D535–D542.

12. Chen,I.M.A., Chu,K., Palaniappan,K., Pillay,M., Ratner,A., Huang,J., Huntemann,M., Varghese,N., White,J.R., Seshadri,R., *et al.* (2019) IMG/M v.5.0: An integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.*, **47**, D666–D677.

13. 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.

14. Wilkinson,M.D., Dumontier,M., Aalbersberg,Ij.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.W., da Silva Santos,L.B., Bourne,P.E., *et al.* (2016) Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 1–9.

15. Chibucos,M.C., Siegele,D.A., Hu,J.C. and Giglio,M. (2017) The evidence and conclusion ontology (ECO): Supporting GO annotations. In *Methods in Molecular Biology*. Humana Press Inc., Vol. 1446, pp. 245–259.

16. Sayers,E.W., Beck,J., Brister,J.R., Bolton,E.E., Canese,K., Comeau,D.C., Funk,K., Ketter,A., Kim,S., Kimchi,A., *et al.* (2020) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **48**, D9–D16.

17. Amid,C., Alako,B.T.F., Balavenkataraman Kadhirvelu,V., Burdett,T., Burgin,J., Fan,J., Harrison,P.W., Holt,S., Hussein,A., Ivanov,E., *et al.* (2020) The European Nucleotide Archive in 2019. *Nucleic Acids Res.*, **48**, D70–D76.

18. Parks,D.H., Chuvochina,M., Waite,D.W., Rinke,C., Skarshewski,A., Chaumeil,P.-A. and Hugenholtz,P. (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, **36**, 996–1004.

19. Karsch-Mizrachi,I., Takagi,T. and Cochrane,G. (2018) The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, **46**, D48–D51.

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

21. Blin,K., Wolf,T., Chevrette,M.G., Lu,X., Schwalen,C.J., Kautsar,S.A., Suarez Duran,H.G., De Los Santos,E.L.C., Kim,H.U., Nave,M., *et al.* (2017) AntiSMASH 4.0 - improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.*, **45**, W36–W41.

22. 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.

23. Buttigieg,P.L., Morrison,N., Smith,B., Mungall,C.J. and Lewis,S.E. (2013) The environment ontology: Contextualising biological and biomedical entities. *J. Biomed. Semantics*, **4**, 1–9.

24. Buttigieg,P.L., Pafilis,E., Lewis,S.E., Schildhauer,M.P., Walls,R.L. and Mungall,C.J. (2016) The environment ontology in 2016: Bridging domains with increased scope, semantic density, and interoperation. *J. Biomed. Semantics*, **7**, 1–12.

25. Menzel,P., Ng,K.L. and Krogh,A. (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.*, **7**, 1–9.

26. Wood,D.E. and Salzberg,S.L. (2014) Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, 1–12.

27. 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.

28. Cook,C.E., Lopez,R., Stroe,O., Cochrane,G., Brooksbank,C., Birney,E. and Apweiler,R. (2019) The European Bioinformatics Institute in 2018: Tools, infrastructure and training. *Nucleic Acids Res.*, **47**, D15–D22.

29. Reimer,L.C., Vetcininova,A., Carbasse,J.S., Söhngen,C., Gleim,D., Ebeling,C. and Overmann,J. (2019) BacDive in 2019: Bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Res.*, **47**, D631–D636.

30. 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.

31. Costello,M.J., Bouchet,P., Boxshall,G., Fauchald,K., Gordon,D., Hoeksema,B.W., Poore,G.C.B., van Soest,R.W.M., Stöhr,S., Walter,T.C., *et al.* (2013) Global Coordination and Standardisation in Marine Biodiversity through the World Register of Marine Species (WoRMS) and Related Databases. *PLoS One*, **8**, e51629.

Paper 3

T. Klemetsen, Robertsen M. Espen, and Willassen P. Nils, “A substantial quality assessment of prokaryotic genomes in the MAR databases reveals an urgent need for submission quality control,” Manuscript for: Oxford University Press - *Bioinformatics*.

Bioinformatics, 2021, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

A substantial quality assessment of prokaryotic genomes in the MAR databases reveals an urgent need for submission quality control

Klemetsen Terje[1,*], Robertsen M. Espen[1] and Willassen P. Nils[1]

[1]Department of Chemistry, Center for Bioinformatics, UiT The Arctic University of Norway, Tromsø, Norway

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Awareness of quality in published genomes is critical for accurate taxonomic classification and functional assignment in bioinformatics. Cost-effective whole-genome and single-cell sequencing techniques and novel methods for binning metagenomic reads drive the INSDC databases into a rapid, unchecked expansion of genome assemblies. The genome quality can be validated through checklists and recommendations but remains to be integrated with public repositories. Genome completeness, contamination, quality score, contig size and fragmentation, rRNAs, and unique tRNAs are parameters considered in quality classifications. We evaluated the entries in the prokaryotic marine databases MarRef and MarDB for these parameters to gain insight into the quality of genome assemblies in public repositories and databases.

Results: The quality of genome assemblies is highly varied among public marine samples, with whole-genome sequencing as the chief contribution to high quality. Additional efforts on single amplified and metagenome-assembled genomes can provide finished and high-quality drafts but mainly distributed in the middle and low end of the genome quality scale. However, some public genomes fail to comply with the requirements of Low-quality drafts, particularly from excessive contamination. To avoid a quality loophole, we propose Very low-quality drafts as a label for genome assemblies not fit current quality categories and suggest improved reporting of quality in published genome assemblies.

Availability and Implementation: The Marine Metagenomics Portal (MMP) which was the source of this study can be found at https://mmp.sfb.uit.no/ and the interactive explorer for metadata is located at https://mmp-visualization.sfb.uit.no/mmp_interactive.

Contact: -
Supplementary information: Supplementary Tables 1, 2 and 3.

1 Introduction

The first finished or complete bacterial genome to be published was Haemophilus influenzae Rd Kw20, already more than 25 years ago (Fleischmann *et al.,* 1995). Since then, the number of prokaryotic genomes submitted to the International Nucleotide Sequence Database Collaboration, INSDC, including the DDBJ, ENA, and NCBI databases, has grown exponentially (Cochrane *et al.,* 2016). By June 2021, the number of prokaryotic genome assemblies in NCBI had reached more than

330.000 genomes. This overall growth is primarily due to advances in sequencing technologies. But also new bioinformatics approaches such as the assembly of genomes from metagenome samples and the development of single-cell sequencing technologies has contributed substantially to the growth.

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 sequence reads 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 may lead to genome misassemblies that warrant additional experimental and bioinformatics analyses to identify and correct before completion and deposition to public archives.

Technological developments have facilitated unprecedented access to the uncultured genomes or ”the microbial dark matter”, using either single-cell or metagenomic sequencing technologies. Although Single Amplification Genome (SAG) and Metagenome Assembled Genome (MAG) approaches have proven robust, several challenges are associated with each. Starting from a single-cell, SAG sequencing is demanding due to PCR artefacts, 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 (Kogawa *et al.,* 2018). Generation of MAGs, on the other hand, requires high sequencing depth and, ideally, a large number of samples with the same richness but different relative species abundance to identify and assemble identical bins. Besides, 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 (Meziti *et al.,* 2021).

While the quality of isolate WGS genomes have traditionally been evaluated using assembly statistics, such as contig and scaffolds lengths N50 and L50.N50 is defined as the sequence length of the shortest contig at 50% of the total assembly length, and L50 as the number of contigs/scaffolds whose summed length is N50 (Salzberg *et al.,* 2012). However, these statistics are less meaningful in the case of assessing the quality of MAGs and SAGs. Evaluation can usually be performed by identifying and counting universal Single Copy Genes (SCGs). These SCGs or ”marker genes” are found ubiquitously across bacterial and archaeal lineages and only once within a genome. Several lists of such SCGs exist and consist mainly of genes encoding for ribosomal proteins and other housekeeping genes (Rinke *et al.,* 2013). Using such lists, one can estimate the completeness and contamination of SAGs, MAGs and WGS draft assemblies. In short, completeness is the number of unique SCGs present divided by the number of expected SCGs in an assembly.

On the other hand, contamination is estimated by counting the number of SCGs present in multiple copies, as only one copy of each SCG is expected to be present per assembly. CheckM (Parks *et al.,* 2015), the most used software for assessing assembly completeness and contamination of prokaryotes, uses ubiquitous and SCGs specific to a genomic lineage within a reference tree. The lineage-specific marker sets determined for all nodes within the reference genome tree by identifying SCGs present in ≥97% of all descendant genomes. In terms of completeness and contamination, the quality of an assembly 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.

Table 1. Quality classification of genome sequences.

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. Presence of the 23S, 16S and 5S rRNA genes and at least 18 tRNAs. Assembly statistics report. Completeness[3] > 90% Contamination[4] < 5%
Near-Complete High-Quality Draft	Multiple fragments where gaps span repetitive regions. May include other quality measurements such as strain heterogenicity. Assembly statistics report. Completeness[3] > 90% Contamination[4] < 5%
Medium-Quality Draft	Many fragments with little to no review of assembly other than reporting of standard assembly statistics. Completeness ≥ 50% Contamination < 10%
Low-Quality Draft	Many fragments with little to no review of assembly other than reporting of standard assembly statistics. Completeness score < 50% Contamination < 10%
Very Low-Quality Draft	Many fragments with little to no review of assembly other than reporting of standard assembly statistics. Contamination ≥ 10%

[1] Q50 = Phred quality score of 50: the probability of one incorrect base call in 100,000 (99.999% base call accuracy).
[2] Assembly statistics, including but not restricted to total assembly length, number of chromosomes and plasmids, number of scaffolds and contigs, contig and scaffold N50, and maximum contig length.
[3] Completness score - the ratio of observed single-copy marker genes to total single-copy marker genes in the 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 the chosen marker gene set (%).

The Genomic Standards Consortium (GSC) developed two standards, the Minimum Information about a Single Amplified Genome (MISAG) and the Minimum Information about a Metagenome-Assembled Genome (MIMAG), for improving the reporting of assembly quality, estimates of genome completeness and contamination, and provide criteria for describing the quality of draft genomes (Bowers *et al.,* 2017). Based upon these standards, the GSC recommends classifying genomes as: “Finished”, “High-quality draft”, “Medium-quality draft”, and “Low-quality draft”. The finished genome is used for high quality manually curated genomes which consist of a validated, contiguous sequence per replicon without gaps or ambiguities with a consensus base calling error rate equivalent to Q50 or better. A High-quality draft is an assembly with a completeness score of > 90% and a contamination score < 5%. The assemblies in this class should also encode the 23S, 16S, and 5S rRNA genes and tRNAs for at least 18 of the 20 possible amino acids. A Medium quality draft is an assembly with an estimated completeness score ≥50% and a contamination score < 10%, while a Low-quality draft should be reported for assemblies with a completeness score <50% and with a

Table 2. Genome statistics (count, assembly, predicted rRNAs and tRNAs) of entries in the MarRef and MarDB databases, subdivided into the three methodologies WGS, MAG and SAG for obtaining genomes. (1) Database and recovery method, (2) Genome count, (3) Assembly length, (4) Number of contigs, (5) Contig length, (6) Assemblies with rRNA[1], (7) rRNA count (5S, 16S, 23S), (8) Assembly with tRNA[2], (9) Number of tRNA in assemblies, (10) Assemblies with ≥ 18 unique tRNA.

(1)	(2)	(3)			(4)		(5)		(6)	(7)	(8)	(9)	(10)
		Avg.	Min.	Max.	Avg.	Min.	Min.	Max.		Avg.		Avg.	
MarRef	**970**												
WGS	953	3985484	490889	9708663	2.0	1	1974	9673108	947	4.62, 4.34, 4.30	953	63.8	940
MAG	17	2601370	593370	9384773	1.3	1	1945	9373345	14	1.33, 1.26, 1.26	17	44.2	14
MarDB	**13237**												
WGS	5034	4386493	89553	1379946	153.1	1	20	8262658	4483	2.90, 2.76, 2.61	5016	59.8	4741
MAG	7180	2329670	10021	1441.639	189.7	1	102	4422561	1436	0.48, 0.47, 0.46	7167	29.7	3362
SAG	1023	1083967	8830	4854236	93.8	2	99	1189816	630	0.70, 0.80, 0.91	1007	23.7	427

[1] Genome assemblies containing a minimum of one 5S, 16S and 23S rRNA coding gene.

[2] Genome assemblies containing a minimum of one tRNA coding gene.

contamination score <10%. In recent reports, the terms "Near-complete" or "Near-complete high-quality" has been introduced to describe MAGs of high quality, which do not fulfil all criteria for being classified as "High-quality" drafts (Parks *et al.,* 2017, Almeida *et al.,* 2019). Near-complete is used for drafts with completeness > 90% and contamination < 5%, often in combination with other quality criteria such as strain heterogeneity.

Another commonly used metric for assessing genome assemblies is the quality score (QS), first introduced by Parks *et al.,* (2017), defined as QS = completeness − 5 × contamination. Draft assemblies with QS ≥ 50 are often considered as being of acceptable quality. The multiplication factor of the contamination means a trade-off, ensuring that partial draft assemblies can only contain minimal contamination.

As a part of increasing the usability of the MAR databases (Klemetsen *et al.,* 2018) and awareness of the quality of the genomes present in databases, we analyzed the genome assemblies for rRNAs, tRNAs, completeness and contamination. MarRef and MarDB are two curated marine prokaryote sequence and contextual databases. While MarRef is a manually curated reference database consisting of complete and finished genomes. MarDB is partially curated and contains draft genomes generated using shotgun sequencing, assembly of metagenomics reads, or draft genomes generated from single-cell amplified sequencing. Both databases consist of 120 different attribute fields, including taxonomy, sampling information, assembly and annotation. The assembly and annotation include assembly length, number of contigs, number of rRNAs (5S, 16S & 23S), number of tRNAs, and quality metrics such as completeness, contamination, strain heterogeneity and quality score (QS).

2 Methods

Definition

The Mar databases operate with entries in a flat format, each entry outlining a genome that defines the DNA molecules in an assembly dedicated to a given taxon. The definition of a finished (or complete) genome in this paper results from an assembly yielding no gaps in chromosomes and appertaining plasmids. Draft genome assemblies are complementary by having gaps between contigs or scaffolds and may present irregularities not representative for complete genomes, e.g. inaccurate size, or contaminated sequence data. The term 'genome assembly' may be abbreviated as 'assembly' dependent on context and applies to all genome conditions.

Datasets

The MarRef and MarDB databases (https://mmp.sfb.uit.no/) are compiled from publicly available databases, including NCBI (Schoch *et al.,* 2020) and ENA (Harrison *et al.,* 2020). The MarRef v5 consists of 970 finished genomes, out of which 953 entries are WGSs, and 17 are MAGs. The MarDB v5 contains 13237 draft genomes of various completeness; 5034 WGSs, 7180 MAGs, and 1023 SAGs.

Prediction of rRNA and tRNA

To achieve as high consistency as possible the 5S, 16S and 23S ribosomal RNAs (rRNAs) were predicted for all entries with the cmsearch function from INFERNAL v.1.1.2 (options -Z 1000 --cut_ga) using the Rfam covariance models of the bacterial rRNAs (Nawrocki and Eddy, 2013). Transfer RNAs (tRNAs) for each entry were predicted using tRNAscan-s.e. v.2.0 using the bacterial tRNA model (option -B) and default parameters (Chan *et al.,* 2019).

Quality assessment

Completeness and contamination for all entries in MarRef and MarDB were estimated with CheckM version 1.07 using the linage_wf workflow (Parks *et al.,* 2015). Quality score (QS) for each genome was calculated as the level of completeness − 5 × contamination (Parks *et al.,* 2017).

Classification of assemblies

The genome assemblies in the MAR databases were classified according to the MIMAG/MISAG standards for describing the quality of SAG and MAG; Finished, High-quality, Medium-quality and Low-quality draft (Bowers *et al.,* 2017). In addition, we also introduced the class Near Complete drafts (Parks *et al.,* 2017, Almeida *et al.,* 2019). This class is frequently used in the literature for MAGs, which do not fulfil the strict requirements for a High-quality draft, as shown in Table 1. However, to capture all entries in the MAR databases, we also needed to include a class called "Very-low quality" draft, which includes all entries with contamination scores of > 10%.

Reevaluation of selected Very-low quality drafts genomes

Selected cases classified as Very-low quality draft genomes were explored using CAT/BAT (Meijenfeldt *et al.,* 2019), running two different analyses. First, we assigned taxonomy using the subprograms "bin" and "add_names" using default parameters. Second, we ran the same analysis only with the parameter f=0.01, recommended to identify mis-binned contigs/contamination.

3 Results

To gain more insight into the quality of publicly available genomes, we performed a thorough analysis of genomes present in the MarRef and MarDB databases. The genomes were classified according to the recovery method, WGS, MAG and SAG, and whether the assembly status was regarded as complete (MarRef) or draft (MarDB) following the quality scheme shown in Table 1.

Assembly versions, lengths and contigs

The number of WGS, MAG and SAG in MarRef and MarDB for version 5 are listed in Table 2. While most assemblies in MarRef represented their initial version published (.1), 51 were published as version .2, and a further ten were version .3. Similarly, for MarDB, the number of entries having an assembly version .2, .3 or .4 amounted to 344, 29 and 11, respectively. The average assembly length of the finished assemblies in MarRef is higher for WGSs than MAGs, 3.96 Mbp, and 2.60 Mbp. However, the minimum and maximum assembly length for the WGSs and MAGs are relatively constant and vary from roughly 0.5 Mbp to 9.5 Mbp. The number of contigs (chromosomes/plasmids) in MarRef varies from 1 to 12, and the number of contigs is on average higher for WGSs than MAGS, 2.0 and 1.3, respectively. In MarDB, the draft WGSs assemblies vary in length from 0.9 Mbp to 16.4 Mbp, with an average assembly length of 4.38 Mbp and close to that of the average finished genome in MarRef (3.96 Mbp). For the MAGs in MarDB, the length varies from 0.10 to 11.44 Mbp, with an average assembly length of 2.33 Mbp. For SAGs, the average length is 1.08 Mbp, and the length varies from 0.089 to 4.85 Mbp. The number of contigs in MarDB varies from 1 to 8951, with an average number of 153.1, 189.7 and 93.8 for WGSs, MAGs and SAGs, respectively.

Predicted rRNAs and tRNAs

The number of rRNAs and tRNAs genes were predicted using INFERNAL and tRNAscan-s.e, respectively, as described in the method section and listed in Table 2. In MarRef, the average number of predicted

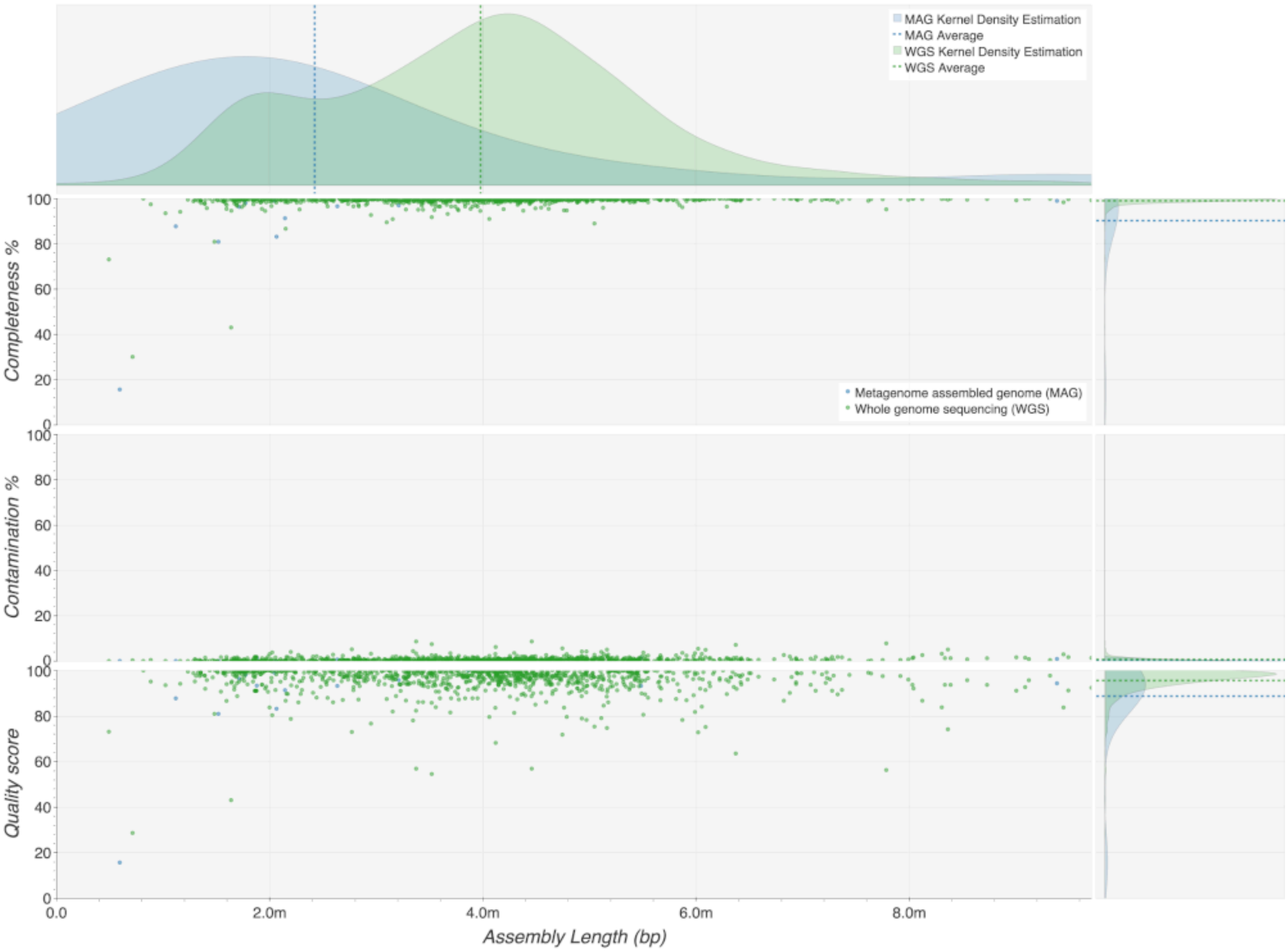


Fig. 1. WGSs (green) and MAGs (blue) from MarRef representing finished genomes are here distributed based on total genome assembly length (x-axis) and estimated (a) completeness, (b) contamination and (c) QS scores as resulting from CheckM. Kernel density estimates of the assembly length and metrics are projected on top and right sides.

5S, 16S and 23S rRNAs genes in assemblies vary from 4.44 in WGSs to 1.28 in MAGs. Only three WGSs and two MAGs did not contain a complete set of 5S, 16S and 23S rRNA. In MarDB, the average number of 5S, 16S, and 23S rRNA sets are 2.76, 0.47 and 0.80 for WGSs, MAGs and SAGs. About 50 % (49.47%) of the entries in MarDB contain more than one 5S, 16S and 23S rRNA.

All of the 970 entries in MarRef contain tRNAs, out of which 954 contain more than 18 unique tRNAs. The number of tRNAs is higher in WGSs than MAGs, on average 63.8 tRNA in WGSs and 44.2 in MAGs. In MarDB, the average number of tRNAs in WGSs, MAGs and SAGs are calculated to 59.8, 29.7, and 23.7, respectively. Of the 13190 entries predicted to contain tRNAs, 8540 (64.5%) have more than 18 unique tRNAs.

Quality assessment: completeness, contamination and quality score

Figures 1 and 2 (and further detailed in Supplementary Table 1) summarise the analysis of completeness, contamination and QS scores for the genomes in the MarRef and MarDB databases.

Although the MarRef database contains finished genomes, the completeness scores for WGS varied between 30 to 100, with contamination scores from 0 to 8.6, resulting in QS scores between 28.7 and 100. MAGs in MarRef scored lower on minimum completeness and contamination with 15.7 and 1.33, respectively, giving QS scores between 15.7 and 100. Similarly, the completeness and contamination scores for WGS type genomes in MarDB ranged from 0 to 100 and 0 to 200, respectively, with resulting QS between 100 and -900. Scores were similar for MAG type genomes except for higher maximum contamination (280) and subsequent minimum QS (-1303). SAG type genomes shared similar completeness scores, but contamination did not exceed 50.8, resulting in QS scores between 99.73 and -160.6. A summary of genome classifications (criteria listed in Table 1) based on the results is presented

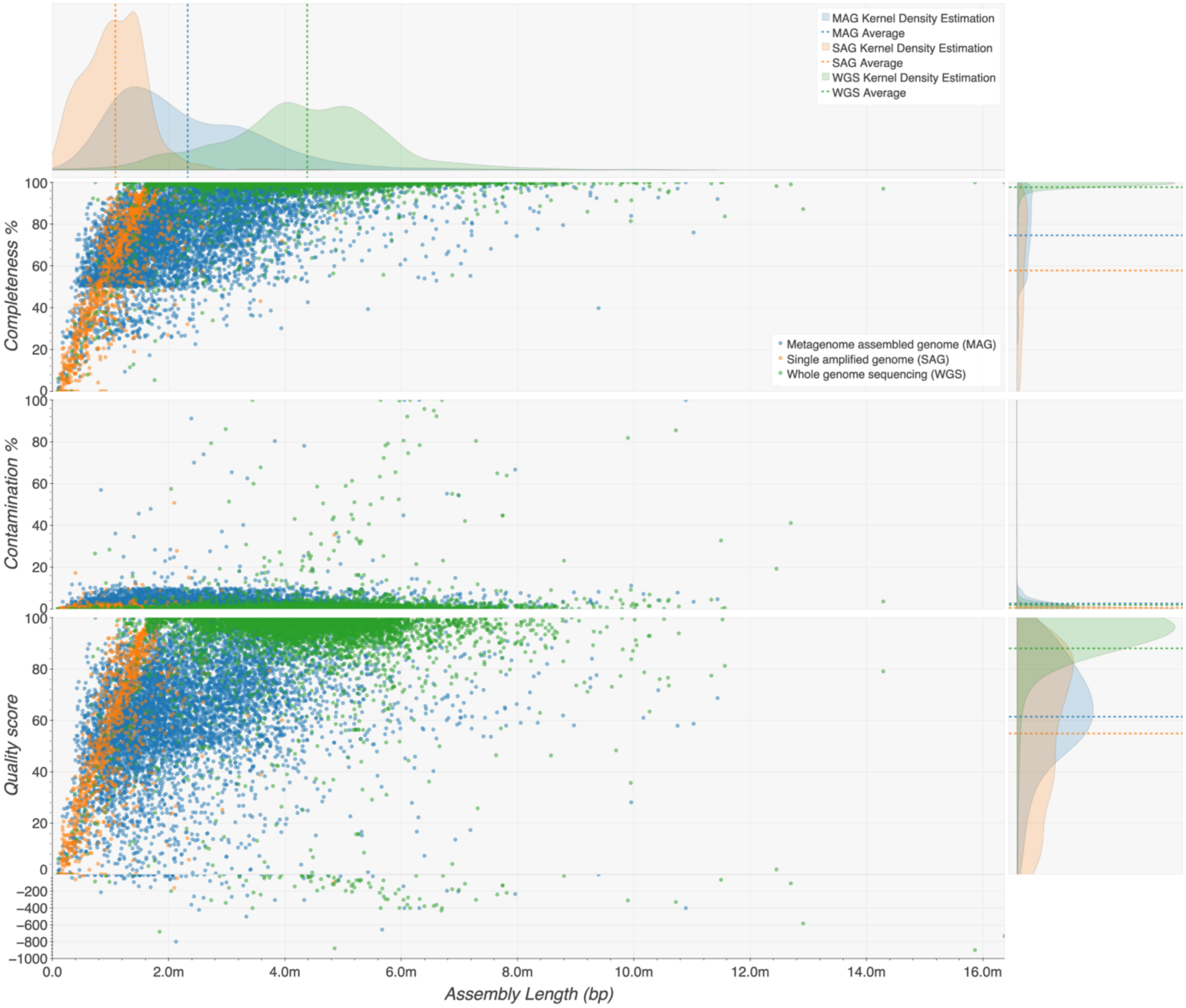


Fig. 2. Distribution of draft genome assemblies from MarDB as WGSs (green), MAGs (blue) and SAGs (orange) on total genome assembly length (x-axis). Estimated (a) completeness (b) and contamination scores and (c) resulting QS.

in Table 3 and shows 220 entries of MarDB fall within the class Very low-quality drafts.

Example cases of Very-low quality drafts genomes

The Very low-quality class introduced in this work is intended to fill the gap where published genomes do not fit according to current classification schemes of the GSC. Exemplified here are three genomes derived as WGS, SAG, and MAG. The SAG-assembly published as Verrucomicrobia bacterium SCGC AAA168-F10 (Martinez-Garcia *et al.,* 2012) (accession num. GCA_000264645.3) had estimated completeness and contamination scores of 84.29 and 33.49 with a strain heterogeneity of 8. GTDB-tk (Parks *et al.,* 2018) classified the assembly within the following order, family and genus; Verrucomicrobiales, Akkermansiaceae and SW10. However, 18.3% of markers had multiple hits. Evaluation of the 1397 contigs with CAT/BAT (Meijenfeldt *et al.,* 2019) resulted in the classification of Verrucomicrobiales with a score of 0.66. Changing the f-parameter to 0.01 delineated six lineages, here listed with gradually lower scores; Verrucomicrobiales bacterium (0.12), Balneola sp. (0.06), Rhodothermaeota bacterium (0.02), unclassified Verrucomicrobiaceae (0.02), Balneolaceae bacterium (0.01), Thaumarchaeota (0.01). The latter being Archaea.

A similar evaluation of the WGS assembly published as a Mumia flava strain MUSC 201 (Lee *et al.,* 2014) (accession num. GCA_000802255.1) resulted in a contamination score of 105.97 and strain heterogeneity of 6.62. With a completeness score of 100, the QS score was calculated to -732.7. Evaluation of 15640 ORFs by CAT/BAT resulted in Burkholderia with a score of 0.40. Changing f=0.01 delineated this assembly into three distinct lineages, here genera with scores; Burkholderia (0.40), Mumia (0.28) and Ralstonia (0.04). While the first and latter are Proteobacteria, the genera Mumia is in the phylum Actinobacteria.

The MAG Alteromonas sp. ESRF-bin4 (assembly accession num. GCA_002632225.1), comprising 1074 contigs, had a calculated 86.95 completeness, 21.52 contamination (-20.65 QS) and 88.49 strain heterogeneity. In estimating taxonomic affiliation 5026, ORFs scored 0.43 for the published genus Alteromonas using CAT/BAT. No further delineation of other taxa resulted after adjusting the f-parameter. Further details of the CAT/BAT output data can be found in Supplementary Table 2.

Table 3. Summed quality classification of entries in MarRef and MarDB. Table also shows in brackets the simulated number of *High-quality* and *Near-complete* classifications of MarRef if the class representing finished genomes were absent.

	Finished	High Quality	Near-Complete	Medium Quality	Low Quality	Very Low Quality
MarRef						
WGS	951	(732)	(204)			
MAG	15	(2)	(11)			
MarDB						
WGS		1386	3253	219	164	118
MAG		11	1372	5271	524	95
SAG		1	118	513	391	7

Trends

As novel type genome submissions such as MAGs and SAGs have significantly increased in recent years, it becomes vital to investigate the effects these particular entries have on genomic reference databases in terms of sequence quality and completeness. To examine this development, we made a plot distinguishing these discrete divisions in the MarDB marine database in recent years (Figure 3). We observe that while the general completeness of WGS based submissions increase over time, MAGS submissions have a downward completeness trend. SAG type entries typically have lower completeness on average; however, an increasing completeness trend overall.

4 Discussion

The development of new sequencing technologies and methods has greatly enhanced the number of finished genomes and draft assemblies in public repositories in the last few years. The methods for recovering the genomes such as WGS, SAG and MAG have their weaknesses and strengths, which are dependent on, among others, sample preparation, sequencing technology, coverage, and methods used to generate the genome consensus sequence. All of the weaknesses and strengths contribute to the overall quality of the consensus sequence. Since reference databases contain annotated genomes or genes used in most taxonomic classification and functional assignment tools, the quality must be as high as possible to avoid misassignment, essential in classifying metagenomic samples using either assembly-based classification or read (or K-mer) mapping approaches.

The contig length varies between the different technologies and databases. While the WGSs in MarRef have an average length of 3.99 Mbp, the WGSs in MarDB is somewhat higher with an average length of 4.38 Mbp. For the MAGs, it is the opposite, where the contig length is 2.60 Mbp in MarRef and 2.33 Mbp in MarDB. Although the differences are in general small, the disparity in average contig length of WGSs in MarRef and MarDB may arise from the fact that the genomes in MarRef are closed and contain less contamination than the MGSs in MarDB. The average contig length of the MAGs in MarRef and MarDB is approximately 65% and 54%, respectively, of the average length of the WGSs in the two databases. The observed difference between WGSs and MAGs in MarRef may reflect that a small MAG is more straightforward to close than larger MAGs. The average length of the SAGs in MarDB is only 1.09 Mbp. The low average length may be due to the multiple displacement amplification (MDA) chemistry itself and challenges for the assemblers to handle the SAG data (Kaster and Sobol, 2020). The MDA process often results in uneven coverage across the genome, and most assemblers rely on even coverage across the genome and therefore perform poorly on SAG datasets which often leads to partial genome recovery. The coverage has been improved by combining multiple sequenced single-cell genomes or enhancing the number of copies of the genome (Kogawa *et al.,* 2018).

The number of contigs in the MarRef entries are, as expected, low and varies from 1 to 12, which reflects that these are gapless chromosomes and replicons. For the admissions in MarDB, the number of contigs goes up to 8951, demonstrating both a considerable diversity in assembly level and contig length. However, these numbers for drafts may change as 6.3% of entries in MarRef, and 2.9% in MarDB represent revised assemblies conveying improved accuracy of genome reproductions. Databases are thus continuously dynamic towards updating assemblies, and low-quality genomes may be updated to better or even finished states. These are ultimately reflected in the MarRef and MarDB databases, but finished genomes in MarDB may still be inconspicuously overlooked as drafts due to manual assessment - limiting this study to consider general trends.

The presence of a minimum of one 5S, 16S and 23S rRNA in addition to 18 unique tRNA adapters are the requirements for the High-Quality draft category (Table 1) (Bowers *et al.,* 2017). In the MarDB database, the High-Quality draft is the most significant attainable assembly class and was assigned 1398 of its total entries. In this quality class, MAG and SAG type genomes were present with 11 and one assemblies. The low percentage present in this category (10.6%) illustrates the fluctuating quality represented by draft genomes as WGS, MAG or SAG. Indeed, close to 40% of WGS genomes have earlier been identified with lacking sets of tRNAs (Land *et al.,* 2014). In fact, 25 genomes labelled as finished in MarRef were not complying with the High-Quality draft requirements. Nine of these did not satisfy the complete rRNA triplet, and a further 16 did not comply with the required number of unique tRNA adapters - for example, the Dokdonia sp. PRO95 (Riedel *et al.,* 2013), which belongs to the Flavobacteriaceae family, has a single 16S rRNA gene while missing both the 5S and 23S.

Similarly, the strain F1 of Staphylothermus marinus (Anderson *et al.,* 2009), an isolate from a hydrothermal vent, contains only 15 unique tRNA adapters. Nonetheless, following the GSC quality requirements, the metrics concerning rRNA and tRNA (total number and number of unique genes) should be mandatory information for databases housing genomes. In classifying genome assemblies as High-quality drafts, the use of rRNA indicators might be speculative compared to estimated completeness. Not only will strict adherence to RNA requirements lead to the rejection of some gap-less, complete genomes, but also categorizing them as Medium-quality drafts.

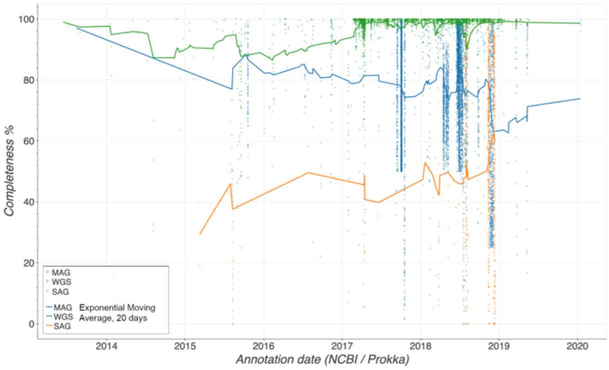


Fig. 3. Progression in completeness since 2013 for marine draft genomes of MarDB.

The resulting completeness scores of WGS, MAG, and SAG in both considered databases (MarRef and MarDB) demonstrated distinct calculated averages (Figure 1, 2 and Supplementary Table 1). As expected, MarRef finished WGSs and MAGs scored higher for completeness than draft genomes, corroborating earlier studies (Land *et al.,* 2014). For example, finished and draft MAGs had average scores of 90.26% and 74.61%, respectively. In MarDB, WGS average completeness was found in the high-end (97.66%) and MAGs in the middle-end scale (shown above). The low-end scale was occupied by SAGs (57.82%), where results corroborate earlier SAG completion estimates (Rinke *et al.,* 2014). However, internal variations were extensive as draft genome assemblies of all genome types were distributed from High-quality to Very low-quality drafts (Table 3). The distribution of completeness scores for draft MAG and SAG genomes in Figure 2a, however, contrast the Medium-quality mark at 50%. While SAG genomes remain prevalent (38% of entries) below this criterion, MAG genomes are present in scarcer cases (841, or 11.71% of entries). The notable distinction suggests quality restrictions are frequently enforced by submitting authors to avoid Low-Quality MAGs. However, the cost and effort going into SAGs might explain why genome assemblies are published regardless of guidelines and quality issues. Still, since 2013 the level of completeness in genomes has generally improved, as illustrated by Figure 3. The development of tools and sequencing techniques and best practices, and experience has likely advanced the representation of WGS and SAG-type genomes. In this context, SAGs may have gained completeness with the developing practice of single-cell sequencing, were more significant proportions of genomes are recovered and even finished without gaps (Woyke *et al.,* 2017). MAGs, on the contrary, illustrate a negative completeness trend over the years and can be the result of sizable metagenome binning projects. For example, one BioProject (accession num. PRJNA391950) represents the metagenomic study of microbial diversity in the subseafloor-crust (Tully *et al.,* 2017). This project lists 195 MAGs and contributes with 23 Very-low quality drafts in our reanalysis. Similarly, 289 genome assemblies classified as Low-Quality drafts, represented by BioProject PRJNA385762, are from metagenomic binning of deep-sea hydrothermal vent samples (Zhou *et al.,* 2019).

Not surprisingly, contamination was found more frequently in drafts of MarDB than finished genomes of MarRef. Contamination in MarRef was found negligible (Figure 1b), only affecting quality scores marginally. Entries surpassing the contamination level of 10% amounted to 220 Very low-quality draft assemblies in MarDB. As shown in Figure 2b, the number of highly contaminated assemblies in this class is notable but not extensive and may be regarded as outliers. Outliers are primarily attributed to WGS and MAGs, with a few exceptions of SAGs. Results here demonstrate the presence of nonconforming genome assemblies regarding classifications schemes presented by Bowers *et al.,* 2017.

All three example cases of Very Low-quality drafts (SAG, WGS and MAG) are well within the threshold by having 33.49%, 105.97% and 21.52% contamination, respectively. Both the SAG and WGS example could be split into other taxa based on contig re-evaluations. Analysis with CAT/BAT (Meijenfeldt *et al.,* 2019) concluded the SAG taxon as Verrucomicrobiales bacterium, same as published (Martinez-Garcia *et al.,* 2012), but with possible contaminants from up to six other taxa, including Archaea. Contamination in SAGs is not uncommon and has received considerable attention during pre-sequencing steps to mitigate its impact on the final sequencing output (Rinke *et al.,* 2014). Still, SAGs published in the early 2010s may not have received rigorous post-assembly evaluation of genome contamination, a topic widely emphasized with improved autonomy in genome quality assessment from programs like CheckM and BUSCO (Parks *et al.,* 2015, Simão *et al.,* 2015). For the highly contaminated WGS example, re-evaluation of the Actinobacteria-species Mumia flava (Lee *et al.,* 2014) indicated contamination with Burkholderia contaminans and Ralstonia of the Proteobacteria phylum. Substantial contamination levels in sequenced reads have previously been identified in WGS isolates of pure cultures (Goig *et al.,* 2020). The Burkholderia contaminans has been described as a contaminating species found in various environments (Vanlaere *et al.,* 2009, Savi *et al.,* 2019). On the other hand, the Very-low quality MAG draft, Alteromonas sp. ESRF-bin4 is an example with moderate to high contamination (21.52%) and high strain heterogeneity (88.49). However, considering strain heterogeneity as an indicator of multiple strains or divergent taxa (Parks *et al.,* 2015), CAT/BAT could not separate other taxa in this case.

5 Conclusion

Genomes reaching public repositories like ENA and NCBI do not consistently categorize within GSC quality guidelines. In general, quality (completeness and contamination) cannot be rigorously inspected without metrics from programs like CheckM, which we propose as mandatory for submission to the INSDC databases. Furthermore, as genome types (WGS, MAG, SAG) can significantly impact what quality to expect, we propose their recovery method as compulsory to document for authors publishing genome assemblies. Regardless of the recovery method, caution must be taken as significant proportions of genome assemblies fall into Medium, Low and Very low-quality drafts. Consequently, adverse conclusions may arise if these quality categories are applied as functional or taxonomic classification references. Ideally, completeness > 90% and contamination < 5% should lead to no less than a minimum QC score > 65 for reference genomes. However, stringent use of rRNA and tRNA must be considered with care as finished genomes without gaps are apt to circumvent these requirements.

Acknowledgements

-

Funding

This project has been funded by The Arctic University of Norway.

Conflict of Interest: none declared.

References

Almeida,A. *et al.* (2019) A new genomic blueprint of the human gut microbiota. *Nature*, **568**, 499–504.

Anderson,I.J. *et al.* (2009) Complete genome sequence of staphylothermus marinus stetter and fiala 1986 type strain F1. *Stand. Genomic Sci.*, **1**, 183–188.

Bowers,R.M. *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.

Chain,P.S.G. *et al.* (2009) Genome project standards in a new era of sequencing. *Science (80-.).*, **326**, 236–237.

Cochrane,G. *et al.* (2016) The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, **44**, D48–D50.

Fleischmann,R.D. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (80-.).*, **269**, 496–512.

Goig,G.A. *et al.* (2020) Contaminant DNA in bacterial sequencing experiments is a major source of false genetic variability. *BMC Biol.*, **18**.

Gurevich,A. *et al.* (2013) QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

Haroon,M.F. *et al.* (2013) Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. *Nature*, **500**, 567–570.

Harrison,P.W. *et al.* (2021) The European Nucleotide Archive in 2020. *Nucleic Acids Res.*, **49**, D82–D85.

Kalyanaraman,A. (2011) Genome Assembly. In, Padua,D. (ed), *Encyclopedia of Parallel Computing*. Springer US, Boston, MA, pp. 755–768.

Kaster,A.K. and Sobol,M.S. (2020) Microbial single-cell omics: the crux of the matter. *Appl. Microbiol. Biotechnol.*, **104**, 8209–8220.

Klemetsen,T. *et al.* (2018) The MAR databases: Development and implementation of databases specific for marine metagenomics. *Nucleic Acids Res.*, **46**.

Kogawa,M. *et al.* (2018) Obtaining high-quality draft genomes from uncultured microbes by cleaning and co-assembly of single-cell amplified genomes. *Sci. Rep.*, **8**, 2059.

Land,M.L. *et al.* (2014) Quality scores for 32,000 genomes. *Stand. Genomic Sci.*, **9**, 20.

Lee,L.H. *et al.* (2014) Mumia flava gen. nov., sp. nov., an actinobacterium of the family Nocardioidaceae. *Int. J. Syst. Evol. Microbiol.*, **64**, 1461–1467.

Lowe,T.M. and Chan,P.P. (2016) tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.*, **44**, W54–W57.

Martinez-Garcia,M. *et al.* (2012) Capturing single cell genomes of active polysaccharide degraders: An unexpected contribution of verrucomicrobia. *PLoS One*, **7**, 35314.

Meziti,A. *et al.* (2021) The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Appl. Environ. Microbiol.*, **87**, 1–15.

Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.

Parks,D.H. *et al.* (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, **36**, 996–1004.

Parks,D.H. *et al.* (2015) CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043–1055.

Parks,D.H. *et al.* (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.*, **2**, 1533–1542.

Riedel,T. *et al.* (2013) Genomics and Physiology of a Marine Flavobacterium Encoding a Proteorhodopsin and a Xanthorhodopsin-Like Protein. *PLoS One*, **8**, 57487.

Rinke,C. *et al.* (2014) Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat. Protoc.*, **9**, 1038–1048.

Rinke,C. *et al.* (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, **499**, 431–437.

Roux,S. *et al.* (2019) Minimum information about an uncultivated virus genome (MIUVIG). *Nat. Biotechnol.*, **37**, 29–37.

Salzberg,S.L. *et al.* (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.*, **22**, 557–567.

Savi,D. *et al.* (2019) Impact of clonally-related Burkholderia contaminans strains in two patients attending an Italian cystic fibrosis centre: A case report. *BMC Pulm. Med.*, **19**, 1–8.

Schoch,C.L. *et al.* (2020) NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database*, **2020**.

Sharon,I. *et al.* (2013) Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.*, **23**, 111–120.

Simão,F.A. *et al.* (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Tully,B.J. *et al.* (2018) A dynamic microbial community with high functional redundancy inhabits the cold, oxic subseafloor aquifer. *ISME J.*, **12**, 1–16.

Vanlaere,E. *et al.* (2009) Taxon K, a complex within the Burkholderia cepacia complex, comprises at least two novel species, Burkholderia contaminans sp. nov. and Burkholderia lata sp. nov. *Int. J. Syst. Evol. Microbiol.*, **59**, 102–111.

Von Meijenfeldt,F.A.B. *et al.* (2019) Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.*, **20**, 1–14.

Woyke,T. *et al.* (2017) The trajectory of microbial single-cell sequencing. *Nat. Methods*, **14**, 1045–1054.

Wrighton,K.C. *et al.* (2012) Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science (80-.).*, **337**, 1661–1665.

Zhou,Z. *et al.* (2020) Genome- and Community-Level Interaction Insights into Carbon Utilization and Element Cycling Functions of Hydrothermarchaeota in Hydrothermal Sediment . *mSystems*, **5**.

Paper 4

T. Klemetsen, N. P. Willassen, and C. Karlsen, "Full-length 16S rRNA gene classification of Atlantic salmon bacteria and effects of using different 16S variable regions on community structure analysis," *Microbiologyopen*, 2019, doi: 10.1002/mbo3.898.

Received: 11 April 2019 | Revised: 7 June 2019 | Accepted: 10 June 2019

DOI: 10.1002/mbo3.898

ORIGINAL ARTICLE

MicrobiologyOpen Open Access WILEY

Full-length 16S rRNA gene classification of Atlantic salmon bacteria and effects of using different 16S variable regions on community structure analysis

Terje Klemetsen[1] | Nils Peder Willassen[1] | Christian René Karlsen[2]

[1]Department of Chemistry, Center for Bioinformatics, UiT The Arctic University of Norway, Tromsø, Norway

[2]Department of fish health, Nofima, Norway

Correspondence
Terje Klemetsen, Department of Chemistry, Center for Bioinformatics, UiT The Arctic University of Norway, Tromsø, Norway.
Email: Terje.klemetsen@uit.no

Funding information
Norges Forskningsråd, Grant/Award Number: 194050/F40; Research Council of Norway, Grant/Award Number: 194050 and F40; UiT The Arctic University of Norway

Abstract

Understanding fish-microbial relationships may be of great value for fish producers as fish growth, development and welfare are influenced by the microbial community associated with the rearing systems and fish surfaces. Accurate methods to generate and analyze these microbial communities would be an important tool to help improve understanding of microbial effects in the industry. In this study, we performed taxonomic classification and determination of operational taxonomic units on Atlantic salmon microbiota by taking advantage of full-length 16S rRNA gene sequences. Skin mucus was dominated by the genera *Flavobacterium* and *Psychrobacter*. Intestinal samples were dominated by the genera *Carnobacterium*, *Aeromonas*, *Mycoplasma* and by sequences assigned to the order Clostridiales. Applying Sanger sequencing on the full-length bacterial 16S rRNA gene from the pool of 46 isolates obtained in this study showed a clear assignment of the PacBio full-length bacterial 16S rRNA gene sequences down to the genus level. One of the bottlenecks in comparing microbial profiles is that different studies use different 16S rRNA gene regions. Comparisons of sequence assignments between full-length and in silico derived variable 16S rRNA gene regions showed different microbial profiles with variable effects between phylogenetic groups and taxonomic ranks.

KEYWORDS

Atlantic salmon, Full-length 16S rRNA gene sequence, microbiota

1 | BACKGROUND

Second-generation sequencing is widely used to assess the composition of the microbial community through partial sequence analysis of the 16S rRNA gene. Different bacterial 16S rRNA gene regions are used by different researchers, which makes it difficult to perform global comparisons of microbiome studies. Discrepancy in bacterial diversity may also be observed between full-length 16S rRNA gene and variable region 16S rRNA gene datasets (Sun, Jiang, Wu, & Zhou, 2013; Wagner et al., 2016). Sequence regions but also PCR primer choice used for short-read amplicon sequencing of different 16S rRNA gene hypervariable regions can affect the accuracy of the inferred community profiles and sensitivity to certain bacterial taxa (Chen et al., 2019; Walker et al., 2015). Therefore, sequences of full-length 16S rRNA genes that cover all the variable regions should potentially increase the accuracy and the resolution of closely related taxa. However, full-length sequences compared to shorter sequences generated by other platforms favorized Proteobacteria and provided a lower taxonomic profiling of the human feces, partly due to sequence accuracy and low coverage of terminal regions in the 16S rRNA databases (Whon et al., 2018). Still, long-read sequencing such as PacBio circular consensus sequencing (CCS) applied on the 16S rRNA gene provides a promising approach to increase the taxonomic resolution of microbial communities. The CCS technology generates a consensus sequence from a single molecule by reading a ligated circular DNA template multiple times, achieving high read accuracy (Travers, Chin, Rank, Eid, & Turner, 2010). However, few studies have investigated advantages and disadvantages of PacBio CCS for such analyses. A recent study showed that PacBio sequencing error rates were in the same range as Roche 454 and MiSeq platforms (Wagner et al., 2016). The authors reported inconsistencies in species-level analysis between full-length 16S rRNA gene sequences obtained from Sanger and PacBio sequencing comparisons and that more sample types were needed to determine whether partial or full-length 16S rRNA gene sequences was superior in terms of taxonomic profiling effectiveness. The first metagenomic marine environmental samples from PacBio CCS sequences provided a superior taxonomic resolution to the species level compared to in silico derived partial regions of the gene (Pootakham et al., 2017).

In fish, both the skin epithelial surface and the gastrointestinal tract are covered by a mucosal layer. The main components of this mucus layer are secreted glycoproteins called mucins, which are differentially regulated between tissue types in Atlantic salmon, *Salmo salar* (Sveen, Grammes, Ytteborg, Takle, & Jørgensen, 2017). These glycosylated proteins might be a highly attractive substrate for the attachment and settlement of microorganisms. Interaction studies between cutaneous microbiota and the fish surface are scarce, but mutualistic relationships (Beklioglu, Telli, & Gozen, 2006) and the existence of a resilient microbiome (Larsen, Bullard, Womble, & Arias, 2015) has been suggested. Salinity acclimation results in turnover of dominant bacterial taxa within the host microbiome that are unrelated to changes in water microbiota (Schmidt, Smith, Melvin, & Amaral-Zettler, 2015), which is also reported for Atlantic salmon (Karlsen et al., 2017; Lokesh & Kiron, 2016). Microorganisms populating the gastrointestinal tract are believed to take part in digestive functions and contribute to fish health (Nayak, 2010). Many fish gut bacteria are not detected from water samples (Sullam et al., 2012), and gut microbiota profiles in Atlantic salmon change between intestinal compartments (Gajardo, Rodiles, et al., 2016), rearing environments (Dehler, Secombes, & Martin, 2017), and diets (Schmidt, Amaral-Zettler, Davidson, Summerfelt, & Good, 2016). Sequencing the 16S rRNA genes is a powerful tool that provides a comprehensive picture of the phylogenetic diversity and composition of the microorganisms present in a sample as many of the microbial groups are absent or difficult to cultivate. Traditionally, diagnosis of suspected Atlantic salmon bacterial infections has relied on clinical signs and symptoms, and microbiological culture-dependent methods. Many bacteria associated with salmonids have been considered difficult to cultivate and more accurate culture-independent diagnostic procedures are developed (Grove, Reitan, Lunder, & Colquhoun, 2008; Sepúlveda, Bohle, Labra, Grothusen, & Marshall, 2013). However, culturing is an important step to better understand effects of microorganisms. Recovering isolates of microbial symbionts is increasing in focus as they can be used to study activity and functional relationships to a host (Esteves, Amer, Nguyen, & Thomas, 2016; KleinJan, Jeanthon, Boyen, & Dittami, 2017). The present study was also designed to recover and identify members of the Atlantic salmon microbial communities, characterized by 16S rRNA gene sequencing, for future studies.

The changing environment in the salmon production, utilization of different feeds, use of closed or semi-closed recirculation systems and transfer to unprotected seawater environment at the final stage of production affects skin and gut of complex microbial communities. Taxonomic profiling that can reveal alterations or deviations from “normal” microbial communities may advance our understanding of any functional effects of detected microbiota. Appropriate methods that accurately generate and analyze the microbial communities would be an important tool to help improve understanding of microbial effects in the aquaculture industry. Here, we applied long-read technology to demonstrate its utility, as a proof of concept, in characterizing the microbial profiles of the skin surface and the bulk intestinal content of Atlantic salmon.

2 | MATERIALS AND METHODS

2.1 | Sample collection and DNA extraction

This study utilized fish from an industry research study conducted at 12 ppt salinity with recirculated water with temperature of 13°C at the Research Station for Sustainable Aquaculture (Sunndalsøra, Norway) in accordance with regulations of the Norwegian Food Safety Authority. As proof of concept, six Atlantic salmon were selected for sampling, anesthetized with benzocaine and killed by a blow to the head. The six sampled individuals were split between two dietary groups. Three had been fed fish meal (mean ± *SEM* body weight was 193 ± 41 g), and three were from a diet where fish meal was substituted with krill meal (mean ± *SEM* body weight was 181 ± 20 g). Fish were not fed for 48 hr prior to sampling. Skin mucus samples were obtained by swiping a cell swiper across the left side of the fish and concentrating mucus, which was aspirated by pipetting. The abdominal cavity was then opened and an incision was made to open the distal intestine. The intestinal content was collected by gently scraping the intestine to collect bulk feces including the intestinal mucus layer. All tissue material was stored in 96% EtOH. DNA was extracted from 200 mg skin mucus samples and 100 mg intestinal samples. The protocol was performed using the PowerLyzer® PowerSoil® DNA Isolation Kit (MoBio) according to the manufacturer's specification with the following amendments: samples after adding Solution C1 were heated at 70°C for 10 min. Samples were homogenized with the mechanical bead beater device Precellys®24 (Bertin Technologies) for 1 × 20 s at 5,000 rpm. The DNA was resuspended in 30 µl of DNase/RNase free molecular water and concentration determined using a Thermo Scientific Nanodrop 2000c.

2.2 | Bacterial isolation and identification

Bacteriology was performed on the same sampled individuals for the total bacterial DNA sequence analysis. Skin mucus (200 mg) and bulk intestinal feces/mucus (100 mg) were separately vortexed and suspended in total of 1 ml 0.9% NaCl saline solution. In addition, gill mucus material and water were similarly suspended in saline solution to retrieve bacterial isolates, included in the Appendix A. A 100 µl aliquot of the content was serial diluted up to 10^{-6} and plated in duplicate onto R2A (BD Difco), representing low-nutrient conditions and MacConkey agar (CM007 Oxoid), representing high-nutrient conditions, under aerobic incubation at 12°C for 9 days. Plates were inspected and colony numbers were counted based on morphological characteristics, that is, pigmentation, colony form, elevation, surface appearance, and texture. The relative distribution in percentage between colony morphologies is provided as an average of each sample type. Representative colonies were selected according to dominant morphologies and then identified by 16S rRNA gene sequence analysis. Briefly, representative colonies were selected for purity plating onto R2A or MacConkey plates. Pure colonies determined for freeze stocks were further expanded in Luria-Bertani (LB) broth (Bertani, 1951) with 3.5% NaCl at 12°C before supplementation with 10% glycerol and stored at −80°C. Genomic DNA isolation was performed using PureLink® Genomic DNA Mini Kit (Invitrogen). PCR amplification of the 16S rRNA gene with primers 27F (5′-AGAGTTTGATCMTGGCTCAG) and 1492R (5′-TACCTTGTTACGACTT) was identical to previous descriptions (Karlsen et al., 2014). Products were visualized following agarose (1.0%) gel electrophoresis and RedSafe (Chembio), before being purified using a QIAquick Gel Extraction Kit (Qiagen) followed by Sanger sequencing (sequenced at GATC Biotech, DNA sequencing services and bioinformatics, Germany). The forward and reverse sequences were assembled in Bioedit (Hall, 1999) and consensus sequences deposited in GenBank (submission: SUB3162162), with accession numbers MG263463-MG263508 (Table A1, Appendix). Sequences were aligned with type strain reference sequences using Sequence Match software from The Ribosomal Database Project II (RDP II) web site (Cole et al., 2014). The phylogenetic relationships between sequences were constructed utilizing selected 16S rRNA sequences of type strains in each genus (Figure A1). Sequences were aligned using the ClustalW algorithm in BioEdit (Hall, 1999). The phylogenetic relationships were determined using maximum likelihood (ML) based on the Tamura 3-parameter model including all coding positions (total of 1,425) with 1,000 bootstrap trials (neighbor-joining tree) in MEGA6 (Tamura, Stecher, Peterson, Filipski, & Kumar, 2013).

2.3 | PCR amplification, barcoding, and PacBio sequencing

To analyze the microbial population associated with the skin mucus and intestine, sequencing of the 16S rRNA gene was performed using the PacBio sequencing technology (Pacific Biosciences). The full-length 16S rRNA gene was amplified using degenerated versions of the universal bacterial 16S rRNA gene primers 27 F (5′- AGRGTTTGATYMTGGCTCAG) and 1492 R (5′-GGYTACCTTGTTACGACTT). In accordance with "Guidelines for Using PacBio® Barcodes for SMRT® Sequencing" guide, a 5 nt (5′-GGTAG) padding sequence was added to each unique 16 nt barcode to allow all barcodes to ligate to the SMRTbell™ adapter with equal efficiency. The utilization of barcodes allowed the multiplex sequencing of amplicons from several samples in one library using SMRT®. Primers were synthesized and HPLC-purified as recommended in PacBio's SMRT guidelines by Invitrogen Custom DNA Oligos (Thermo Fisher Scientific). Barcoded 16S rRNA amplicons were obtained by a two-step amplification protocol using Phusion® High-Fidelity PCR Master Mix (Thermo Fisher Scientific). The first PCR was performed in triplicate with the 27 F and 1492 R universal bacterial 16S rRNA gene primers using 100 ng of extracted total DNA, 1 × Phusion Master Mix, 0.5 µmol/L 16S-F forward primer, 0.5 µmol/L 16S-R reverse primer, in a 50 µl reaction volume. Samples were prepared on ice and amplified in the thermocycler with the block preheated to 98°C. The reactions were performed using the following cycling conditions: preincubation at 98°C for 2 min, followed by 25 cycles of denaturation at 98°C for 10 s, annealing at 55°C for 15 s, elongation at 72°C for 60 s, and a final extension step at 72°C for 3 min. Triplicate samples were pooled, and amplification was verified by 1% agarose gel electrophoresis before the reaction products were purified with an Invitrogen™ PureLink™ PCR Purification Kit. The purified PCR products were diluted and triplicates of 1 ng DNA was used as template for the second amplification reactions to generate padded barcoded products, with reagent concentrations as described above. Product amplification was as above with the following changes: 14 cycles with annealing at 60°C for 15 s. Triplicates were pooled, and products were verified by agarose gel electrophoresis and padded barcoded 16S rRNA gene amplicons from the reaction were purified using PureLink™ PCR Purification Kit and quantified using a NanoDrop spectrophotometer. Purified barcoded amplicons from the 12 samples (skin and intestine from 6 fish) were then pooled in equimolar concentrations, and 250 ng of DNA was used for library preparation at the Norwegian Sequencing Centre (www.sequencing.uio.no). Briefly, library was prepared using PacBio 2 kb library preparation protocol. Size selection was performed using Ampure beads. Adapters were ligated onto the barcoded amplicons, and the library was sequenced on a PacBio RSII system using the P6-C4 polymerase and chemistry with a 360-min movie time, using one SMRT cell.

2.4 | Sequence data analysis

Raw reads were filtered and demultiplexing using RS_subreads.1 pipeline on SMRT Portal (software version 2.3) with the following settings: minimum number of passes = 1, minimum predicted accuracy = 0.90, and minimum barcode score = 30. Read sequences were then prepared by filtering to a window length between 1,000 and 1,600 nt using PRINSEQ v.0.20.4 (Schmieder & Edwards, 2011) and reoriented in accordance with SILVA v.132 SSU (Pruesse et al., 2007) with the USEARCH v.10 (Edgar, 2010) orient function. The hypervariable regions v3-v4 and v5-v6 in the filtered data were extracted after using BLASTN v.2.6.0 (Altschul, Gish, Miller, Myers, & Lipman, 1990) (word size 4, gap opening penalty 0, *E*-value 0.01) to mark the flanking regions as described elsewhere (Pootakham et al., 2017). Sequences positively identified with both v-regions were advanced to generate two new trimmed data subsets for the v3-v4 and v5-v6 regions. Sequences were discarded if a flanking region could not be determined, suggesting a missing or partial v-region for a given sequence.

LCAClassifier (Lanzén et al., 2012) was applied to obtain the taxonomic mapping of the datasets (sample acronyms I1-I6 and SM1-SM6) and their respective data subsets. These were individually aligned against the SilvaMod database by MEGABLAST (Morgulis et al., 2008) (identity cutoff = 75.0, E-value cutoff = 0.001) as recommended prior to applying LCAClassifier. Finally, the analysis was carried out using default settings in the LCAClassifier program. Numerical data output from taxonomic classification was ordered based on ranks of taxonomy and the assignments obtained for the 12 datasets and data subsets. This was used to calculate the variation in taxon mapping between full-length PacBio CCS 16S rRNA gene sequences and their respective v-regions. To detail taxonomic variations as principal components linked to the sampling sites of intestine and skin mucus only the full-length 16S rRNA gene sequences were considered. The class level was used due to a 98.92% or greater successful assignment of the filtered full-length sequences of any sample. The sample dataset sizes were downscaled to sample SM2, which had the lowest number of successfully assigned sequences. The scaled values of class data, treating zero as "NA," were uploaded to ClustViz (Metsalu & Vilo,) with parameters set to not perform row scaling and the method set to SVD with imputation. Rarefaction analysis on the obtained operational taxonomic units (OTUs) was conducted using MicrobiomeAnalyst (Dhariwal et al., 2017). Unscaled OTU counts from mapping of reads with USEARCH were provided alongside their taxonomy lineage to genus level. Filtering of data was set using default parameters. Data were not rarefied or transformed, but total sum scaling was applied. The rarefaction curve was obtained with the filtered data using 20 steps.

Cultivability was determined using the 43,910 pooled full-length sequences combined with the 46 plate-isolated strains. Next, the USEARCH function cluster_fast was applied to cluster all sequences within an identity threshold of 97%. Sequences in clusters containing at least one of the cultivated strains were included when computing the cultivability percentage. OTUs were also determined for the pooled dataset of 43,910 full-length 16S rRNA gene sequences positive for v3v4 and v5v6 regions. USEARCH was applied following dereplication of the dataset, clustering OTUs at a 97% identity threshold, keeping parameters at default. 31,634 (~72%) of the full-length sequences were successfully reassigned to the 10 OTU representatives found and counted according to sample origin. The sum of sample specific assignments was scaled to fit the smallest sample size of SM2 and incorporated as pie charts in the phylogenetic inference described below. The OTU centroid sequences were further used to infer phylogenetic relationship between these representatives along with the 46 plate-isolated strains and 10 reference type strains. The 66 sequences were aligned with Clustal Omega v.1.2.1 (Sievers & Higgins, 2014) using default nucleotide parameters. The complete alignment was inferred using NeighborNet network with Uncorrected P distances in SplitsTree v.4.13.1 (Huson & Bryant, 2006).

3 | RESULTS

3.1 | Characteristics of full-length 16S rRNA gene sequencing

Full-length 16S rRNA gene sequences were amplified from DNA extracted from sampled bulk intestinal content and skin mucus of six Atlantic salmon. Amplicons were generated using a two-step PCR approach with asymmetrical primers during the second round of amplification for a more cost-effective way to multiplex amplicons from several samples. A total of 110,818 reads were obtained with an average read length of 23,881 nt. Of these were 76,723 assembled and demultiplexed into CCS reads with an average accuracy of 98.56% and a mean length of 1,252 nt. The average number of full passes of the CCS reads was 15.6. The number of processed full-length 16S rRNA sequences per sample ranged from 1,483 to 7,634 reads (Table 1). The microbiota composition was determined by a phylotyping approach directly allocating sequences into taxonomic groups based on bitscore and identity threshold. Of the trimmed reads used for assignment, an average of 96.9% of the skin mucus and 78.8% of the intestine were aligned to the taxonomic level order. According to the taxonomic resolution from order to genus, this method assigned more reads to taxonomic references for skin mucus samples compared to intestinal samples (Figure 1). The discriminant taxon, based on the generated reads from the intestine samples was the prominent Clostridiales that became unassigned at higher resolution ranks. The rarefaction analysis (Figure A2) of clustered OTUs showed that the bacterial communities of the intestines are less diverse compared to the skin mucus samples. Convergence were reached (sample I4, I5, I6, SM1, and SM3) but seven out of twelve samples, including both intestine and skin mucus, did not converge properly in the analysis.

3.2 | Comparisons of microbial compositions

Hierarchical clustering strengthened by principal component analyses (Figure 2) revealed that bacteria communities clustered to tissue types sampled from the Atlantic salmon in both the taxonomic rank of order to genus. Skin mucus microbiota profiles are tightly clustered. The intestinal microbiota profiles are dispersed but separates into krill meal diet (acronym I1 – I3) with predominate *Aeromonas* and *Mycoplasma* at the genus level, and fish meal diet (acronym I4 – I6) with predominant *Carnobacterium* at the genus level. The most dominant

TABLE 1 Sample names, acronyms, and PacBio CCS sequence characteristics

	Sample	Acronym	PacBio CCS reads	Total assignments
Krill meal	Fish 1 Skin Mucus	SM1	4,481	3,675
	Fish 1 Intestine	I1	4,681	4,025
	Fish 2 Skin Mucus	SM2	1,483	1,226
	Fish 2 Intestine	I2	4,810	4,195
	Fish 3 Skin Mucus	SM3	2,816	2,424
	Fish 3 Intestine	I3	7,634	6,664
Fish meal	Fish 4 Skin Mucus	SM4	4,138	3,440
	Fish 4 Intestine	I4	4,971	4,300
	Fish 5 Skin Mucus	SM5	4,022	3,155
	Fish 5 Intestine	I5	5,375	4,581
	Fish 6 Skin Mucus	SM6	4,945	3,960
	Fish 6 Intestine	I6	2,633	2,265

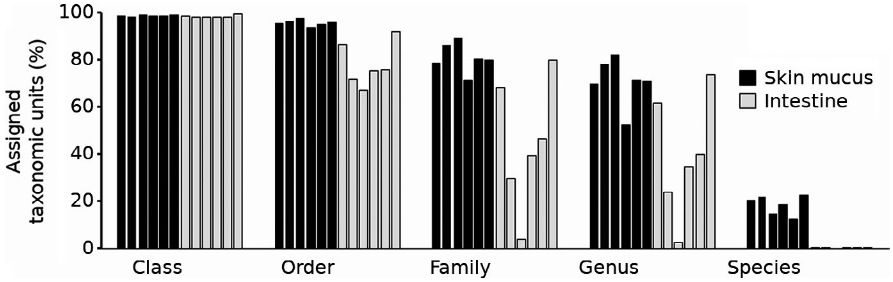


FIGURE 1 Percentage of taxonomic units assigned as reads at the class, order, family, genus, and species levels. Black bars = sampled fish skin mucus (fish no. 1 – 6). Gray bars = sampled intestine (fish no. 1 – 6)

bacteria orders of the skin mucus samples were Flavobacteriales (51.2%), Pseudomonadales (40.7%), Burkholderiales (2.3%), and Alteromonadales (1.3%). Members of the Flavobacteriales assigned to the genus level were dominated by *Flavobacterium* > *Chryseobacterium* > Leeuwenhoekiella > *Bizionia* > Gillisia. Genus members within the order of Pseudomonadales were dominated by *Psychrobacter* > *Pseudoalteromonas*. Sequences within remaining orders were not assigned down to genus level. The most dominant bacterial orders of the intestinal samples were Lactobacillales (35.0%) with *Carnobacterium* at the genus level, Clostridiales (26.1%), Aeromonadales (16.0%) with *Aeromonas* at the genus level, Mycoplasmatales (4.1%) with *Mycoplasma* at the genus level. Sequences within Clostridiales were not assigned down to the genus level.

To determine whether longer sequences of the 16S rRNA gene would be advantageous to assign sequences to the lower ranks of taxonomic affiliation, partial sequences spanning the v3v4 and v5v6 regions were extracted in silico from the full-length 16S rRNA gene sequence dataset. The proportion of assigned sequences between the datasets at the class, order, family, genus, and species levels were evaluated (Figure 3a). Using the SilvaMod database showed that overall sequence assignments varied between the full-length 16S rRNA sequences and partial 16S rRNA sequences in addition to differences in the proportions of the assigned sequences at the different taxonomic ranks (Table 2). The v3v4 sequences had the highest proportion of assignment at the family and genus levels. At the species level, the full-length dataset had the highest proportion of assigned sequences (mean 8.9%).

Figure 3b gives insight into the distribution of genus diversity in the different data subsets. The number of genus in full-length, v3v4, and v5v6 datasets were 44, 60, and 63, respectively. The v3v4/v5v6 sequences included 47 genera, while the full-length/v3v4 sequences included 39 genera and the full-length/v5v6 sequences included 39 genera. Thirty-six genera were present in all datasets. The relative proportion of total sequences assigned to the 36 jointly shared genera was 49.04%, 59.62%, and 43.44% for the full, v3v4 and v5v6 datasets, respectively. In contrast, the proportion of sequences assigned to the 42 remaining ancillary genera ranged between 0.01% and 0.05% (Figure 3b). To further compare taxonomic profiling, the number of sequences assigned to each bacterial taxon was identified (Table 2). This revealed differences in efficiency of read assembly in the different datasets. The discriminant taxon at the order level was Clostridiales with 64.5% and 63.2% higher assignments in the v3v4 and v5v6 data subsets, respectively. Discrepancy is also seen between short sequence length data subsets. An apparent example, which is observed down to the genus level, is sequences assigned to *Carnobacterium* where 23.1% more v3v4 sequences are assigned compared to the full-length, while the v5v6 data-subset had only 0.3% assigned compared to the full-length dataset. The data further suggest differences in the proportion of assigned sequences between sequence length and the taxonomic rank genus and species. Both v3v4 and v5v6 region sequences have a higher proportion assigned to the genus *Flavobacterium* compared to the full-length dataset. This is opposite to species level where no v3v4 or v5v6 sequences are assigned while 8.3% of the full-length sequences are assigned to *Flavobacterium frigidarium*.

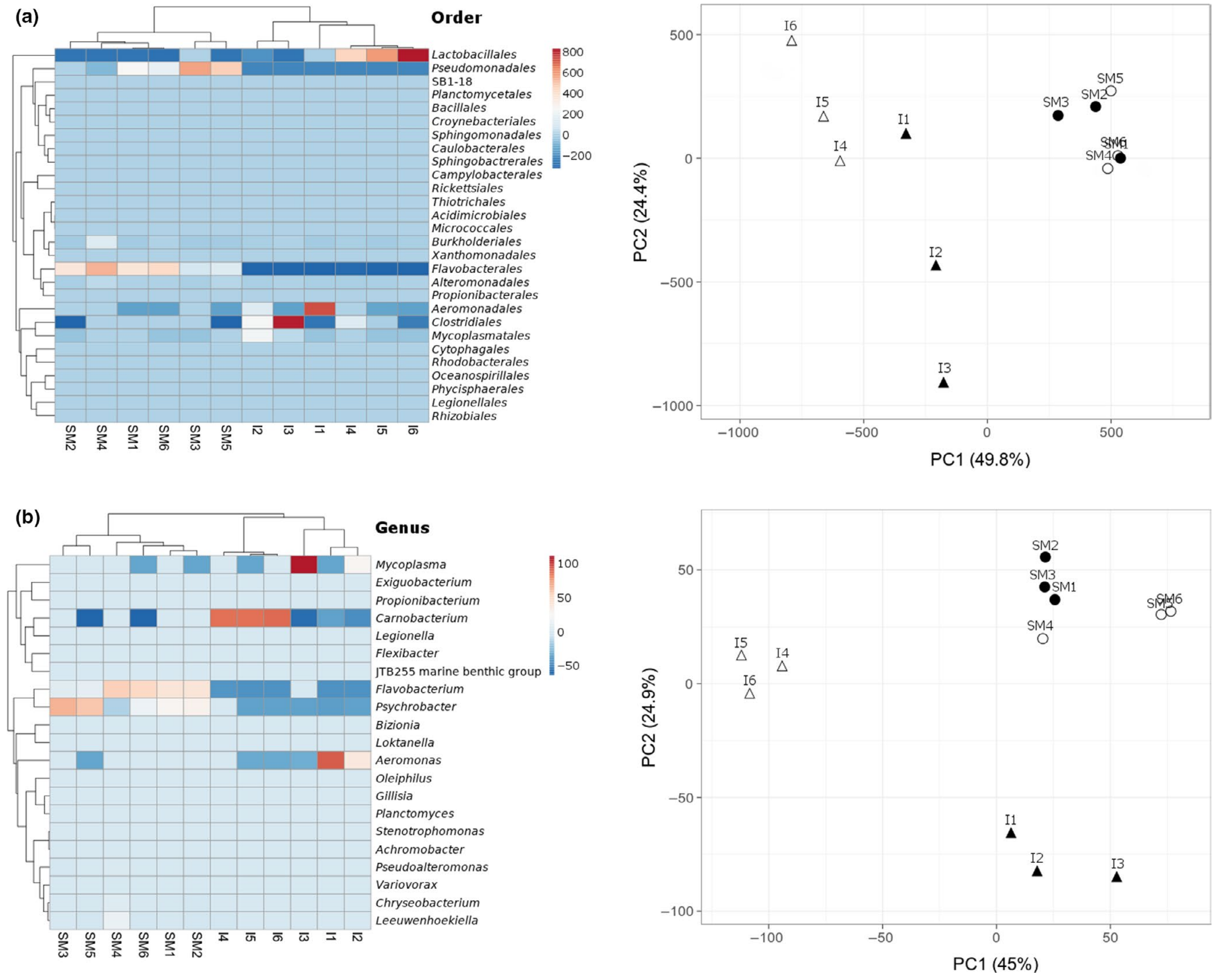


FIGURE 2 A hierarchically clustered heatmap of the microbial profiles of Atlantic skin mucus and intestine based on SILVA database taxonomy assigned to (a) order and (b) genus levels. Dendrogram at the top of the heatmap shows the clustering of the sample types, skin mucus (SM1-SM6), and intestine (I1-I6). The dendrogram at the left side shows the distribution of bacteria. The color scale depicts the normalized relative abundance of each rank level. Principal component analysis (PCA) plot of scaled microbiota profiles representing both skin mucus (circles) and intestinal samples (triangles) of Atlantic salmon is depicted in the far right for both order and genus levels. Filled (black) symbols for fish fed krill meal diet and unfilled (white) symbols for fish fed fish meal based diet

3.3 | Bacteria recovered by culture-dependent methods

The culture-based method aimed to provide representative isolates within anticipated genera to compare sequence information against the recovered CCS pool of sequences. Bacterial colonies were phenotypically categorized, and a total of 46 representative colonies were further identified and allocated to seven different genera based on the comparative 16S rRNA gene sequence alignment (Table A1). The relative distribution between colony morphologies on plates retrospectively identified by phylogeny is provided as an average of each sample type in Table A2, Appendix. Isolate information, sample origin, diet group, growth medium used and the most closely related type strains to each genus group is shown in Figure A1, Appendix. Skin mucus samples resulted in *Flavobacterium*, *Psychrobacter*, and *Exigubacterium* isolates on R2A plates, and *Pseudomonas* and *Shewanella* on MacConkey plates. *Carnobacterium* dominated intestinal samples on both media plates. Representative colonies aligning to either genus *Flavobacterium*, *Carnobacterium*, or *Exigubacterium* were identical at the 16S rRNA gene level. Variants within the 16S rRNA gene level were found for isolates within each genus *Chryseobacterium*, *Psychrobacter*, *Pseudomonas*, and *Shewanella*.

3.4 | Assignment of CCS reads to the bacterial isolates

The distribution of the 46 identified bacteria isolates among the pool of sequences is shown in Figure 4, where the displayed OTUs

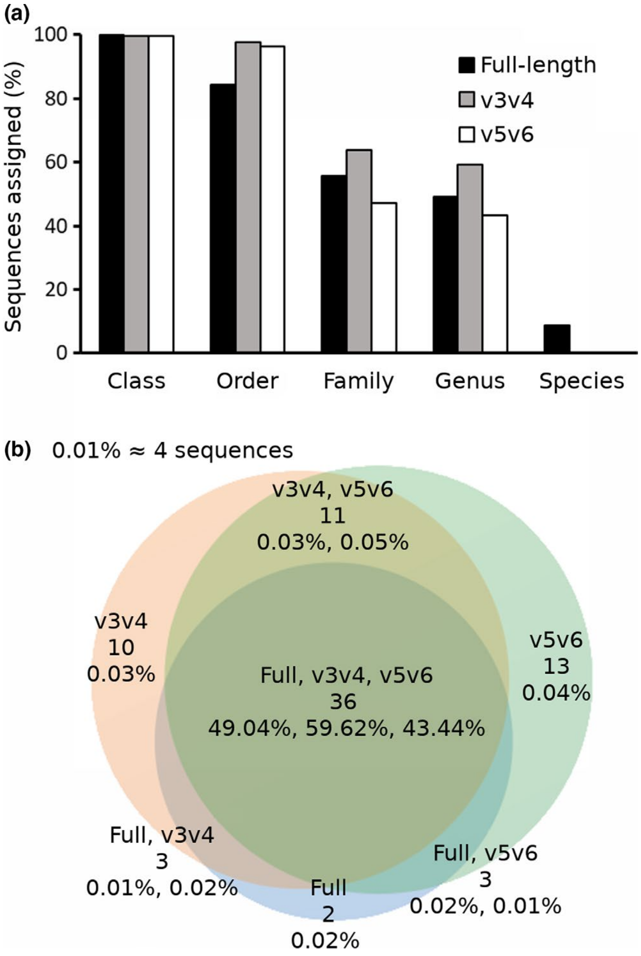


FIGURE 3 Percentage of assigned sequences at the class, order, family, genus, and species levels from full-length (black bars), v3v4 (gray bars), and v5v6 (white bars) datasets (a). Venn diagram of genus level assignments illustrating the number of taxa shared in and between each dataset (b). Top line shows dataset(s), middle line shows the number of unique genus taxa shared among dataset(s), bottom line shows the relative abundance of sequences assigned to the involved taxa and the given dataset(s)

TABLE 2 Assignments for pooled data of full-length 16S rRNA, v3v4 and v5v6 regions

Or. Fa. Ge. Sp.	Full	v3v4	v5v6
Aeromonadales	3,242	+14	+5
Aeromonadaceae	2,801	+369	+302
Aeromonas	2,598	+420	+407
Alteromonadales	260	+13	+22
Pseudoalteromonadaceae	126	+103	+93
Pseudoalteromonas	94	+118	+117
Bacillales	34	+170	+7
Burkholderiales	591	+35	+28
Alcaligenaceae	387	+171	+142
Clostridiales	8,163	+5,264	+5,157
Ruminococcaceae	0	+11	+176
Enterobacteriales	0	0	+72
Flavobacteriales	9,391	+119	−3
Flavobacteriaceae	7,877	+1,219	+1,117
Chryseobacterium	154	+155	+141
Chryseobacterium marinum	84	−84	−84
Flavobacterium	6,500	+1,281	+1,058
Flavobacterium frigidarium	3,802	−3,802	−3,802
Leeuwenhoekiella	256	+235	+187
Bizionia	19	+19	+24
Others	13	+31	+40
Lactobacillales	7,420	−83	−217
Carnobacteriaceae	6,306	+873	−6,283
Carnobacterium	5,630	+1,303	−5,615
Mycoplasmatales	955	+70	+61
Mycoplasmataceae	756	+217	+201
Mycoplasma	571	+251	+343
Pseudomonadales	6,953	+66	+74
Moraxellaceae	6,114	+619	+523
Psychrobacter	5,645	+756	+790
Rhodobacterales	44	+20	+7
Rhodobacteraceae	23	+32	+19
Sphingobacteriales	56	+16	+10
Saprospiraceae	23	+25	+26

Note: Sum of assignments where the sub-datasets of v-regions are displayed relative to the full dataset. Only assignments above 0.1% (>44 no. of sequences) of the total 43,910 sequences represented in each of the three datasets is shown.
Abbreviations: Fa., family; Ge., Genus; Or., order; Sp., Species.

in tree branches are representative CCS within each allocated taxon. As expected, not all taxa detected by sequencing are cultivatable under the conditions used. The proportion of cultivable bacteria was investigated by assigning 16S rRNA gene sequences from the obtained isolates to the full-length 16S rRNA CCS generated reads. Sequences (CCS reads) in cluster with 97% sequence identity to isolates accounted for 3.99% of the CCS reads, that is, cultivability of 3.99%. Comparing the relative abundance of OTUs assigned to the genus level shows domination of *Flavobacterium* and *Psychrobacter* in skin mucus and *Carnobacterium* in the intestinal samples. Two of the most abundant genera in the skin mucus, *Psychrobacter* and *Flavobacterium*, are represented by eight isolates, each (Figure 4). They are clearly distinctive to their respective branch, although the *Psychrobacter* appear to include subgroups that cannot be discriminated based on their 16S rRNA gene sequences (Figure A1). An exception within the tree is created by OTU 10 *Psychrobacter* where the CCS is placed between the genera *Psychrobacter* and *Flavobacterium*. Reads belonging to the genus *Exiguobacterium* was not detected in the intestine and is represented with ≤4 CCS reads in the skin mucus samples. *Shewanella* CCS reads are represented by ≤2 sequences in two samples. No CCS reads belong to the genus *Pseudomonas*.

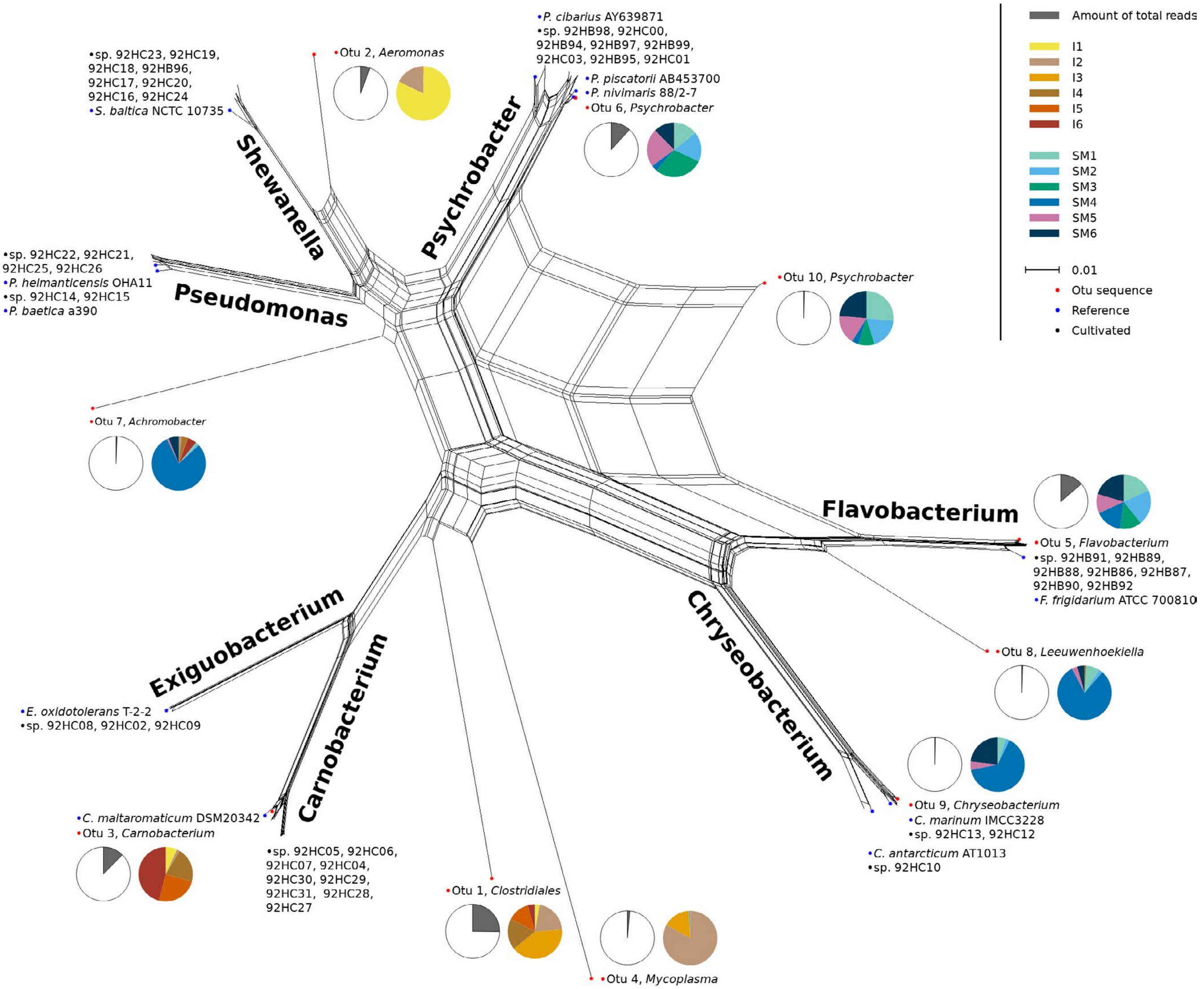


FIGURE 4 Phylogenetic relationships as NeighborNet network of representative OTUs (red), cultivated bacteria sequences (black), and reference strains (blue). Pie charts detail the given OTU by indicating the proportion of reassigned full-length 16S reads (gray) and how these are distributed from the samples (colored). The sample values are shown scaled to the minimal sample SM2. Scale bar shows distance as number of nucleotide substitutions per site

4 | DISCUSSION

The skin mucus microbial community was dominated by Proteobacteria within the genus *Flavobacterium* and *Psychrobacter*. This corroborates previous findings reporting the dominance of a few Proteobacteria-affiliated phylotypes in Atlantic salmon skin mucus from both controlled experiments (Lokesh & Kiron, 2016; Minniti et al., 2017) and commercial production systems (Karlsen et al., 2017). The *Flavobacterium* genus has several important fish pathogens, primarily found in freshwater hatchery-reared fish (Starliper, 2011), that may adversely affect Atlantic salmon (Loch & Faisal, 2015). *Psychrobacter* spp. are also associated with marine fish species (Ramírez & Romero, 2017a, 2017b; Småge, Frisch, Brevik, Watanabe, & Nylund, 2016). Prominent intestinal bacteria were of the order Clostridiales, Aeromonadales, and Lactobacillales, which are repeatedly reported from the Atlantic salmon intestine (Catalán, Villasante, Wacyk, Ramírez, & Romero, 2017; Gajardo, Jaramillo-Torres, et al., 2016; Llewellyn et al., 2015; Zarkasi et al., 2014). The number of fish in this trial makes it difficult to provide any conclusive insight into any dietary effect. Of note is the indication of more predominant *Carnobacterium* in the fish fed fish meal. This corroborates a previous study that also substituted fish meal with krill meal followed by culture-dependent techniques on gut microbiota (Ringø et al., 2006) where *Carnobacterium* was present in non-krillmeal-fed fish. A large proportion of the Clostridiales CCS 16S rRNA gene sequences could not be taxonomically assigned above the order rank, indicating the presence of so far uncharacterized bacteria.

Bacteria in the seawater column often inhabit a nutrient poor environment and a large number of the marine bacteria that occur in this environment are suggested to be nonculturable using standard culture-based techniques (Giovannoni & Stingl, 2005). Bacteria associated with the Atlantic salmon integument are also considered difficult to cultivate. This includes pathogens such as *Tenacibaculum* (Olsen et al., 2011) and *Moritella* (Grove et al., 2008), and culture-based diagnostics of Atlantic salmon are considered unreliable (Grove et al., 2008). Early studies reported relative high cultivability of bacteria from salmonid intestines (Huber et al., 2004; Spanggaard et al., 2000), but also inconsistencies between culture-dependent and molecular-based methods (Hovda, Lunestad, Fontanillas, & Rosnes, 2007). Our approached did not aim to assess culturability, but was applied to recover Atlantic salmon isolates for future studies. Dominant bacteria in the intestine within the order Clostridiales and OTUs assigned to *Aeromonas* and *Mycoplasma* were not isolated, likely due to the growth media and aerobic condition used. Still, an overall ratio of culturability of 3.99% was demonstrated based on the presence/absence of all assigned CCS 16S rRNA gene sequences corresponding to cultivated isolates with 97% sequence identity. This dropped to 0.13% at 99% sequence identity. Both the dominating genera *Flavobacterium* and *Psychrobacter* are studied for their degrading properties (Lasa & Romalde, 2017; Loch & Faisal, 2015), and *Carnobacterium* is isolated from a range of environments (Leisner, Laursen, Prévost, Drider, & Dalgaard, 2007). It is possible that these genera are well-adapted to grow on the laboratory cultivation media used in this study.

Like many genera associated with the Atlantic salmon, the phylogenetic assignment is often based on the 16S rRNA gene. However, there is a lack of properly defined bacterial species within the aquatic environment that in general hamper our ability to understand and organize bacterial diversity to these environments. There are also technological limitations that may impact on the ecological data description of these analyses (Schmidt, Matias, & Mering, 2015). Related to Atlantic salmon, many of the dominant groups or recovered isolates cannot be discriminated at the species level by their 16S rRNA gene sequences (Grove et al., 2010; Småge et al., 2015). Furthermore, our data suggest that the salmon contains more than one species within observed genera such as *Psychrobacter*. Covering the full-length sequences of 16S rRNA genes is expected to be advantageous for inferring phylogenetic affiliations and provide a more precise microbial community profiling of Atlantic salmon. However, our partial gene sequences were assigned in a higher abundance compared to the full-length sequences. Only at the species level was full-length sequences assigned in a higher proportion (mean 8.9% compared to 0.01% and 0.005% for v3v4 and v5v6 data subsets, respectively). Effects on abundance profiling and discrepancy with an overestimation in bacterial diversity between full-length 16S rRNA gene and partial variable datasets are in similar accordance with other studies (Singer et al., 2016; Sun et al., 2013; Wagner et al., 2016; Whon et al., 2018). The taxonomic resolution in amplicon sequencing might be affected by several factors. In our study, where we generated different sequence lengths in silico one possibility is differences of the intravariability within each targeted hypervariable region (v3v4 and v5v6) in comparison with the full-length sequence (Kumar, Brooker, Dowd, & Camerlengo, 2011). Another influencing factor could be the reference sequences in the choice of database used (Werner et al., 2012). Evaluating the species richness, we could not fully describe the community in all samples. The proportion of pooled sequences and the sample sizes obtained can be contributing factors to this lack of convergence where potentially important taxa have not been identified. Using the pooled Atlantic salmon community, we observed discrepancy in community structure and phylogenetic resolution across multiple taxonomic levels between full-length and partial sequences. Differences were revealed more clearly in some of the phylogenetic lineages. In our data, especially Clostridiales at the order level but also genera within Flavobacteriaceae and Carnobacteriaceae. This highlights that taxonomic affiliation using the bacterial 16S rRNA gene should be concluded with care. The 36 genera shared among the full-length and partial data subsets compose most of the assigned sequences (49.04%, 59.62%, and 43.44% of the full-length, and v3v4 and v5v6 data subsets, respectively) and greatly outnumber the proportion of assigned sequences in the 42 ancillary genera identified from one or two of the combined datasets. This high number of genera that is derived from a small part of the sequence sets might indicate that some genera represent false positives, caused by the shorter stretches of the 16S rRNA linked to a lack of resolution in the sequence regions. Collectively, these data suggest that sequence length and part of variable region used on the 16S rRNA gene will influence the outcome of the obtained microbial profile, in a way that is different between taxonomic rank and phylogenetic group. Although this study suggests confounding effects when using different variable regions of the 16S rRNA gene to characterize microbial profiles, it also has key limitations. The sample size should have been increased to better allow robust assessment of experimental effects. The microbiota in context to associations and interactions in animal trials are likely to be complex. Increasing the number of samples may better take in considerations concerning factors such as individual effects caused by husbandry or disease status (Moore & Stanley, 2016). Methodological errors or inadequacy to lyse and extract DNA from bacterial cells between tissue types have also a potential for introducing biases into the results. In addition, the taxonomic resolution in amplicon sequencing might be affected by several additional factors such as PCR conditions used to amplify the product (Lorenz, 2012), primer specificity (Beckers et al., 2016; Tremblay et al., 2015) and possible contamination in low microbial biomass samples (Eisenhofer et al., 2019). To further narrow the gaps and identify the best approach for microbial profiling of Atlantic salmon, future studies should compare primer sets targeting different sequence regions combined with sequencing technology platforms using several tissue types and DNA from a mock community containing a known number of species.

By comparing the distribution of representative OTUs to 16S rRNA gene sequences from cultivated isolates, we aimed to asses any anomaly or difference in the taxonomic classification. The sequences were clearly assigned down to the genus level, except OTU 10 which was placed between the genera *Psychrobacter* and *Flavobacterium*. An interesting observation is that *Exiguobacterium* and *Shewanella* are observed in relative high abundance by the cultivation method compared to the extraction of DNA. *Pseudomonas* was at most represented by five sequences in the pooled v3v4 data subset of the generated 16S rRNA CCS reads. Methodological errors or inadequacy to lyse and extract DNA from bacterial cells, primer specificity, or PCR conditions could be possible explanations. However, because all cultivable taxa were detectable with the nonbarcoded version of the primers it is unlikely to account for an almost complete absence of a genus. Another explanation is that the sequencing depth may have been too low as indicated by the rarefaction analysis. Bacteria not detected by molecular methods may also be site or human-derived plate contaminants or cultivation procedures may have facilitated growth of these bacteria. The phenomenon of cultivable isolates not being detected in corresponding gene libraries are reported from a wide distribution of marine sample types such as seawater (Eilers, Pernthaler, Glöckner, & Amann, 2000), sponges (Esteves et al., 2016), and algae (KleinJan et al., 2017).

In this study, both skin mucus and intestinal samples were successfully utilized to generate full-length 16S rRNA gene sequences by the PacBio CCS technology. A high proportion of reads from skin surface samples were allocated down to the level genus. In contrast, intestinal samples dominated by reads assigned to the order Clostridiales were lost at higher resolutions. This highlights the need to further expand the microbial 16S rRNA gene catalogue from underrepresented marine taxa into reference databases. Our data also suggest that different variable regions and sequence length of the 16S rRNA gene will influence the microbial profile differently by taxonomic rank and phylogenetic group. To identify the best taxonomic profiling effectiveness between different variable regions and full-length 16S rRNA gene would need further validation. However, at present, one of the bottlenecks in comparing microbial profiles is due to the different 16S rRNA gene regions used in different studies. Using the full-length 16S rRNA gene sequence has the potential to become a tool for more precise microbial community profiling that better allows global comparisons of microbiome studies.

ACKNOWLEDGMENTS

This study was supported by UiT The Arctic University of Norway and through the Research Council of Norway (RCN) by the Nofima basic research grant (194050/F40). The publication charges for this article have been funded by a grant from the publication fund of UiT The Arctic University of Norway.

CONFLICT OF INTERESTS

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

CK initiated the study and performed the laboratory work. TK conducted and performed the bioinformatic analyses. NW coordinated the work. CK and TK drafted the manuscript and all authors provided critical feedback and helped shape the research, analysis, and manuscript.

ETHICS STATEMENT

Fish in this study was based on post mortem sampling of material from fish harvested from a different industry research study for other purposes. Fish at the Research Station for Sustainable Aquaculture (Sunndalsøra, Norway) was reared by trained professionals in accordance with guidelines and regulations of the Norwegian Food Safety Authority.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed in this study are available in the BioProject repository at https://www.ncbi.nlm.nih.gov/bioproject/PRJEB28410.

ORCID

Terje Klemetsen https://orcid.org/0000-0002-2024-1798

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Beckers, B., Op De Beeck, M., Thijs, S., Truyens, S., Weyens, N., Boerjan, W., & Vangronsveld, J.. (2016). Performance of 16s rDNA Primer Pairs in the Study of Rhizosphere and Endosphere Bacterial Microbiomes in Metabarcoding Studies. *Frontiers in Microbiology*, *7*, 650. https://doi.org/10.3389/fmicb.2016.00650

Beklioglu, M., Telli, M., & Gozen, A. G. (2006). Fish and mucus-dwelling bacteria interact to produce a kairomone that induces diel vertical migration in *Daphnia*. *Freshwater Biology*, *51*(12), 2200–2206. https://doi.org/10.1111/j.1365-2427.2006.01642.x

Bertani, G. (1951). Studies on lysogenesis I. : The mode of phage liberation by lysogenic *Escherichia coli*. *Journal of Bacteriology*, *62*(3), 293–300.

Catalán, N., Villasante, A., Wacyk, J., Ramírez, C., & Romero, J. (2017). Fermented Soybean Meal Increases Lactic Acid Bacteria in Gut Microbiota of Atlantic Salmon (Salmo salar). *Probiotics and Antimicrobial Proteins*, *10*(3), 566–576. https://doi.org/10.1007/s12602-017-9366-7

Chen, Z., Hui, P. C., Hui, M., Yeoh, Y. K., Wong, P. Y., Chan, M. C. W., … Chan, P. K. S. (2019). Impact of preservation method and 16S rRNA hypervariable region on gut microbiota profiling. *mSystems*, *4*(1), e00271–e00318. https://doi.org/10.1128/mSystems.00271-18

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., … Tiedje, J. M. (2014). Ribosomal database project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*(D1), D633–D642. https://doi.org/10.1093/nar/gkt1244

Dehler, C. E., Secombes, C. J., & Martin, S. A. M. (2017). Environmental and physiological factors shape the gut microbiota of Atlantic salmon parr (*Salmo salar* L.). *Aquaculture (Amsterdam, Netherlands)*, *467*, 149–157. https://doi.org/10.1016/j.aquaculture.2016.07.017

Dhariwal, A., Xia, J., Chong, J., Agellon, L. B., Habib, S., & King, I. L. (2017). MicrobiomeAnalyst: A web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic*

Acids Research, *45*(W1), W180–W188. https://doi.org/10.1093/nar/gkx295

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461.

Eilers, H., Pernthaler, J., Glöckner, F. O., & Amann, R. (2000). Culturability and in situ abundance of pelagic bacteria from the North sea. *Applied and Environmental Microbiology*, *66*(7), 3044–3051. https://doi.org/10.1128/AEM.66.7.3044-3051.2000 https://doi.org/10.1128/AEM.66.7.3044-3051.2000

Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R., & Weyrich, L. S. (2019). Contamination in low microbial biomass microbiome studies: Issues and recommendations. *Trends in Microbiology*, *27*(2), 105–117. https://doi.org/10.1016/j.tim.2018.11.003

Esteves, A. I. S., Amer, N., Nguyen, M., & Thomas, T. (2016). Sample processing impacts the viability and cultivability of the sponge microbiome. *Frontiers in Microbiology*, *7*, 499. https://doi.org/10.3389/fmicb.2016.00499

Gajardo, K., Jaramillo-Torres, A., Kortner, T. M., Merrifield, D. L., Tinsley, J., Bakke, A. M., & Krogdahl, Å. (2016). Alternative protein sources in the diet modulate microbiota and functionality in the distal intestine of Atlantic salmon (*Salmo salar*). *Applied and Environmental Microbiology*, *83*(5), e02615–e02616. https://doi.org/10.1128/aem.02615-16

Gajardo, K., Rodiles, A., Kortner, T. M., Krogdahl, Å., Bakke, A. M., Merrifield, D. L., & Sørum, H. (2016). A high-resolution map of the gut microbiota in Atlantic salmon (*Salmo salar*): A basis for comparative gut microbial research. *Scientific Reports*, *6*, 30893. https://doi.org/10.1038/srep30893

Giovannoni, S. J., & Stingl, U. (2005). Molecular diversity and ecology of microbial plankton. *Nature*, *437*, 343. https://doi.org/10.1038/nature04158

Grove, S., Reitan, L. J., Lunder, T., & Colquhoun, D. (2008). Real-time PCR detection of *Moritella viscosa*, the likely causal agent of winter-ulcer in Atlantic salmon *Salmo salar* and rainbow trout *Oncorhynchus mykiss*. *Diseases of Aquatic Organisms*, *82*(2), 105–109. https://doi.org/10.3354/dao01972

Grove, S., Wiik-Nielsen, C. R., Lunder, T., Tunsjø, H. S., Tandstad, N. M., Reitan, L. J., ... Colquhoun, D. J. (2010). Previously unrecognised division within *Moritella viscosa* isolated from fish farmed in the North Atlantic. *Diseases of Aquatic Organisms*, *93*(1), 51–61. https://doi.org/10.3354/dao02271

Hall, T. A. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95–98.

Hovda, M. B., Lunestad, B. T., Fontanillas, R., & Rosnes, J. T. (2007). Molecular characterisation of the intestinal microbiota of farmed Atlantic salmon (*Salmo salar* L.). *Aquaculture*, *272*, 581–588. https://doi.org/10.1016/j.aquaculture.2007.08.045

Huber, I., Spanggaard, B., Appel, K. F., Rossen, L., Nielsen, T., & Gram, L. (2004). Phylogenetic analysis and *in situ* identification of the intestinal microbial community of rainbow trout (*Oncorhynchus mykiss*, Walbaum). *Journal of Applied Microbiology*, *96*(1), 117–132. https://doi.org/10.1046/j.1365-2672.2003.02109.x

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*(2), 254–267. https://doi.org/10.1093/molbev/msj030

Karlsen, C., Ellingsen, A. B., Wiik-Nielsen, C., Winther-Larsen, H. C., Colquhoun, D., & Sørum, H. (2014). Host specificity and clade dependent distribution of putative virulence genes in *Moritella viscosa*. *Microbial Pathogenesis*, *77*, 53–65. https://doi.org/10.1016/j.micpath.2014.09.014

Karlsen, C., Ottem, K. F., Brevik, Ø. J., Davey, M., Sørum, H., & Winther-Larsen, H. C. (2017). The environmental and host-associated bacterial microbiota of Arctic seawater-farmed Atlantic salmon with ulcerative disorders. *Journal of Fish Diseases*, *40*(11), 1645–1663. https://doi.org/10.1111/jfd.12632

KleinJan, H., Jeanthon, C., Boyen, C., & Dittami, S. M. (2017). Exploring the cultivable *Ectocarpus* microbiome. *Frontiers in Microbiology*, *8*, 2456. https://doi.org/10.3389/fmicb.2017.02456

Kumar, P. S., Brooker, M. R., Dowd, S. E., & Camerlengo, T. (2011). Target region selection is a critical determinant of community fingerprints generated by 16S pyrosequencing. *PLoS ONE*, *6*(6), e20956. https://doi.org/10.1371/journal.pone.0020956

Lanzén, A., Jørgensen, S. L., Huson, D. H., Gorfer, M., Grindhaug, S. H., Jonassen, I., ... Urich, T. (2012). CREST – Classification Resources for Environmental Sequence Tags. *PLoS ONE*, *7*(11), e49334. https://doi.org/10.1371/journal.pone.0049334

Larsen, A., Bullard, S., Womble, M., & Arias, C. (2015). Community structure of skin microbiome of Gulf killifish, *Fundulus grandis*, is driven by seasonality and not exposure to oiled sediments in a Louisiana salt marsh. *Microbial Ecology*, *70*(2), 534–544. https://doi.org/10.1007/s00248-015-0578-7

Lasa, A., & Romalde, J. L. (2017). Genome sequence of three *Psychrobacter* sp. strains with potential applications in bioremediation. *Genomics Data*, *12*, 7–10. https://doi.org/10.1016/j.gdata.2017.01.005

Leisner, J. J., Laursen, B. G., Prévost, H., Drider, D., & Dalgaard, P. (2007). *Carnobacterium*: Positive and negative effects in the environment and in foods. *Fems Microbiology Reviews*, *31*(5), 592–613. https://doi.org/10.1111/j.1574-6976.2007.00080.x

Llewellyn, M. S., McGinnity, P., Dionne, M., Letourneau, J., Thonier, F., Carvalho, G. R., ... Derome, N. (2015). The biogeography of the atlantic salmon (*Salmo salar*) gut microbiome. *ISME Journal*, *10*(5), 1280–1284. https://doi.org/10.1038/ismej.2015.189

Loch, T. P., & Faisal, M. (2015). Emerging flavobacterial infections in fish: A review. *Journal of Advanced Research*, *6*(3), 283–300. https://doi.org/10.1016/j.jare.2014.10.009

Lokesh, J., & Kiron, V. (2016). Transition from freshwater to seawater reshapes the skin-associated microbiota of Atlantic salmon. *Scientific Reports*, *6*, 19707. https://doi.org/10.1038/srep19707

Lorenz, T. C. (2012). Polymerase chain reaction: Basic protocol plus troubleshooting and optimization strategies. *Journal of Visualized Experiments*, *63*, e3998–e3998. https://doi.org/10.3791/3998

Metsalu, T., & Vilo, J. (2015). ClustVis: A web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Research*, *43*(W1), W566–W570. https://doi.org/10.1093/nar/gkv468

Minniti, G., Hagen, L. H., Porcellato, D., Jørgensen, S. M., Pope, P. B., & Vaaje-Kolstad, G. (2017). The skin-mucus microbial community of farmed Atlantic salmon (*Salmo salar*). *Frontiers in Microbiology*, *8*, 2043. https://doi.org/10.3389/fmicb.2017.02043

Moore, R. J., & Stanley, D. (2016). Experimental design considerations in microbiota/inflammation studies. *Clinical & Translational Immunology*, *5*(7), e92–e92. https://doi.org/10.1038/cti.2016.41

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., & Schäffer, A. A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics*, *24*(16), 1757–1764. https://doi.org/10.1093/bioinformatics/btn322

Nayak, S. K. (2010). Role of gastrointestinal microbiota in fish. *Aquaculture Research*, *41*(11), 1553–1573. https://doi.org/10.1111/j.1365-2109.2010.02546.x

Olsen, A. B., Nilsen, H., Sandlund, N., Mikkelsen, H., Sørum, H., & Colquhoun, D. J. (2011). *Tenacibaculum* sp. associated with winter ulcers in sea-reared Atlantic salmon *Salmo salar. Diseases of Aquatic Organisms*, *94*(3), 189–199.

Pootakham, W., Mhuantong, W., Yoocha, T., Putchim, L., Sonthirod, C., Naktang, C., ... Tangphatsornruang, S. (2017). High resolution profiling of coral-associated bacterial communities using full-length 16S rRNA sequence data from PacBio SMRT sequencing system. *Scientific Reports*, *7*(1), 2774. https://doi.org/10.1038/s41598-017-03139-4

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glockner, F. O. (2007). SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, *35*(21), 7188–7196. https://doi.org/10.1093/nar/gkm864

Ramírez, C., & Romero, J. (2017a). Fine flounder (Paralichthys adspersus) microbiome showed important differences between wild and reared specimens. *Frontiers in Microbiology*, *08*, 271. https://doi.org/10.3389/fmicb.2017.00271

Ramírez, C., & Romero, J. (2017b). The Microbiome of Seriola lalandi of Wild and Aquaculture Origin Reveals Differences in Composition and Potential Function. *Frontiers in Microbiology*, *8*, 1844. https://doi.org/10.3389/fmicb.2017.01844

Ringø, E., Sperstad, S., Myklebust, R., Mayhew, T. M., Mjelde, A., Melle, W., & Olsen, R. E. (2006). The effect of dietary krill supplementation on epithelium-associated bacteria in the hindgut of Atlantic salmon (*Salmo salar* L.): A microbial and electron microscopical study. *Aquaculture Research*, *37*(16), 1644–1653. https://doi.org/10.1111/j.1365-2109.2006.01611.x

Schmidt, T. S. B., Matias, R. J. F., & Mering, C. (2015). Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environmental Microbiology*, *17*(5), 1689–1706. https://doi.org/10.1111/1462-2920.12610

Schmidt, V., Amaral-Zettler, L., Davidson, J., Summerfelt, S., & Good, C. (2016). Influence of fishmeal-free diets on microbial communities in Atlantic salmon (*Salmo salar*) recirculation aquaculture systems. *Applied and Environmental Microbiology*, *82*(15), 4470–4481. https://doi.org/10.1128/aem.00902-16

Schmidt, V. T., Smith, K. F., Melvin, D. W., & Amaral-Zettler, L. A. (2015). Community assembly of a euryhaline fish microbiome during salinity acclimation. *Molecular Ecology*, *24*(10), 2537–2550. https://doi.org/10.1111/mec.13177

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863–864. https://doi.org/10.1093/bioinformatics/btr026

Sepúlveda, D., Bohle, H., Labra, Á., Grothusen, H., & Marshall, S. H. (2013). Design and evaluation of a unique RT-qPCR assay for diagnostic quality control assessment that is applicable to pathogen detection in three species of salmonid fish. *BMC Veterinary Research*, *9*, 183–183. https://doi.org/10.1186/1746-6148-9-183

Sievers, F., & Higgins, D. G. (2014). Clustal omega, accurate alignment of very large numbers of sequences. In D. J. Russell (Ed.), *Multiple Sequence Alignment Methods* (pp. 105–116). Totowa, NJ: Humana Press.

Singer, E., Bushnell, B., Coleman-Derr, D., Bowman, B., Bowers, R. M., Levy, A., … Woyke, T. (2016). High-resolution phylogenetic microbial community profiling. *The Isme Journal*, *10*, 2020. https://doi.org/10.1038/ismej.2015.249

Småge, S. B., Brevik, Ø. J., Duesund, H., Ottem, K. F., Watanabe, K., & Nylund, A. (2015). Tenacibaculum finnmarkense sp. nov., a fish pathogenic bacterium of the family Flavobacteriaceae isolated from Atlantic salmon. *Antonie Van Leeuwenhoek*, *109*(2), 273–285. https://doi.org/10.1007/s10482-015-0630-0

Småge, S. B., Frisch, K., Brevik, Ø. J., Watanabe, K., & Nylund, A. (2016). First isolation, identification and characterisation of *Tenacibaculum maritimum* in Norway, isolated from diseased farmed sea lice cleaner fish *Cyclopterus lumpus* L. *Aquaculture*, *464*, 178–184. https://doi.org/10.1016/j.aquaculture.2016.06.030

Spanggaard, B., Huber, I., Nielsen, J., Nielsen, T., Appel, K. F., & Gram, L. (2000). The microflora of rainbow trout intestine: A comparison of traditional and molecular identification. *Aquaculture*, *182*(1), 1–15. https://doi.org/10.1016/S0044-8486(99)00250-1

Starliper, C. E. (2011). Bacterial coldwater disease of fishes caused by *Flavobacterium psychrophilum*. *Journal of Advanced Research*, *2*(2), 97–108. https://doi.org/10.1016/j.jare.2010.04.001

Sullam, K. E., Essinger, S. D., Lozupone, C. A., O'Connor, M. P., Rosen, G. L., Knight, R., & Kilham, S. S. (2012). Environmental and ecological factors that shape the gut bacterial communities of fish: A meta-analysis. *Molecular Ecology*, *21*(13), 3363–3378. https://doi.org/10.1111/j.1365-294X.2012.05552.x

Sun, D.-L., Jiang, X., Wu, Q. L., & Zhou, N.-Y. (2013). Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity. *Applied and Environmental Microbiology*, *79*(19), 5962–5969. https://doi.org/10.1128/aem.01282-13

Sveen, L. R., Grammes, F. T., Ytteborg, E., Takle, H., & Jørgensen, S. M. (2017). Genome-wide analysis of Atlantic salmon (*Salmo salar*) mucin genes and their role as biomarkers. *PLoS ONE*, *12*(12), e0189103. https://doi.org/10.1371/journal.pone.0189103

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, *30*(12), 2725–2729. https://doi.org/10.1093/molbev/mst197

Travers, K. J., Chin, C.-S., Rank, D. R., Eid, J. S., & Turner, S. W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, *38*(15), e159–e159. https://doi.org/10.1093/nar/gkq543

Tremblay, J., Singh, K., Fern, A., Kirton, E. S., He, S., Woyke, T., … Tringe, S. G. (2015). Primer and platform effects on 16S rRNA tag sequencing. *Frontiers in Microbiology*, *6*, 771–771. https://doi.org/10.3389/fmicb.2015.00771

Wagner, J., Coupland, P., Browne, H. P., Lawley, T. D., Francis, S. C., & Parkhill, J. (2016). Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. *BMC Microbiology*, *16*(1), 274. https://doi.org/10.1186/s12866-016-0891-4

Walker, A. W., Martin, J. C., Scott, P., Parkhill, J., Flint, H. J., & Scott, K. P. (2015). 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome*, *3*(1), 26. https://doi.org/10.1186/s40168-015-0087-4

Werner, J. J., Koren, O., Hugenholtz, P., DeSantis, T. Z., Walters, W. A., Caporaso, J. G., … Ley, R. E. (2012). Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *The ISME Journal*, *6*(1), 94–103. https://doi.org/10.1038/ismej.2011.82

Whon, T. W., Chung, W.-H., Lim, M. Y., Song, E.-J., Kim, P. S., Hyun, D.-W., … Nam, Y.-D. (2018). The effects of sequencing platforms on phylogenetic resolution in 16 S rRNA gene profiling of human feces. *Scientific Data*, *5*, 180068–180068. https://doi.org/10.1038/sdata.2018.68 https://doi.org/10.1038/sdata.2018.68

Zarkasi, K. Z., Abell, G. C. J., Taylor, R. S., Neuman, C., Hatje, E., Tamplin, M. L., … Bowman J. P. (2014). Pyrosequencing-based characterization of gastrointestinal bacteria of Atlantic salmon (*Salmo salar* L.) within a commercial mariculture system. *Journal of Applied Microbiology*, *117*(1), 18–27.

How to cite this article: Klemetsen T, Willassen NP, Karlsen CR. Full-length 16S rRNA gene classification of Atlantic salmon bacteria and effects of using different 16S variable regions on community structure analysis. *MicrobiologyOpen*. 2019;e898. https://doi.org/10.1002/mbo3.898

APPENDIX A

TABLE A1 Novel bacterial strains isolated in this study presented with reconstructed phylogenetic genus groups, 16S accession number, isolation source, and culture medium

Isolates	GenBank acc. no.	Tissue	Culture medium
Flavobacterium sp. 92HB86	MG263463	Skin	R2A
Flavobacterium sp. 92HB87	MG263464	Skin	R2A
Flavobacterium sp. 92HB88	MG263465	Skin	R2A
Flavobacterium sp. 92HB89	MG263466	Skin	R2A
Flavobacterium sp. 92HB90	MG263467	Intestine	R2A
Flavobacterium sp. 92HB91	MG263468	Intestine	R2A
Flavobacterium sp. 92HB92	MG263469	Gill	R2A
Flavobacterium sp. 92HB93	MG263470	Gill	R2A
Psychrobacter sp. 92HB94	MG263471	Gill	R2A
Psychrobacter sp. 92HB95	MG263472	Gill	R2A
Psychrobacter sp. 92HB97	MG263473	Skin	R2A
Psychrobacter sp. 92HB98	MG263474	Skin	R2A
Psychrobacter sp. 92HB99	MG263475	Skin	R2A
Psychrobacter sp. 92HC00	MG263476	Skin	R2A
Psychrobacter sp. 92HC01	MG263477	Intestine	R2A
Psychrobacter sp. 92HC03	MG263478	Skin	R2A
Exiguobacterium sp. 92HC02	MG263480	Skin	R2A
Exiguobacterium sp. 92HC08	MG263481	Gill	R2A
Exiguobacterium sp. 92HC09	MG263482	Gill	R2A
Chryseobacterium sp. 92HC10	MG263487	Gill	R2A
Chryseobacterium sp. 92HC11	MG263488	Gill	R2A
Chryseobacterium sp. 92HC12	MG263489	Gill	R2A
Chryseobacterium sp. 92HC13	MG263490	Gill	R2A
Pseudomonas sp. 92HC14	MG263491	Skin	MacConkey
Pseudomonas sp. 92HC15	MG263492	Skin	MacConkey
Pseudomonas sp. 92HC21	MG263493	Skin	MacConkey
Pseudomonas sp. 92HC22	MG263494	Skin	MacConkey
Pseudomonas sp. 92HC25	MG263495	Skin	MacConkey
Pseudomonas sp. 92HC26	MG263496	Gill	MacConkey
Shewanella sp. 92HB96	MG263479	Water	R2A
Shewanella sp. 92HC16	MG263497	Skin	MacConkey
Shewanella sp. 92HC24	MG263498	Skin	MacConkey
Shewanella sp. 92HC17	MG263499	Skin	MacConkey
Shewanella sp. 92HC18	MG263500	Skin	MacConkey
Shewanella sp. 92HC19	MG263501	Skin	MacConkey
Shewanella sp. 92HC20	MG263502	Intestine	MacConkey
Shewanella sp. 92HC23	MG263503	Skin	MacConkey
Carnobacterium sp. 92HC27	MG263504	Intestine	MacConkey
Carnobacterium sp. 92HC28	MG263505	Intestine	MacConkey
Carnobacterium sp. 92HC29	MG263506	Intestine	MacConkey
Carnobacterium sp. 92HC30	MG263507	Intestine	MacConkey
Carnobacterium sp. 92HC31	MG263508	Intestine	MacConkey
Carnobacterium sp. 92HC04	MG263483	Intestine	R2A
Carnobacterium sp. 92HC05	MG263484	Intestine	R2A
Carnobacterium sp. 92HC06	MG263485	Intestine	R2A
Carnobacterium sp. 92HC07	MG263486	Intestine	R2A

TABLE A2 Overview of the most abundant bacteria assigned to the genus level present in Atlantic salmon sampled skin mucus and intestine. Their relative abundance across sample types are obtained either from sequence-based profiles (CCS reads) or from culture depended method assessing colony morphologies. Values are relative abundance in percentage, mean ± *SD*

	Extracted DNA				Bacteria on culture medium							
Tissue	**Skin**		**Intestine**		**Skin**				**Intestine**			
Diet	**Fish meal**	**Krill meal**	**Fish meal**	**Krill meal**	**Fish meal**		**Krill meal**		**Fish meal**		**Krill meal**	
Growth medium					**R2A**	**MC**	**R2A**	**MC**	**R2A**	**MC**	**R2A**	**MC**
Flavobacterium	50.5 ± 16.1	46.7 ± 12.3	0.9 ± 0.3	1.0 ± 1.1	54.5 ± 29.9		82.3 ± 8.1		0.5 ± 0.6			
Psychrobacter	34.1 ± 20.6	49.6 ± 14.7	0.7 ± 0.4	0.5 ± 0.5	45.5 ± 29.8		17.0 ± 7.0		0.3 ± 0.6			
Carnobacterium	0.3 ± 0.1	0.2 ± 0.2	97.4 ± 0.5	11.0 ± 8.3					99.1 ± 0.3	99.5 ± 0.7	100	100
Aeromonas	0.1 ± 0.2	0.1 ± 0.1	0.3 ± 0.3	44.0 ± 38.9								
Chryseobacterium	8.7 ± 0.7	2.4 ± 1.9	0.2 ± 0.1	0.1 ± 0.2								
Mycoplasma	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.1	42.9 ± 45.1								
Leeuwenhoekiella	4.0 ± 6.2	0.4 ± 0.5	0.1 ± 0.1									
Pseudoalteromonas	1.2 ± 0.6	0.3 ± 0.3										
Pseudomonas						2.4 ± 2.1		28.0 ± 11.3				
Shewanella						97.6 ± 2.1		72.0 ± 11.3		0.5 ± 0.7		
Exiguobacterium	0.1 ± 0.0						0.8 ± 1.1					

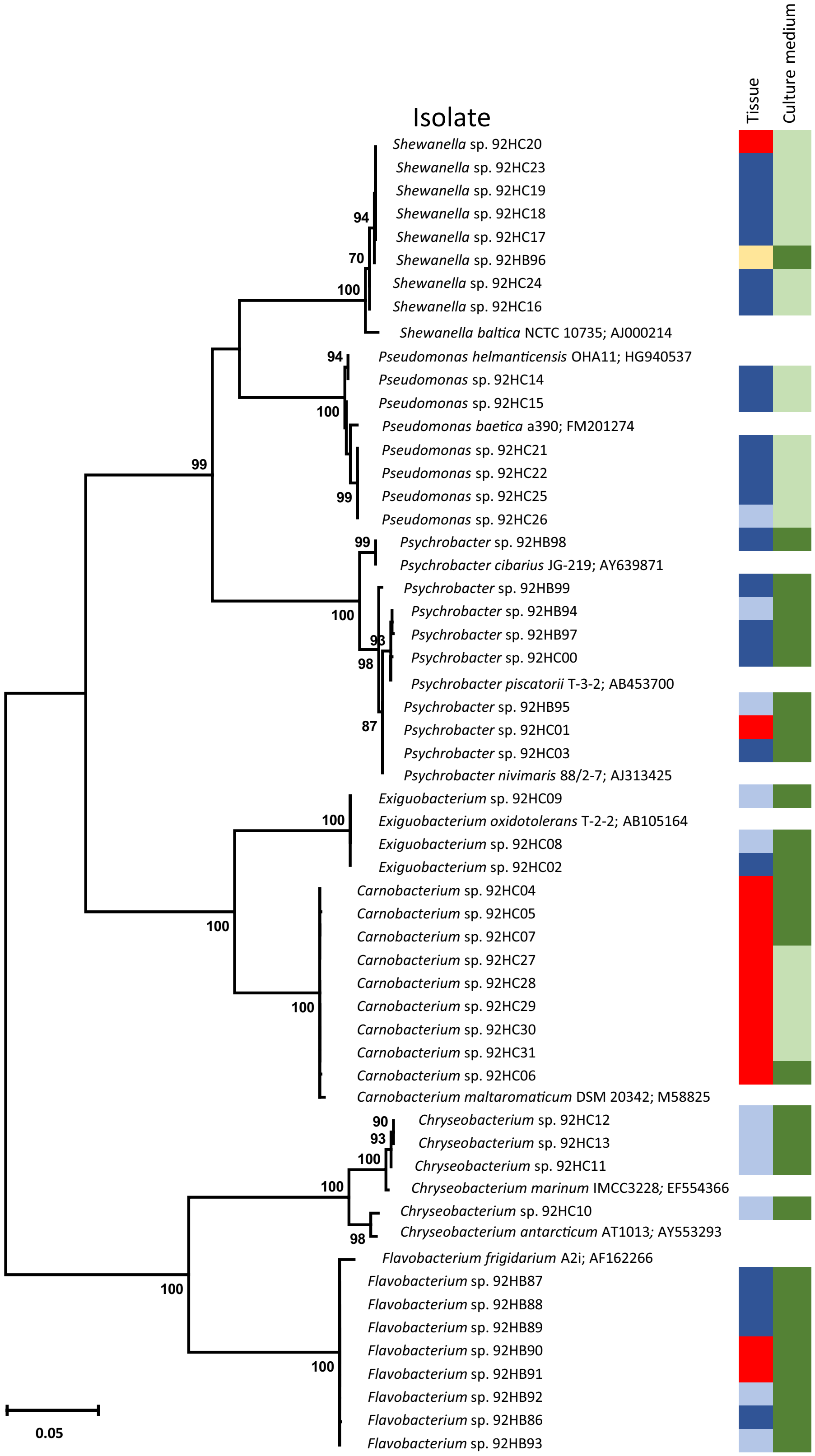

Tissue
Culture medium
Isolate
Shewanella sp. 92HC20
Shewanella sp. 92HC23
Shewanella sp. 92HC19
94
Shewanella sp. 92HC18
Shewanella sp. 92HC17
70
Shewanella sp. 92HB96
100
Shewanella sp. 92HC24
Shewanella sp. 92HC16
Shewanella baltica NCTC 10735; AJ000214
94
Pseudomonas helmanticensis OHA11; HG940537
Pseudomonas sp. 92HC14
100
Pseudomonas sp. 92HC15
Pseudomonas baetica a390; FM201274
99
Pseudomonas sp. 92HC21
Pseudomonas sp. 92HC22
99
Pseudomonas sp. 92HC25
Pseudomonas sp. 92HC26
99
Psychrobacter sp. 92HB98
Psychrobacter cibarius JG-219; AY639871
Psychrobacter sp. 92HB99
100
Psychrobacter sp. 92HB94
93
Psychrobacter sp. 92HB97
98
Psychrobacter sp. 92HC00
Psychrobacter piscatorii T-3-2; AB453700
Psychrobacter sp. 92HB95
87
Psychrobacter sp. 92HC01
Psychrobacter sp. 92HC03
Psychrobacter nivimaris 88/2-7; AJ313425
Exiguobacterium sp. 92HC09
100
Exiguobacterium oxidotolerans T-2-2; AB105164
Exiguobacterium sp. 92HC08
Exiguobacterium sp. 92HC02
Carnobacterium sp. 92HC04
Carnobacterium sp. 92HC05
100
Carnobacterium sp. 92HC07
Carnobacterium sp. 92HC27
Carnobacterium sp. 92HC28
100
Carnobacterium sp. 92HC29
Carnobacterium sp. 92HC30
Carnobacterium sp. 92HC31
Carnobacterium sp. 92HC06
Carnobacterium maltaromaticum DSM 20342; M58825
90
Chryseobacterium sp. 92HC12
93
Chryseobacterium sp. 92HC13
100
Chryseobacterium sp. 92HC11
100
Chryseobacterium marinum IMCC3228; EF554366
Chryseobacterium sp. 92HC10
98
Chryseobacterium antarcticum AT1013; AY553293
Flavobacterium frigidarium A2i; AF162266
Flavobacterium sp. 92HB87
100
Flavobacterium sp. 92HB88
Flavobacterium sp. 92HB89
Flavobacterium sp. 92HB90
100
Flavobacterium sp. 92HB91
Flavobacterium sp. 92HB92
0.05
Flavobacterium sp. 92HB86
Flavobacterium sp. 92HB93

FIGURE A1 Unrooted neighbor-joining tree constructed from maximum likelihood distances between 16S rRNA gene sequences obtained from isolates of this study with the most closely aligned type strain obtained from the RDP. Type strains are displayed with strain ID followed by the GenBank 16S rRNA gene accession number. GenBank accession number for isolates obtained in this study is provided in Table A1. Bootstrap values ≥ 70% based on 1,000 replicates are indicated, and the scale bar represents the number of substitutions per site. The right color-coding columns of the phylogenetic clades show the tissue type each strain was isolated from and culture medium used. Skin is dark blue, gill is light blue, intestine is red, and water is yellow. Culture medium R2A is dark green, and MacConkey is light green

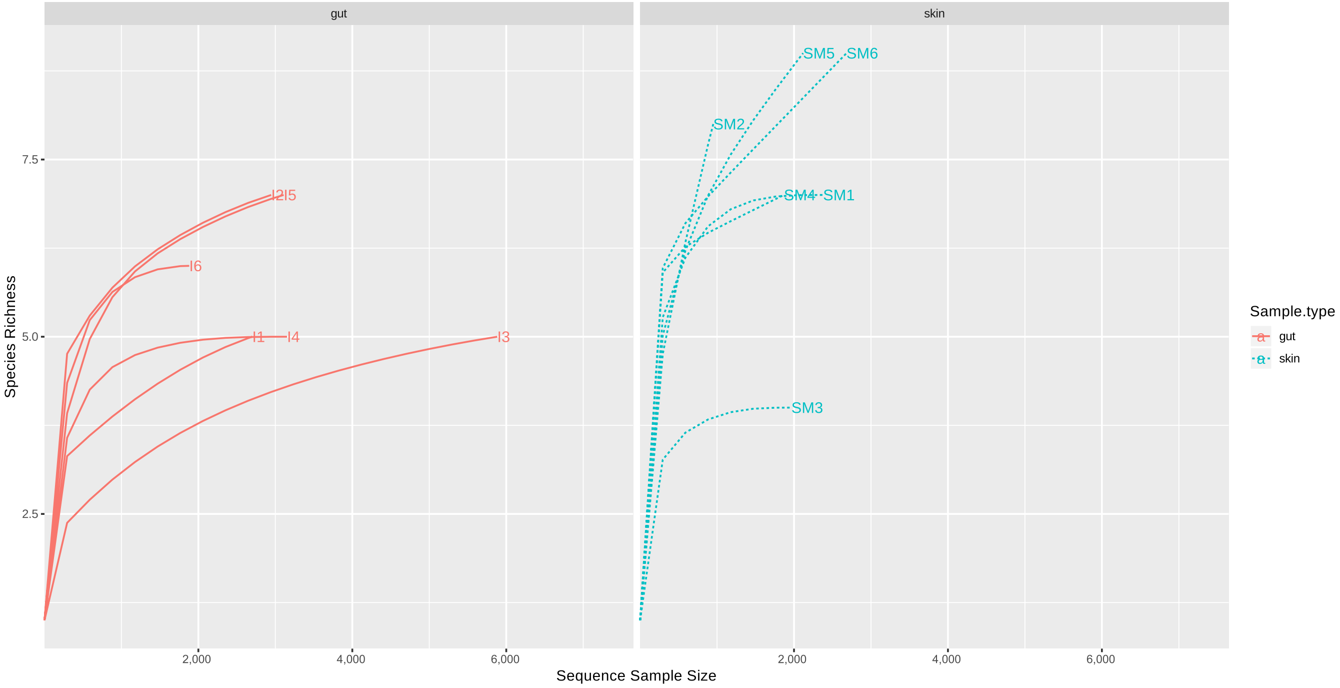


FIGURE A2 Rarefaction plots based on OTU analysis for intestine and skin samples. Underlaying data represent the number of sequences in each dataset mapped to OTUs clustered with a 97% identity threshold. Plot was generated using total sum scaling of filtered data showing 20 steps

Paper 5

T. Klemetsen, C. R. Karlsen, and N. P. Willassen, “Phylogenetic Revision of the Genus Aliivibrio: Intra- and Inter-Species Variance Among Clusters Suggest a Wider Diversity of Species,” *Front. Microbiol.*, vol. 0, p. 272, Feb. 2021, doi: 10.3389/FMICB.2021.626759.

ORIGINAL RESEARCH
published: 18 February 2021
doi: 10.3389/fmicb.2021.626759

Phylogenetic Revision of the Genus *Aliivibrio*: Intra- and Inter-Species Variance Among Clusters Suggest a Wider Diversity of Species

Terje Klemetsen[1]*, Christian R. Karlsen[2] and Nils P. Willassen[1]

[1] Department of Chemistry, Center for Bioinformatics, UiT The Arctic University of Norway, Tromsø, Norway, [2] Department of Fish Health, Nofima, Aas, Norway

OPEN ACCESS

Edited by:
Marcelino T. Suzuki,
Sorbonne Universités, France

Reviewed by:
Maxime Bruto,
Centre National de la Recherche Scientifique (CNRS), France
Isao Yumoto,
National Institute of Advanced Industrial Science and Technology (AIST), Japan

***Correspondence:**
Terje Klemetsen
terje.klemetsen@uit.no

Specialty section:
This article was submitted to Aquatic Microbiology, a section of the journal Frontiers in Microbiology

Received: 06 November 2020
Accepted: 27 January 2021
Published: 18 February 2021

Citation:
Klemetsen T, Karlsen CR and Willassen NP (2021) Phylogenetic Revision of the Genus Aliivibrio: Intra- and Inter-Species Variance Among Clusters Suggest a Wider Diversity of Species. Front. Microbiol. 12:626759. doi: 10.3389/fmicb.2021.626759

Genus *Aliivibrio* is known to harbor species exhibiting bioluminescence as well as pathogenic behavior affecting the fish farming industry. Current phylogenetic understanding of *Aliivibrio* has largely remained dormant after reclassification disentangled it from the *Vibrio* genus in 2007. There is growing evidence of wider diversity, but until now the lack of genomes and selective use of type strains have limited the ability to compare and classify strains firmly. In this study, a total of 143 bacterial strains, including 51 novel sequenced strains, were used to strengthen phylogenetic relationships in *Aliivibrio* by exploring intra-species and inter-species relations. Multilocus sequence analysis (MLSA), applying the six housekeeping genes *16S ribosomal RNA* (*rRNA*), *gapA*, *gyrB*, *pyrH*, *recA*, and *rpoA*, inferred 12 clades and a singular branch in *Aliivibrio*. Along with four new phylogenetic clades, the MLSA resolved prior inconsistencies circumscribing *Aliivibrio wodanis* and formed a unique clade we propose as the novel species *Aliivibrio* sp. "friggae." Furthermore, phylogenetic assessment of individual marker genes showed *gyrB*, *pyrH*, and *recA* superior to the 16S rRNA gene, resolving accurately for most species clades in *Aliivibrio*. In this study, we provide a robust phylogenetic groundwork for *Aliivibrio* as a reference point to classification of species.

Keywords: *Vibrionaceae*, *Aliivibrio*, phylogeny, multilocus sequence analysis, marine bacteria, species group coherence, marker gene

INTRODUCTION

The family of *Vibrionaceae* contains a large number of bacterial species, many of which are described from marine habitats (Thompson et al., 2004). A comprehensive study of the family and its evolutionary history suggested a common ancestor dating back to the Devonian era some 600 million years ago (Sawabe et al., 2007). *Vibrionaceae* is versatile, delineated and holds 22 distinct phylogenetic clades with highly diverse species of which several are harbored within the genus *Aliivibrio* (Sawabe et al., 2007, 2013). *Aliivibrio* is a firmly established genus, separate from *Vibrio* (Ast et al., 2009; Boyd et al., 2015). The genus harbors bioluminescent bacteria that have symbiotic relationships with aquatic organisms (Bongrand and Ruby, 2019), but also includes pathogens of marine animals (Hjerde et al., 2015; Kashulin et al., 2017).

Aliivibrio fischeri is studied extensively for its bioluminescence and symbiotic capability with marine squids and fishes (Visick, 2009). Exploration of *A. fischeri* has focused on revealing and understanding mechanisms of host adaptation, biofilm formation, flagellar function, quorum sensing and subsequent pathways to express its observed phenotypes (Visick, 2009; Verma and Miyashiro, 2013). The ability to form bioluminescent symbiosis with marine hosts has additionally been observed in *Aliivibrio logei* and *Aliivibrio* sp. "thorii" (Ast et al., 2009), while *Aliivibrio sifiae* is capable of forming independent bioluminescent colonies on marine agar (Yoshizawa et al., 2010). *Aliivibrio logei* is associated to skin of farmed fish (Benediktsdóttir et al., 1998), but also found in shellfish (Bang et al., 1978) and in the intestine of fish residing in the seas of Bering and Okhotsk (Bazhenov et al., 2019). *Aliivibrio finisterrensis* has been isolated from free living clams (*Ruditapes philippinarum*) and shown to be seasonally present in the hindgut of Tasmanian farmed Atlantic salmon (Beaz-Hidalgo et al., 2010; Hatje et al., 2014). *Aliivibrio wodanis* is associated to ulcerative skin problems of farmed fish (Karlsen et al., 2014), and *Aliivibrio salmonicida* is the causative agent of the seasonal cold-water vibrosis (Egidius et al., 1986). Strains of *Aliivibrio* sp. "thorii," *A. sifiae* and the non-luminescent *A. finisterrensis* have been given less attention. With the description of *A. finisterrensis* (Beaz-Hidalgo et al., 2010) and *A. sifiae* (Yoshizawa et al., 2010), the number of *Aliivibrio* species is currently six (Parte, 2018).

Utilization of the 16S ribosomal RNA (rRNA) gene marker-sequence is the prevalent method of inferring evolutionary relationships between taxa. However, 16S rRNA gene sequences may not provide interspecies resolution, and it is deemed a poor marker for resolving phylogenetically distinct species of *Vibrionaceae* (Sawabe et al., 2013; Ashok Kumar et al., 2020). Alternative marker genes are used to improve the resolution of species and the accuracy of classification in PCR based methods. Markers with a low degree of sequence conservation are favorable (Lan et al., 2016). The same study additionally found the *coaE* marker gene to phylogenetically mimic the genome wide amino acid identity in *Bacillus*. Other studies have applied the *fur* gene to further increase discriminatory power within *Vibrionaceae* (Machado and Gram, 2015), while *glnAI* has been proposed as an improvement for *Bifidobacteriaceae* compared to the 16S rRNA marker (Killer et al., 2020). Although single markers can provide phylogenetic resolution of species, a combination is often necessary to increase discriminatory power and expose monophyletic groups. This became evident following the use of multilocus sequence analysis (MLSA) reclassifying *Aliivibrio* (Urbanczyk et al., 2007), and in the identification of *Aliivibrio* sp. "thorii" (Ast et al., 2009). To further improve the accuracy of species identification, genome sequencing and genome-wide analysis were introduced (Konstantinidis and Tiedje, 2005). Indeed, genomic taxonomy suggests a correction of the whole *Vibrionaceae* family to change its parent order from the *Vibrionales* to the *Enterobacterales* (Parks et al., 2018).

As methods for more accurate classification has advanced, corrections of strains representative for species within *Aliivibrio* (formerly *Photobacterium* and later *Vibrio*) have occurred several times. For example, strain ATCC 15382, formerly classified as *Vibrio logei*, has later been suggested as a representative of *A. wodanis* (Ast et al., 2009). The same study additionally linked luminescent strain SR6 and SA12 to *A. wodanis*. However, bioluminescence in *A. wodanis* has not been described (Lunder et al., 2000; Hjerde et al., 2015). Furthermore, *A. sifiae* was published by introducing strain H1-1^T and H1-2 (Yoshizawa et al., 2010), while a set of 11 strains by Ast et al. (2009) were informally named as the "sifiae" clade. Without comparison to the described type strains there is no evidence for this classification.

Understanding the ecology and evolution of *Aliivibrio* requires robust and accurate genus- and species-level taxa. The present taxonomic classification results from representative type strains of selected species, eluding the species concepts of genomically coherent groups of microorganisms (Rosselló-Mora and Amann, 2001). The aim of the study was to firmly establish phylogenetic knowledge about the *Aliivibrio* genus with new sequenced data and to analyze marker genes for accurate classification of *Aliivibrio* species.

MATERIALS AND METHODS

Samples and Data Preparation

In this study, marker genes from 143 bacterial strains obtained from in-house sequenced genomes and GenBank (Benson et al., 2017; RRID:SCR_004860; **Supplementary Table 1**) were used to inferring the evolutionary relationships. Among these, 134 were represented by *Aliivibrio* strains and for comparison with neighboring genera, four strains of *Vibrio* and *Photobacterium* were included. The outgroup was represented by *Photorhabdus luminescens* subsp. *laumondii* TT01^T in line with prior phylogenetic studies (Urbanczyk et al., 2007; Ast et al., 2009). The type strains of *A. finisterrensis* DSM 23419^T and *A. logei* ATCC 29985^T, and 51 additional *Aliivibrio* isolates were genome-sequenced for this study. Strains and isolates sequenced in this study are available from the authors upon request.

For sequencing, *Aliivibrio* isolates, were revived from cryopreserved glycerol stocks and cultured in Luria-Bertani (LB) broth supplied with 3.5% w/v sodium chloride at 12°C. Genomic DNA was extracted using the Qiagen DNeasy blood and tissue kit protocol for Gram-negative bacteria. Sequencing libraries were prepared using the Nextera XT DNA Library Preparation Kit (Illumina) according to the manufacturer's protocol. The fragment size distribution was verified to 500–1,000 bp using the Agilent 2100 Bioanalyzer System. Libraries were multiplexed and sequenced on an Illumina MiSeq instrument (RRID:SCR_016379), using either MiSeq Reagent kits v2 (500 cycles) or v3 (600 cycles), yielding an average of 3.8 million reads per bacterial isolate (**Supplementary Table 2**).

All sequence reads were quality controlled and each genome was *de novo* assembled using the CLC Genomics Workbench (RRID:SCR_011853) version 8.0.3. Briefly, paired-end reads were imported using the built-inn CLC pipeline removing failed reads. Reads were further quality trimmed with an ambiguous limit of 2 and a quality limit of 0.05 while reads shorter than 15 bases were removed. *De novo* assembly was performed with default parameters, auto-detecting paired distances and performing scaffolding. A cutoff for minimum contig length was set to 500 bp. On average, the *de novo* assembly gave 343 scaffolds (N50 of 62,646) with total assembly lengths between 3.7 and 5.2 Mb, and an average coverage of 247.43x (**Supplementary Table 2**).

The assembled genomes were annotated using Prokka (Seemann, 2014; RRID:SCR_014732) version 1.13 on the Galaxy platform (Afgan et al., 2016; RRID:SCR_006281) with a default parameter setting. Annotated genomes were screened for the 16S rRNA, *gapA* (P0A9B2), *gyrB* (P0A2I3), *pyrH* (P65933), *recA* (P65977), and *rpoA* (Q664U6) genes, and identified sequences extracted. Public sequence data under the *Aliivibrio* taxa (taxonomy ID 511678) which contained all six genes were gathered from GenBank. Locus tag identifiers from the 90 public strains used in this study are listed in **Supplementary Table 1**.

Sequences in each gene locus were aligned individually using MUSCLE (Edgar, 2004; RRID:SCR_011812) version 3.8.31 with default parameters for nucleotide sequences. Gene regions were selected according to Sawabe et al. (2007). However, flanking ends in each alignment were recursively trimmed. Briefly, flanking positions with gaps occurring in more than 5% of the alignment sequences were trimmed using Aliview (Larsson, 2014; RRID:SCR_002780) version 1.2.6. *Vibrio cholerae* strain N16961 (ungapped gene numbering) was used as reference to the trimmed alignments with the following gene name, gene position range, and reference locus tag: 16S rRNA, 252-1422, and VCr001; *gapA*, 89-862, and VC2000; *gyrB*, 308-1496, and VC0015; *pyrH*, 21-624, and VC2258, *recA*, 69-865 VC0543; and *rpoA*, 20-950, and VC2571. The concatenated sequence of the six trimmed fragments (16S rRNA-*gapA-gyrB-pyrH-recA-rpoA*) produced a multilocus sequence alignment (MLSA) of 5,473 positions.

Analysis

Phylogenetic relationships between bacterial strains included in this study were constructed on the basis of the concatenated MLSA. A network graph was created in SplitsTree4 (Huson and Bryant, 2005; RRID:SCR_014734) version 4.13.1 by applying the Jukes-Cantor (JC69) distance correction between sequences while NeighborNet was used as network model.

To analyze the evolutionary variance within species and the average evolutionary distance between species, strains were assigned to designated groups based on the network model (**Figure 1**). Briefly, the concatenated gene loci dataset was imported as a nucleotide dataset into MEGA X (Kumar et al., 2018; RRID:SCR_000667) version 10.1.7. Fourteen groups were assigned by the Sequence Data Explorer in MEGA using *P. luminescens* subsp. *laumondii* TT01^T as outgroup and *Aliivibrio* sp. appey-12 as a singular group. Distance estimation with SE was calculated for within and between groups of species under the conditions: uniform rates among sites and pairwise deletion was used while applying the JC69 as substitution model. All estimations were statistically tested with 1,000 bootstrap replications. MEGA was further used to construct a Neighbor-joining tree using the MLSA, 16S rRNA, *gapA*, *gyrB*, *pyrH*, *recA*, *rpoA*, concatenated *recA-rpoA, gyrB-rpoH*, and *gyrB-recA* with equivalent parameters as given for the distance measurements. Resulting newick files from each inferred tree was compared topologically against the MLSA using the MutualClusteringInfo algorithm in the TreeDist R package (Smith, 2020). Due to conflicting overlap between sequences in the 16S rRNA alignment, *Photobacterium angustum* strains ATCC 33977 and S14 were removed from the datasets only prior to tree construction and topological comparison.

Sequence identities between and within designated groups were estimated as described for evolutionary distances using single genes as well as the MLSA and a 5-gene concatemer excluding the 16S rRNA gene. The python script identity.py[1] was created to calculate sequence identities. In short, the script evaluates all sequences pairwise, removing any gapped positions before calculating the identity as a percentage. For each dataset (single genes and MLSA) the identity values were enlisted in either of two subsets; those associated with the same genus (within-genera) or any different genera (between-genera). The same procedure was repeated to differentiate subsets of intra- and inter- species identity values. Values not within and between hitherto described species were filtered. Distributions were plotted as $\log_{10}$ transformed histograms to simplify identification of overlapping data. Subsets became colored based on their affiliation as intra- or inter- subsets. GC content was calculated using Biopython GC (Cock et al., 2009; RRID:SCR_007173).

RESULTS AND DISCUSSION

Phylogeny of *Aliivibrio* Reveals 12 Distinct Species Clades

In this study, a MLSA scheme based on six concatenated genes (16S rRNA gene, *gapA*, *gyrB*, *pyrH*, *recA*, and *rpoA*) were used to infer the phylogeny and evolutionary relationships in the *Aliivibrio* genus. Based on the MLSA sequence data both the inferred phylogenetic network (**Figure 1**) and phylogenetic tree (**Supplementary Figure 1**) remained congruent with only minor topology inconsistencies. Twelve individual clades were identified of which seven corresponded to clades described in earlier studies: *A. fischeri* (Urbanczyk et al., 2007), *A. finisterrensis* (Beaz-Hidalgo et al., 2010), *Aliivibrio* sp. "thorii" (Ast et al., 2009), *A. sifiae* (Yoshizawa et al., 2010), *A. wodanis* (Lunder et al., 2000), *A. logei* (Bang et al., 1978), and *A. salmonicida* (Egidius et al., 1986). The results corroborate the reclassification by Urbanczyk et al. (2007), the wider description of *Aliivibrio* (Ast et al., 2009) and confirm the presence of *Aliivibrio* sp. "thorii," *A. finisterrensis* and *A. sifiae* to the genus. Sixteen strains could not be affiliated to any of the described clades.

[1] github.com/tkl014/aliivibrio_identity

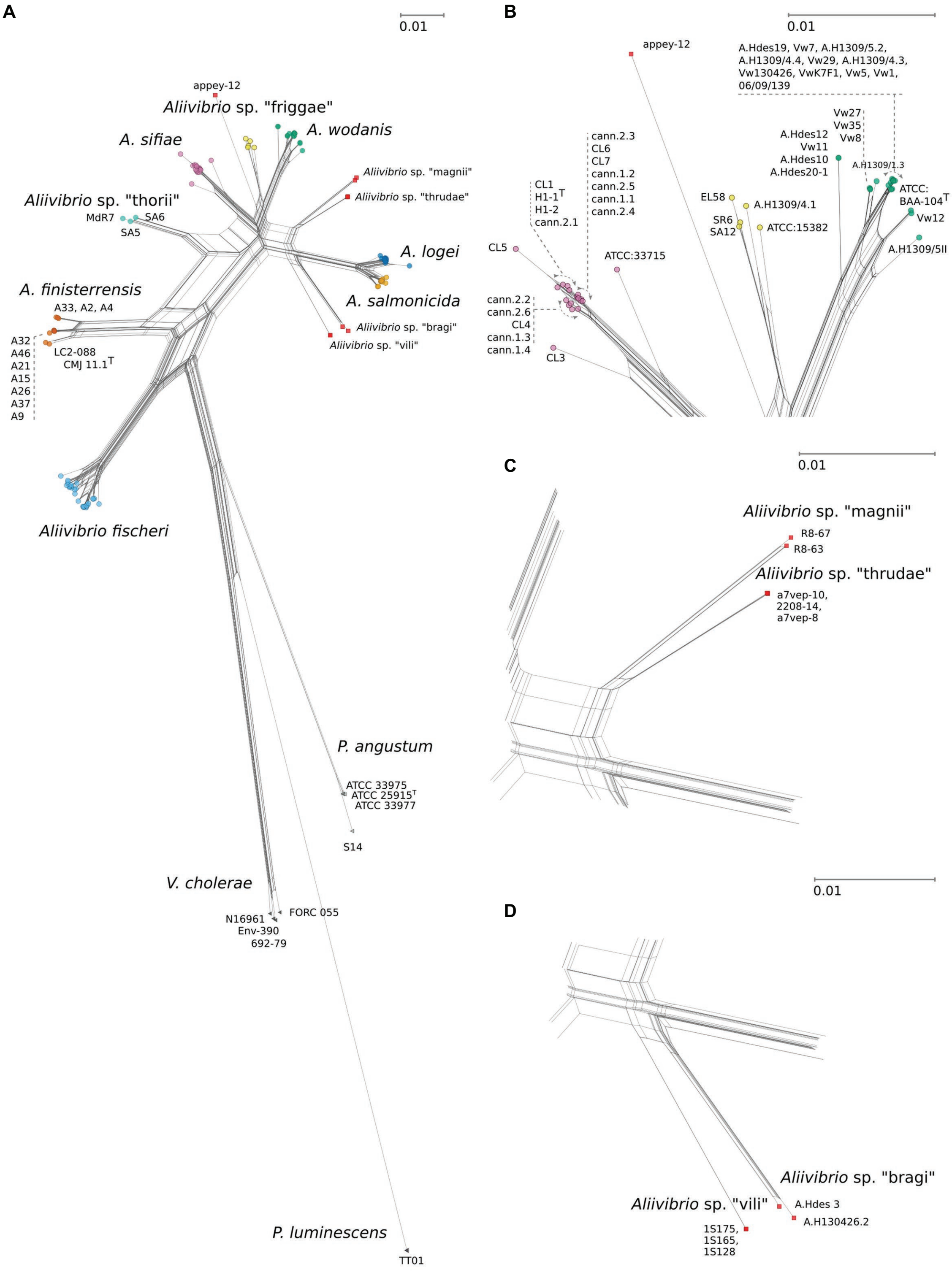


FIGURE 1 | (A) JC69 corrected NeighborNet comprising 143 strains including *Aliivibrio*, *Vibrio cholerae*, *Photobacterium angustum*, and with *Pluminescens luminescens* acting as outgroup. Network is based on a concatenated alignment of the 16S ribosomal RNA (rRNA) gene, *gapA*, *gyrB*, *pyrH*, *recA*, and *rpoA* spanning 5,473 nt positions. Arrows indicate relative placements according to listed strains. Scale bars are relative to the given panels and represent the number of base substitutions per site. **(B)** Details on *Aliivibrio sifiae*, *Aliivibrio* sp. "friggae," and *Aliivibrio wodanis* groups. **(C)** Detailed view of "magnii" and "thrudae" clades. **(D)** Strain details for the "bragi" and "vili" clades. See **Figure 3**, for details on *Aliivibrio fischeri, Aliivibrio logei*, *and Aliivibrio salmonicida*.

These strains gave rise to one singular branch (appey-12) and five clades in which we suggest the provisionally names "friggae," "magnii," "thrudae," "bragi," and "vili" (**Figure 1**) in order to provide working names in line with other species within *Aliivibrio* genus that have derived their names after Norse mythology gods (*A. wodanis*, *A. logei*, *A. sifiae,* and *Aliivibrio* sp. "thorii"). Clades inferred by the neighbor-joining approach (**Supplementary Figure 1**) had, except for *Aliivibrio* sp. "magni" and *Aliivibrio* sp. "thrudae," robust support values. The friggae clade consists of five strains and includes the SR6, SA12 and ATCC 15382, previous classified as wodanis (Ast et al., 2009), EL58, and A.H1309/4.1. These strains were isolated from different hosts such as fishes (Atlantic salmon and Pacific cod), gorgonian coral and bobtail squids. The "thrudae" clade represent three strains isolated from lumpfish while the "magnii" clade represent two isolates from amphipods. Filtered seawater is the source of all three strains composing the "vili" clade. Lastly, strains in the "bragi" clade originate from skin ulcer and head kidney samples isolated from Atlantic salmon.

In-Depth Analysis of Species Groups and Distances

This study utilizes several strains from each species group in *Aliivibrio* rather than a single representative type strain. This approach provides a more robust statistical measure of group affiliations (intra-species), their circumference, and observed interrelations between groups (inter-species). The results showed fluctuating evolutionary variances in *Aliivibrio*, *Vibrio*, and *Photobacterium* (**Figure 2**) that closely reflect the circumference of clades in the phylogenetic tree (**Supplementary Figure 1**). *Aliivibrio* species ranged from the narrow intra-species variance of *A. salmonicida* (0.0012) to the nine times wider variance of *A. fischeri* (0.0109). Still, these measurements are expected to be inaccurate for clades represented by a low number of strains such as *Aliivibrio* sp. "vili" and *Aliivibrio* sp. "thrudae."

Analysis of inter-species distances using the JC69 substitution model is shown in **Table 1**. These values reflect the distance and error value deviation between species groups as they appear in the phylogenetic tree diagram (**Supplementary Figure 1**). In *Aliivibrio*, the average evolutionary inter-species distance was 0.060 where the general defined species groups diverged by ≥0.041. The most distant species groups were between *A. fischeri* and *Aliivibrio* sp. "bragi" (0.104), while the smallest distance was between *A. salmonicida* and *A. logei* (0.013), which make them the closest neighboring species in *Aliivibrio* (**Figure 3A**). It is noteworthy that the low sequence variances of *A. salmonicida* and *A. logei* contrast the number of different host species. Extended sampling of environments might result in the emergence of one interchangeable species rather than two, as a full genome ANI approach has reported representatives of both *A. salmonicida* and *A. logei* as "s__Aliivibrio salmonicida_A" (Parks et al., 2020). This is intriguing as *A. salmonicida* causes cold-water vibriosis and *A. wodanis* causes wodanosis and/or winter ulcer in Atlantic salmon, while *A. logei* is not known to be a salmon pathogen. The phylogenetic structure of *A. fischeri* reflects differences in the host species, but also colonization behavior (Bongrand et al., 2016). Comparable to the *A. fischeri* phylogenetic clades, *A. salmonicida*, *A. wodanis*, and *A. logei* similarities may relate to colonization effectiveness while their differences may be related to behavioral specialization and dependent on environmental factors such as temperature (Nishiguchi, 2000; Hatje et al., 2014; Sunagawa et al., 2015). When considering the 5-gene average GC content (**Table 2**), higher values (42.5–43.1%) were measured compared to the

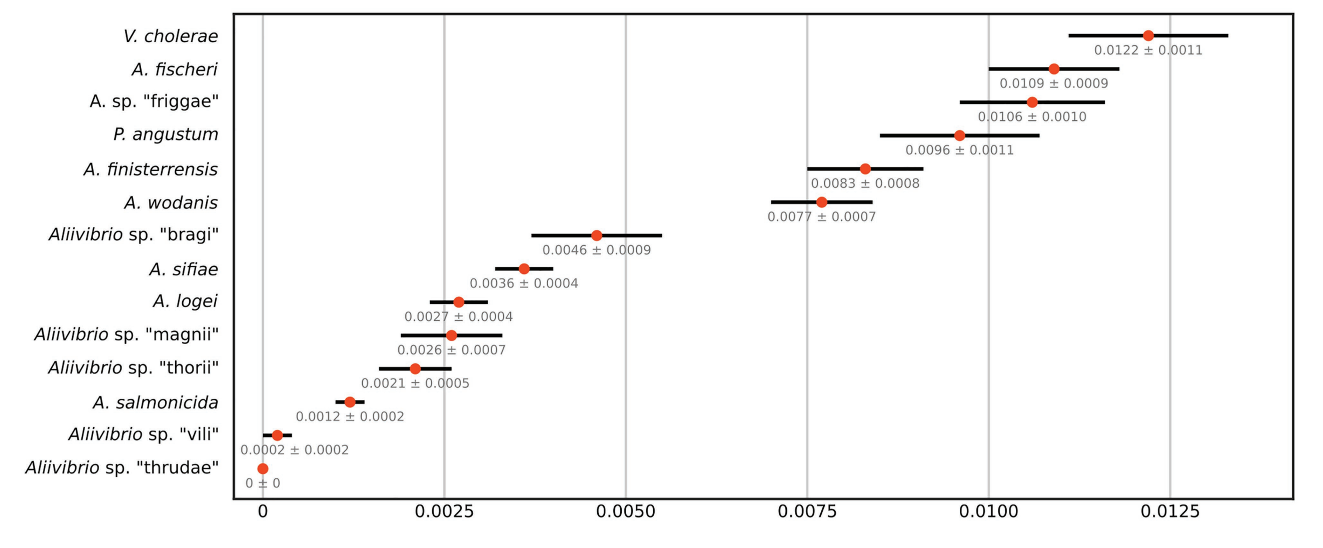


FIGURE 2 | JC69 corrected intra-species distances of the MLSA dataset as number of base substitutions per site. Orange centroids reflect the average over all sequence pairs compared within each group and are drawn with SE values from 1,000 bootstrap replicates.

TABLE 1 | JC69 corrected inter-species distances (lower) of the MLSA dataset as number of base substitutions per site with SE values (upper) from 1,000 bootstrap replicates.

	Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1.	*A. finisterrensis*		0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.003	0.004	0.006	0.008	0.006
2.	*A. fischeri*	0.073		0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.006	0.008	0.006
3.	*A. logei*	0.086	0.100		0.001	0.003	0.003	0.003	0.003	0.004	0.003	0.003	0.004	0.003	0.006	0.008	0.006
4.	*A. salmonicida*	0.085	0.097	0.013		0.003	0.003	0.003	0.003	0.004	0.003	0.003	0.004	0.003	0.006	0.008	0.006
5.	*A. sifiae*	0.078	0.095	0.050	0.051		0.003	0.003	0.003	0.003	0.003	0.002	0.003	0.003	0.006	0.008	0.006
6.	*Aliivibrio* sp. “vili”	0.083	0.101	0.043	0.045	0.046		0.003	0.003	0.004	0.003	0.003	0.003	0.003	0.006	0.008	0.006
7.	*Aliivibrio* sp. “thrudae”	0.077	0.098	0.040	0.041	0.046	0.039		0.003	0.003	0.002	0.002	0.003	0.003	0.006	0.008	0.006
8.	*Aliivibrio* sp. “bragi”	0.088	0.104	0.041	0.041	0.047	0.040	0.041		0.004	0.003	0.003	0.003	0.003	0.006	0.008	0.006
9.	*Aliivibrio* sp. appey-12	0.082	0.104	0.067	0.067	0.050	0.065	0.051	0.064		0.003	0.003	0.004	0.003	0.006	0.008	0.006
10.	*Aliivibrio* sp. “magnii”	0.083	0.101	0.046	0.045	0.050	0.038	0.031	0.036	0.056		0.003	0.003	0.003	0.006	0.008	0.006
11.	*Aliivibrio* sp. “friggae”	0.076	0.096	0.053	0.053	0.036	0.048	0.030	0.047	0.042	0.037		0.003	0.002	0.006	0.008	0.006
12.	*Aliivibrio* sp. “thorii”	0.062	0.093	0.065	0.065	0.046	0.053	0.056	0.059	0.072	0.062	0.051		0.003	0.006	0.008	0.006
13.	*A. wodanis*	0.080	0.097	0.056	0.058	0.041	0.052	0.044	0.052	0.050	0.047	0.030	0.055		0.006	0.008	0.006
14.	*P. angustum*	0.156	0.148	0.164	0.164	0.165	0.164	0.161	0.163	0.169	0.163	0.161	0.165	0.161		0.008	0.006
15.	*P. luminescens*	0.262	0.253	0.262	0.260	0.262	0.262	0.260	0.260	0.264	0.262	0.259	0.260	0.260	0.259		0.007
16.	*V. cholerae*	0.177	0.166	0.186	0.185	0.184	0.184	0.188	0.183	0.190	0.186	0.186	0.186	0.183	0.185	0.239	

Calculated average between *Aliivibrio* groups was found to be 0.060. Inter-species identity values are additionally provided in **Supplementary Table 3**.

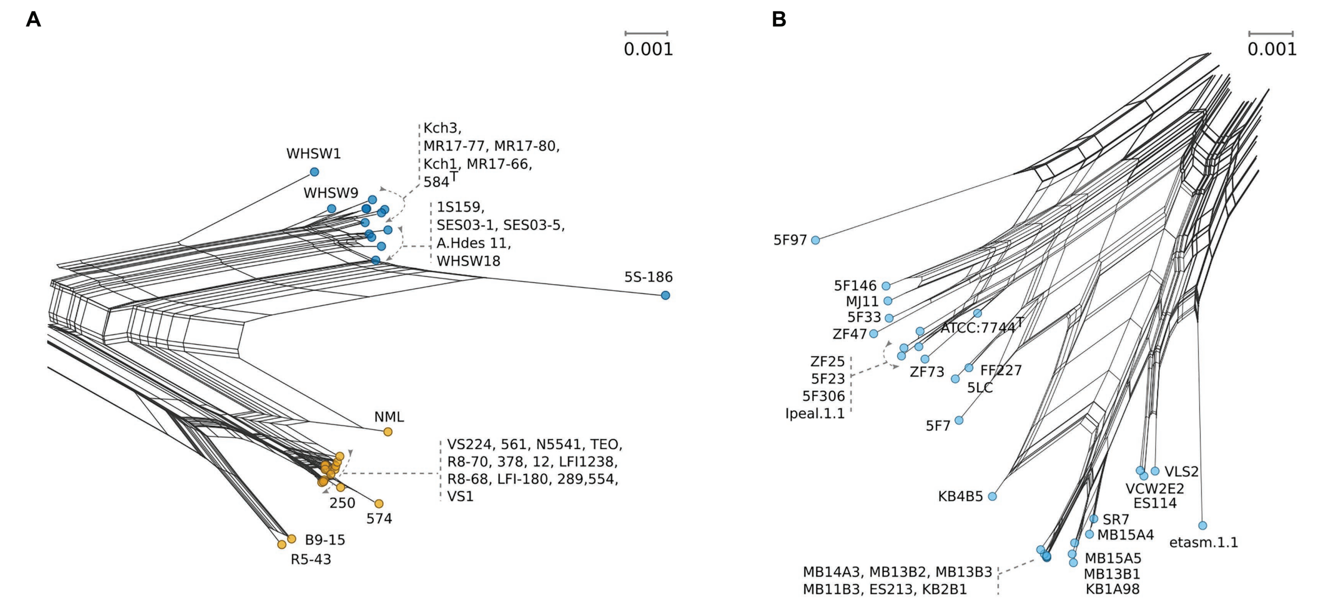


FIGURE 3 | Detailed selections from NeighborNet **Figure 1** with uniform scales showing number of base substitutions per site. **(A)** Strain details for *A. logei* (blue) and *A. salmonicida* (yellow). **(B)** Detailed view for the *A. fischeri* clade (pale blue).

reported genome average of *A. salmonicida* (39.8%), *A. wodanis* (39.4%) and *A. fischeri* (39.0%; Hjerde et al., 2008, 2015; Califano et al., 2015). Indeed, the genome traits of *A. salmonicida* are suggestive of host specificity adaptation (Hjerde et al., 2008), which do not rule out that other lineages of *Aliivibrio* may have evolved by similar ecological strategies. Genetic drifts for specialization of Aliivibrios could involve genes with lower GC content than the house keeping genes, such as the MLSA scheme used in this study.

TABLE 2 | Minimum sequence identities for *Aliivibrio* from trimmed and ungapped sequence pairs within intra-species groups.

Group	Strains	MLSA identity (%)	16S identity (%)	5-gene identity (%)	5-gene GC (avg, %)
A. finisterrensis	12	≥ 98.30	≥ 99.74	≥ 97.76	43.134
A. fischeri	30	≥ 97.88	≥ 98.80	≥ 97.44	42.898
A. logei	14	≥ 99.19	≥ 98.80	≥ 99.14	43.088
A. salmonicida	18	≥ 99.45	≥ 99.66	≥ 99.39	42.647
A. sifiae	19	≥ 97.77	≥ 99.46	≥ 97.25	42.502
Aliivibrio sp. "vili"	2	≥ 99.96	100	≥ 99.95	42.917
Aliivibrio sp. "thrudae"	3	100	100	100	42.894
Aliivibrio sp. "bragi"	3	≥ 99.54	99.83	99.46	42.742
Aliivibrio sp. "magnii"	2	≥ 99.74	100	99.67	42.778
Aliivibrio sp. "friggae"	5	99.71–98.17	≥ 98.72	99.74–98.00	42.679
Aliivibrio sp. "thorii"	3	99.96–99.49	99.83–97.87	≥ 99.93	42.537
Aliivibrio wodanis	22	≥ 98.11	≥ 99.66	≥ 97.62	42.653
Aliivibrio	134	≥ 89.42	≥ 96.65	≥ 87.05	42.786

The MLSA void of the 16S rRNA gene is here defined as a 5-gene concatemer. *Aliivibrio* sp. appey-12 is included in the full set of *Aliivibrio*. For individual gene measurements see **Supplementary Table 4**.

Assessment of Sequence Identity Suggests Improvements in Operational Taxonomic Assignment to *Aliivibrio* Using the *pyrH* or *rpoA* Marker

Gene sequence identity is frequently used as distance measurements in homolog comparisons and for clustering operational taxonomic units (OTUs; Edgar, 2017). Here, the inferred phylogeny of *Aliivibrio* was measured interchangeably for species and genera. Estimates based on the 16S rRNA gene sequences show ≥98.80% intra-species identity and ≥96.65% for *Aliivibrio* (**Table 2**). Similarly, Urbanczyk et al. (2007) reported ≥97.4 intra-species identity among four representative *Aliivibrio* species, illustrating the limited resolving power that corroborate previous assessments by Sawabe et al. (2013). Sequence identity showed no overlap between the within-genera and between-genera subset for the genes *rpoA*, *pyrH*, and the concatenated MLSA (**Figure 4A**). Gaps between subsets of *rpoA*, *pyrH*, and MLSA were 5.31, 2.45, and 2.72%, respectively. The two extremes of *rpoA*; *P. angustum* ATCC 25915^T and *A. sifiae H1-2* (between genera) shared 90.67% identity, while *Aliivibrio* sp. ATCC 15382 and *A. fischeri* 5LC (between species in same genera) shared 96.01% identity. Hence, the *rpoA*, *pyrH*, and MLSA datasets have capability to discriminating genera, but dataset overlaps are unreliable at species level differentiation (**Figure 4B**). Genus-level overlap was observed for the 16S rRNA gene (1171 bp, covering hyper variable regions v3–v8). This causes potentially

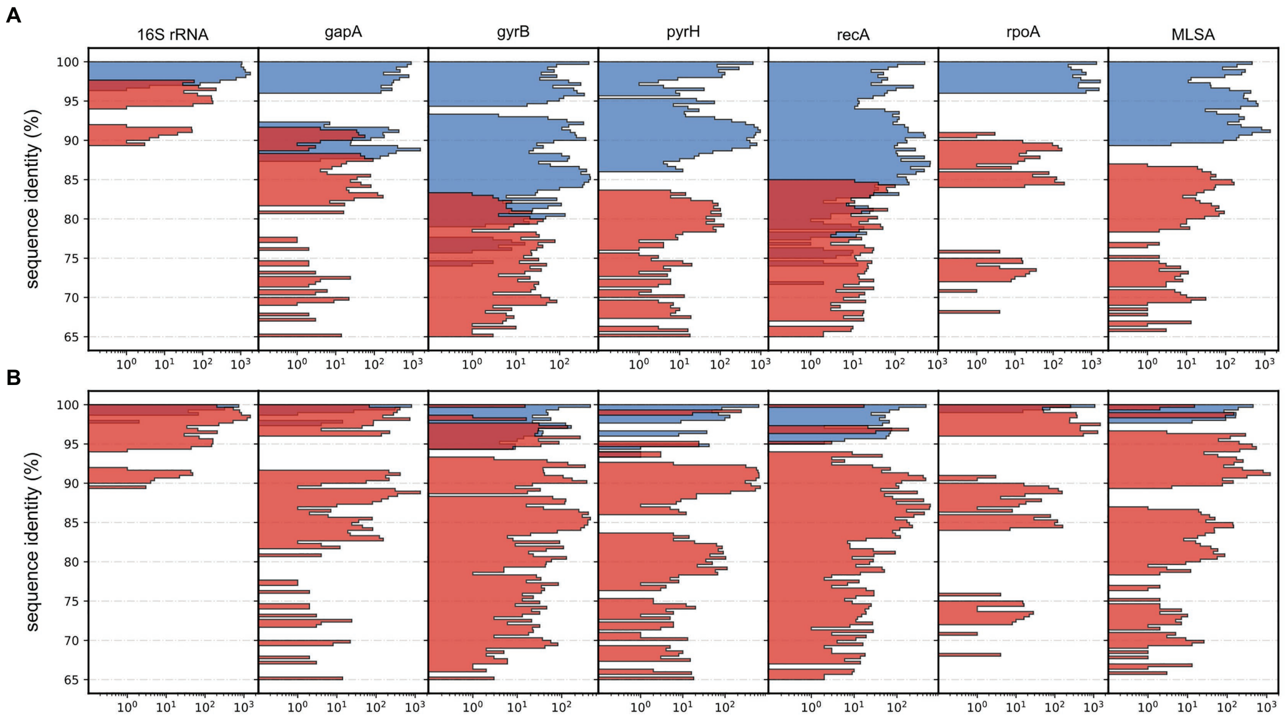


FIGURE 4 | Histograms depicting distributions of pairwise sequence identities for each applied gene and the multilocus sequence analysis (MLSA). Data counts (*x*-axis) are log10 transformed for improved visualization and presented as colored subsets based on relative affiliation. **(A)** Between-genera (red) and within-genera (blue). **(B)** Intra-species (red) and inter-species (blue).

inaccurate or erroneous classification and OTU clustering of Aliivibrios using near full length 16S rRNA gene sequences. Utilization of the MLSA dataset resulted in lower intra-species identity (≥97.77%) and considerably reduced identity for the whole *Aliivibrio* genus (≥89.42%), indicating improved resolution. Also, the 5-gene concatemer, without the 16S rRNA gene, resulted in low identity values (**Table 2**). However, estimates for *A. logei* and *Aliivibrio* sp. "thorii" were found contradicting, suggesting the 16S rRNA gene to be favorable for classification of some species.

Reasonable Phylogenetic Resolving Power in Either *recA*, *pyrH*, or *gyrB* Demonstrate a Potential to Classify Aliivibrios Similarly as the Full MLSA Tree

The ability of individual gene markers to classify monophyletic groups with shared topology to the MLSA scheme was assessed. Marker gene *recA* produced a polyphyletic group of *Aliivibrio* sp. "thorii" and *Aliivibrio* sp. "friggae" (indicated in **Figure 5A**) that shared 67% of the MLSA topology with a misplacement of *Aliivibrio* sp. appey-12. Similar ability to resolve species was observed by *pyrH* (sharing 56.66% of the MLSA topology) in which individual strains of *Aliivibrio* sp. "friggae," *A. sifiae* and *A. salmonicida* mixed with neighboring clades (**Figure 5B**) – similar to previously reported *pyrH* discrepancies in the *Vibrio* group (Pascual et al., 2010). Marker *gyrB* shared 73.51% of the MLSA topology. Manual assessment of the tree revealed *A. wodanis* and *Aliivibrio* sp. "friggae" as polyphyletic (**Figure 5C**) while *Aliivibrio* sp. appey-12 interfere with the *A. sifiae* clade. Still, *gyrB*, *pyrH*, and *recA* markers show significant improvement for *Aliivibrio* classification compared to the 16S rRNA, *gapA*, and *rpoA* which had 54.66, 53.73, and 53.73% topological similarity to the MLSA, respectively. Visual inspection of the resulting trees from 16S rRNA and *gapA* (**Supplementary Figures 2, 3**) show concerns in resolving *A. sifiae*, *A. logei* and smaller clades like *Aliivibrio* sp. "thorii." Furthermore, the relative wide identity gap between genera described earlier becomes apparent for the *rpoA* marker tree (**Supplementary Figure 4**). It firmly discriminates *Aliivibrio* from *Vibrio* and *Photobacterium*, but show similar conserved nature as 16S rRNA for highly similar strains like *A. salmonicida* and *A. logei*.

Although, none of the discussed markers showed discriminatory power equal to that of the MLSA, paired combinations of concatenated *gyrB*, *pyrH*, *recA*, and *rpoA* marker sets were tested. Comparable classification to the MLSA was found in *recA-rpoA* (data not shown), *gyrB-pyrH* (data not shown), and *gyrB-recA* (**Figure 5D**). Based on mutual clustering information trees from these markers shared 68.17, 77.45, and 80.46% of the MLSA topology, respectively. High topological similarity by *gyrB* produced a discriminating power of *gyrB-recA* that best resembled the MLSA phylogeny for *Aliivibrio*.

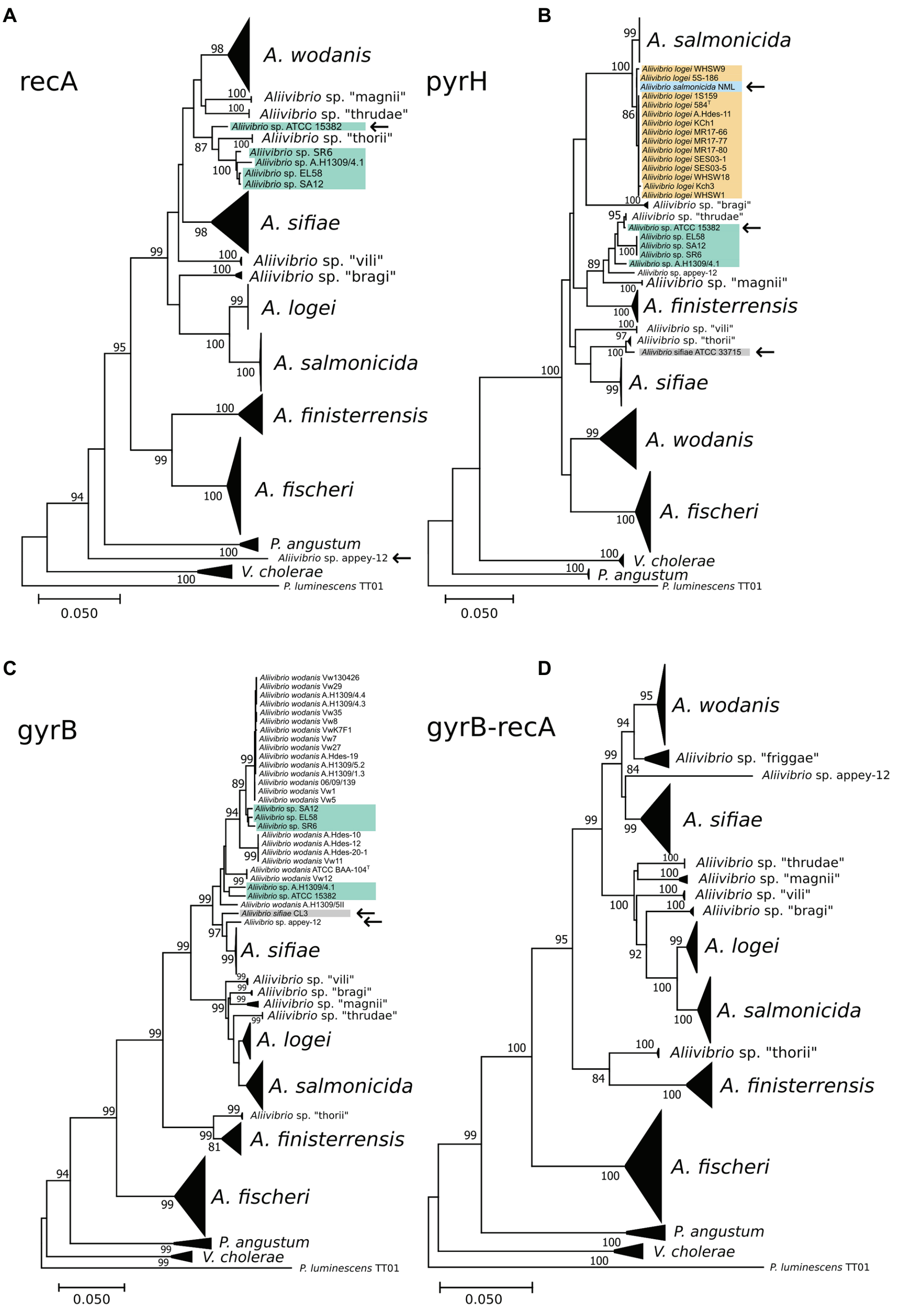


FIGURE 5 | Phylogenetic reconstruction of the marker genes *recA* **(A)**, *pyrH* **(B)**, *gyrB* **(C)**, and *gyrB-recA* **(D)**. Collapsed clades represent conserved classification similarly as in the full MLSA tree. Strains interfering in neighboring or other clades are marked by an arrow and strains of *Aliivibrio* sp. "friggae" has teal background color.

CONCLUSION

In this study, using 143 strains we have applied MLSA to gain new insight into the evolutionary structure and relationships in the *Aliivibrio* genus. Five new clades (friggae, vili, magnii, thrudae, and bragi) and one singular branch was identified in addition to the seven earlier described clades. These presented clades can be illustrated as a snapshot of current knowledge using available data. Future sampling will likely be expanding the complexity and number of clades in genus *Aliivibrio*.

In this study the discrepancy in intra-species variance for some clades, highly identical sequences may be attributed to a bias toward singular sampling origins or repetitive sampling of target species. The different *Aliivibrio* clades, independent of host range, show different inter-species sequence variances but may also be globally distributed with little sequence variations. This underline the need to include a sufficient number of strains that represent the population for each species, and not only type strains in taxonomic studies.

Host-associated microbiomes can influence their host's welfare and health and there is an ongoing effort to identify individual members and their contributions. To help, insight into the evolutionary structure of a genus would be beneficial. Here, the MSLA scheme generated is successfully used to infer a significant insight into the evolutionary relationships in the *Aliivibrio* genus. As a major member of the *Vibrionaceae* family in marine environments (Sunagawa et al., 2015; Machado and Gram, 2017), it includes members pathogenic to aquaculture species. Therefore, awareness is required to reinforce phylogenetic relationship of strains within the genus *Aliivibrio*. Continuing the MLSA approach would be of great value to further investigate the distribution and prevalence in the marine environment and to improve the accuracy of clinical diagnosis when *Aliivibrio* is detected from farmed aquatic animals. The identification of *Aliivibrio* species determined by the *gyrB-recA* approach could be an alternative and even more cost-effective way for a rapid and informative molecular method down to the taxonomic level of species within the *Aliivibrio*.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.ncbi.nlm.nih.gov/, PRJEB34882.

AUTHOR CONTRIBUTIONS

TK conducted and performed the bioinformatic analyses. NW coordinated the work. TK and CK drafted the manuscript. TK, CK, and NW authors provided critical feedback and helped shape the research, analysis, and manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by UiT The Arctic University of Norway and through the Research Council of Norway (RCN) by the Nofima basic grant (194050/F40). The publication charges for this article have been funded by a grant from the publication fund of UiT The Arctic University of Norway.

ACKNOWLEDGMENTS

We would like to thank Concetta De Santi, Seila Pandur and Juan Fu at the Center for Bioinformatics, UiT The Arctic University of Norway who revived the cultures, extracted DNA and performed the sequencing. Additional thanks to Erik Hjerde for giving helpful feedback on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.626759/full#supplementary-material

REFERENCES

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., et al. (2016). The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* 44, W3–W10. doi: 10.1093/nar/gkw343

Ashok Kumar, J., Vinaya Kumar, K., Avunje, S., Akhil, V., Ashok, S., Kumar, S., et al. (2020). Phylogenetic relationship among Brackishwater vibrio species. *Evol. Bioinforma.* 16:117693432090328. doi: 10.1177/1176934320903288

Ast, J. C., Urbanczyk, H., and Dunlap, P. V. (2009). Multi-gene analysis reveals previously unrecognized phylogenetic diversity in *Aliivibrio*. *Syst. Appl. Microbiol.* 32, 379–386. doi: 10.1016/j.syapm.2009.04.005

Bang, S. S., Baumann, P., and Nealson, K. H. (1978). Phenotypic characterization of *Photobacterium logei* (sp. nov.), a species related to *P. fischeri*. *Curr. Microbiol.* 1, 285–288. doi: 10.1007/BF02601683

Bazhenov, S. V., Khrulnova, S. A., Konopleva, M. N., and Manukhov, I. V. (2019). Seasonal changes in luminescent intestinal microflora of the fish inhabiting the Bering and Okhotsk seas. *FEMS Microbiol. Lett.* 366:fnz040. doi: 10.1093/femsle/fnz040

Beaz-Hidalgo, R., Doce, A., Balboa, S., Barja, J. L., and Romalde, J. L. (2010). *Aliivibrio finisterrensis* sp. nov., isolated from Manila clam, Ruditapes philippinarum and emended description of the genus Aliivibrio. *Int. J. Syst. Evol. Microbiol.* 60, 223–228. doi: 10.1099/ijs.0.010710-0

Benediktsdóttir, E., Helgason, S., and Sigurjónsdóttir, H. (1998). *Vibrio* spp. isolated from salmonids with shallow skin lesions and reared at low temperature. *J. Fish Dis.* 21, 19–28. doi: 10.1046/j.1365-2761.1998.00065.x

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2017). GenBank. *Nucleic Acids Res.* 45, D37–D42. doi: 10.1093/nar/gkw1070

Bongrand, C., Koch, E. J., Moriano-Gutierrez, S., Cordero, O. X., McFall-Ngai, M., Polz, M. F., et al. (2016). A genomic comparison of 13 symbiotic *Vibrio fischeri* isolates from the perspective of their host source and colonization behavior. *ISME J.* 10, 2907–2917. doi: 10.1038/ismej.2016.69

Bongrand, C., and Ruby, E. G. (2019). The impact of *Vibrio fischeri* strain variation on host colonization. *Curr. Opin. Microbiol.* 50, 15–19. doi: 10.1016/j.mib.2019.09.002

Boyd, E. F., Carpenter, M. R., Chowdhury, N., Cohen, A. L., Haines-Menges, B. L., and Kalburge, S. S., et al. (2015). Post-genomic analysis of members of the family *Vibrionaceae*. *Microbiol. Spectr.* 3:10. doi:10.1128/microbiolspec.VE-0009-2014

Califano, G., Franco, T., Gonçalves, A. C. S., Castanho, S., Soares, F., Ribeiro, L., et al. (2015). Draft genome sequence of Aliivibrio fischeri strain 5LC, a bacterium retrieved from gilthead sea bream (*Sparus aurata*) larvae reared in aquaculture. *Genome Announc.* 3, e00593–e00615. doi: 10.1128/genomeA.00593-15

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Edgar, R. (2017). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. bioRxiv [Preprint]. doi:10.1101/192211

Egidius, E., Wiik, R., Andersen, K., Hoff, K. A., and Hjeltnes, B. (1986). *Vibrio salmonicida* sp. nov., a new fish pathogen. *Int. J. Syst. Bacteriol.* 36, 518–520. doi: 10.1099/00207713-36-4-518

Hatje, E., Neuman, C., Stevenson, H., Bowman, J. P., and Katouli, M. (2014). Population dynamics of *Vibrio* and *Pseudomonas* species isolated from farmed Tasmanian Atlantic Salmon (*Salmo salar* L.): a seasonal study. *Microb. Ecol.* 68, 679–687. doi: 10.1007/s00248-014-0462-x

Hjerde, E., Karlsen, C., Sørum, H., Parkhill, J., Willassen, N. P., and Thomson, N. R. (2015). Co-cultivation and transcriptome sequencing of two co-existing fish pathogens *Moritella viscosa* and *Aliivibrio wodanis*. *BMC Genomics* 16:447. doi: 10.1186/s12864-015-1669-z

Hjerde, E., Lorentzen, M., Holden, M. T., Seeger, K., Paulsen, S., Bason, N., et al. (2008). The genome sequence of the fish pathogen *Aliivibrio salmonicida* strain LFI1238 shows extensive evidence of gene decay. *BMC Genomics* 9:616. doi: 10.1186/1471-2164-9-616

Huson, D. H., and Bryant, D. (2005). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

Karlsen, C., Vanberg, C., Mikkelsen, H., and Sørum, H. (2014). Co-infection of Atlantic salmon (*Salmo salar*), by *Moritella viscosa* and *Aliivibrio wodanis*, development of disease and host colonization. *Vet. Microbiol.* 171, 112–121. doi: 10.1016/j.vetmic.2014.03.011

Kashulin, A., Seredkina, N., and Sørum, H. (2017). Cold-water vibriosis. The current status of knowledge. *J. Fish Dis.* 40, 119–126. doi: 10.1111/jfd.12465

Killer, J., Mekadim, C., Bunešová, V., Mrázek, J., Hroncová, Z., and Vlková, E. (2020). Glutamine synthetase type I (glnAI) represents a rewarding molecular marker in the classification of bifidobacteria and related genera. *Folia Microbiol.* 65, 143–151. doi: 10.1007/s12223-019-00716-0

Konstantinidis, K. T., and Tiedje, J. M. (2005). Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol.* 187, 6258–6264. doi: 10.1128/JB.187.18.6258-6264.2005

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Lan, Y., Rosen, G., and Hershberg, R. (2016). Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains. *Microbiome* 4:18. doi: 10.1186/s40168-016-0162-5

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* 30, 3276–3278. doi: 10.1093/bioinformatics/btu531

Lunder, T., Sorum, H., Holstad, G., Steigerwalt, A. G., Mowinckel, P., and Brenner, D. J. (2000). Phenotypic and genotypic characterization of *Vibrio viscosus* sp. nov. and *Vibrio wodanis* sp. nov. isolated from Atlantic salmon (*Salmo salar*) with "winter ulcer". *Int. J. Syst. Evol. Microbiol.* 50, 427–450. doi: 10.1099/00207713-50-2-427

Machado, H., and Gram, L. (2015). The fur gene as a new phylogenetic marker for Vibrionaceae species identification. *Appl. Environ. Microbiol.* 81, 2745–2752. doi: 10.1128/AEM.00058-15

Machado, H., and Gram, L. (2017). Comparative genomics reveals high genomic diversity in the genus Photobacterium. *Front. Microbiol.* 8:1204. doi: 10.3389/fmicb.2017.01204

Nishiguchi, M. K. (2000). Temperature affects species distribution in symbiotic populations of *Vibrio* spp. *Appl. Environ. Microbiol.* 66, 3550–3555. doi: 10.1128/AEM.66.8.3550-3555.2000

Parks, D. H., Chuvochina, M., Chaumeil, P. A., Rinke, C., Mussig, A. J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for bacteria and Archaea. *Nat. Biotechnol.* 38, 1079–1086. doi: 10.1038/s41587-020-0501-8

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P. -A., et al. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36, 996–1004. doi: 10.1038/nbt.4229

Parte, A. C. (2018). LPSN-list of prokaryotic names with standing in nomenclature (Bacterio.net), 20 years on. *Int. J. Syst. Evol. Microbiol.* 68, 1825–1829. doi: 10.1099/ijsem.0.002786

Pascual, J., Macián, M. C., Arahal, D. R., Garay, E., and Pujalte, M. J. (2010). Multilocus sequence analysis of the central clade of the genus vibrio by using the 16S rRNA, recA, pyrH, rpoD, gyrB, rctB and toxR genes. *Int. J. Syst. Evol. Microbiol.* 60, 154–165. doi: 10.1099/ijs.0.010702-0

Rosselló-Mora, R., and Amann, R. (2001). The species concept for prokaryotes. *FEMS Microbiol. Rev.* 25, 39–67. doi: 10.1111/j.1574-6976.2001.tb00571.x

Sawabe, T., Kita-Tsukamoto, K., and Thompson, F. L. (2007). Inferring the evolutionary history of vibrios by means of multilocus sequence analysis. *J. Bacteriol.* 189, 7932–7936. doi: 10.1128/JB.00693-07

Sawabe, T., Ogura, Y., Matsumura, Y., Feng, G., Amin, A. R., Mino, S., et al. (2013). Updating the vibrio clades defined by multilocus sequence phylogeny: proposal of eight new clades, and the description of *Vibrio tritonius* sp. nov. *Front. Microbiol.* 4:414. doi: 10.3389/fmicb.2013.00414

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Smith, M. R. (2020). Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. *Bioinformatics* 36, 5007–5013. doi: 10.1093/bioinformatics/btaa614

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., et al. (2015). Structure and function of the global ocean microbiome. *Science* 348:1261359. doi: 10.1126/science.1261359

Thompson, F. L., Iida, T., and Swings, J. (2004). Biodiversity of Vibrios. *Microbiol. Mol. Biol. Rev.* 68, 403–431. doi: 10.1128/mmbr.68.3.403-431.2004

Urbanczyk, H., Ast, J. C., Higgins, M. J., Carson, J., and Dunlap, P. V. (2007). Reclassification of *Vibrio fischeri*, *Vibrio logei*, *Vibrio salmonicida* and *Vibrio wodanis* as *Aliivibrio fischeri* gen. nov., comb. nov., *Aliivibrio logei* comb. nov., *Aliivibrio salmonicida* comb. nov. and *Aliivibrio wodanis* comb. nov. *Int. J. Syst. Evol. Microbiol.* 57, 2823–2829. doi: 10.1099/ijs.0.65081-0

Verma, S. C., and Miyashiro, T. (2013). Quorum sensing in the squid-vibrio symbiosis. *Int. J. Mol. Sci.* 14, 16386–16401. doi: 10.3390/ijms140816386

Visick, K. L. (2009). An intricate network of regulators controls biofilm formation and colonization by *Vibrio fischeri*: MicroReview. *Mol. Microbiol.* 74, 782–789. doi: 10.1111/j.1365-2958.2009.06899.x

Yoshizawa, S., Karatani, H., Wada, M., Yokota, A., and Kogure, K. (2010). *Aliivibrio sifi* ae sp. nov., luminous marine bacteria isolated from seawater. *J. Gen. Appl. Microbiol.* 56, 509–518. doi: 10.2323/jgam.56.509

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.