Marine

Microbial Biodiversity,

Bioinformatics & Biotechnology

Grant agreement n°287589

Acronym: Micro B3

Start date of project: 01/01/2012, funded for 48 month

Deliverable 5.13

Complete MB3-IS Documentation

Version: 1.0

Circulated to: Authors and Coordinator

Approved by: Prof. Frank Oliver Glöckner, 04/02/2016

Expected Submission Date: 31/12/2015

Actual Submission Date: 04/02/2016

Lead Party for Deliverable: MPI Bremen
Mail: rkottman@mpi-bremen.de Tel.: +49(0)4212028974

Public (PU)	X
Restricted to other programme participants (including the Commission Services) (PP)	
Restricted to a group specified by the consortium (including the Commission Services) (RE)	
Confidential, only for members of the consortium (including the Commission Services) (CO)	

The Micro B3 project has received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement no 287589 (Joint Call OCEAN.2011-2: Marine microbial diversity – new insights into marine ecosystems functioning and its biotechnological potential).
The Micro B3 project is solely responsible for this publication. It does not represent the opinion of the EU. The EU is not responsible for any use that might be made of data appearing herein.

Summary

The complete Micro B3 Information System (Micro B3-IS) is a set of modular and interoperable software components which together implement a “bioinformatics scientific discovery workflow” from data generation to gaining new insights. The Micro B3-IS was successfully used for Micro B3’s Ocean Sampling Day (OSD) and the accompanying MyOSD citizen science campaigns. The modular and interoperable approach of Micro B3-IS builds on shoulders of existing European infrastructures such as European Nucleotide Archive and SeaDataNet to not re-invent existing components. The capability to also use one or more of the Micro B3-IS components for different purposes other than OSD or MyOSD is another advantage of this modular approach. If researchers want to perform own sampling campaigns with the aim to follow the “Marine microbial biodiversity, bioinformatics and biotechnology (M2B3) data reporting and service standards” (M2B3) they can build on the OSD Registry sample registration component and also make use of the OSD Smartphone App. Each bioinformatics pipeline like e.g. EBI’s Metagenome Portal can be used independently of all other components. The PostBIS database extension is directly usable for any kind of DNA or Protein data. The ProX tool for large-scale network-visualization can be used for many kinds of different large-scale biological networks. The tools for the ecological analysis of microbial diversity data GUSTAME and MASAME are usable stand-alone. These are just examples to illustrate the value of the modular architecture of Micro B3-IS.

Additionally, all but one component are open source with a permissive license (Apache 2 License). This allows everyone to make free use of every component for all kind of purposes. First, it allows scientists to scrutinize the software w.r.t. scientific quality. Second, any commercial entity can use any component for free on a legally clear and save basis for any business goal.

Table of Content

Contents

Introduction

This deliverable gives a complete overview of the major results of the Micro B3 Information System (Micro B3-IS). The general definition of an Information System is:

> *"**Information system**, an integrated set of components for collecting, storing, and processing data and for delivering information, knowledge, and digital products."*

More specifically in the context of Micro B3 this information system gives users access to an integrated view of microbial diversity and function in the marine environment in order enable users from biotechnology as well as ecosystems research to exploit information on microbial communities by effectively managing, analyzing, and sharing genomic and metagenomic data.

Contrary to the general understanding that information systems are developed for- and within single organizations, the Micro B3-IS has to work across organizations. Therefore, it is not a single application nor is there the single web page. From the very beginning it was designed to be modular with interoperable components[1] for specific aspects of the general Micro B3 OSD and MyOSD scientific discovery workflow (Figure 2). This had from the beginning the advantage to build on shoulders of existing European infrastructures such as European Nucleotide Archive, SeaDataNet and Pangaea and to not re-invent existing components. The capability to also use one or more of the Micro B3-IS components for different purposes other than OSD or MyOSD is another advantage of this modular approach. If researchers want to perform own sampling campaigns with the aim to follow the M2B3 Standard they can build on the OSD Registries sample registration component and also make use of the OSD Smartphone app. The ProX tool for large-scale network-visualization can be used for many kinds of different networks. PostBIS is directly usable for any kind of DNA or Protein data. These are just examples to illustrate the value of the modular architecture of Micro B3-IS.

Additionally, almost all components are open source under a permissive license (mostly Apache 2 License). This allows everyone to make free use of every component for all kind of purposes. First, it allows scientists to scrutinize the software w.r.t. scientific quality. Second, any commercial entity can use any component for free on a legally clear and save basis for any business goals.

[1] The term "component" has different meanings in different contexts. In order to match the use of the term "component" in the definition of information system "an individual software component is a software package, a web service, a web resource, or a module that encapsulates a set of related functions (or data)" (see also https://en.wikipedia.org/w/index.php?title=Component-based_software_engineering&oldid=702983327)

Overview

The whole of Micro B3 could be described in many ways. Over time it turned out that an effective way to describe the main aspects of MB3-IS is from a “data workflow for analytics” perspective. This perspective is used e.g. in the design of Big Data systems as depicted in Figure 1.

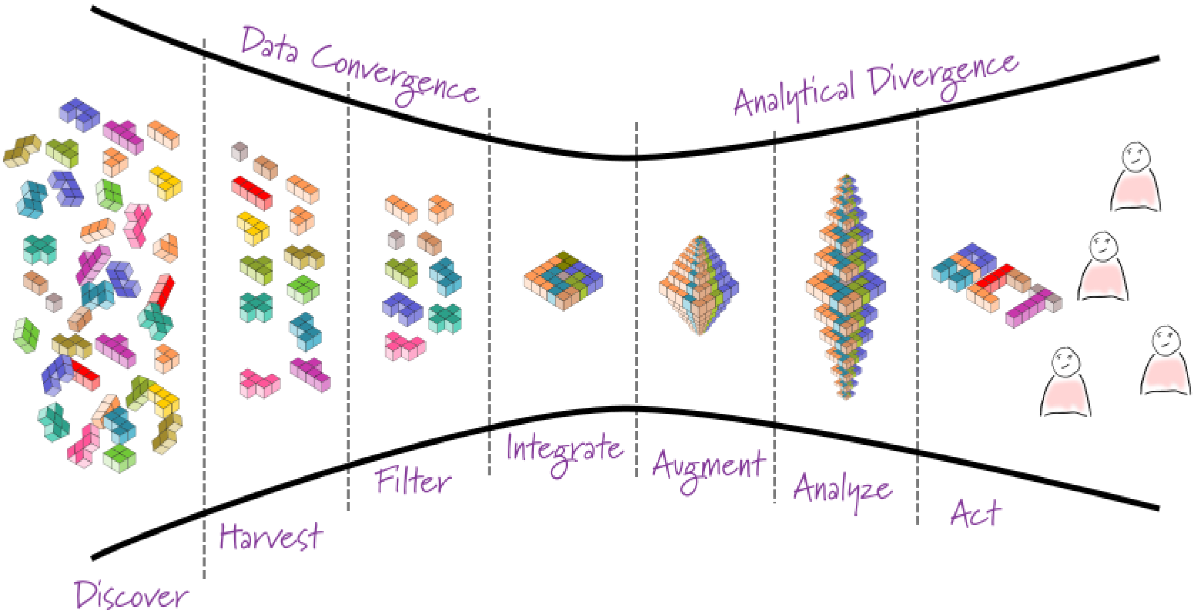

Figure 1: Data workflow for analytics perspective in big data systems. Courtesy of M. Fowler (http://martinfowler.com/articles/bigData/#analytics-workflow)

It basically divides every data workflow in three parts:

1. Data gathering from heterogeneous resources
2. Integration
3. Analysis

The data gathering part is characterized by data convergence i.e. different kind of data is unified towards an integrated view on data. The steps in this part are to 1) discover and/or generate data 2) harvest the data from where they are and 3) filter and unify the data. Whereas the integration part is in the center, it does not imply that all data needs to be integrated. However, an integrated view on the relevant data simplifies further analysis in the second part which can also include simple and complex visualizations of data summaries. Hence, it is characterized by analytical divergence i.e. new data is generated as the result from many different kind of analysis on the integrated data. The steps in the analysis part involve 1) augment and 2) analyze data. The last “Act” step is special because it has different meanings in different contexts. In the typical business case, act means that the analysis reveals new insights which let business act upon e.g. by changing marketing strategy. In a more scientific context it can either simply mean that the data is taken for further investigation out site the current data workflow. Or it can mean that the result of an analysis reveals new insights about the studied object. Therefore, the last step can be seen as "act of acquiring knowledge".

Overall, it is a simplified scheme, which leaves out complex data management and lifecycle aspects. However, on this level of abstraction it matches some typical scientific discovery workflows which generate data through field studies and experiments (discover), take the data from successful experiments (harvest), check and clean the data (filter), combine the data with other data from other sources or previous experiments (integrate), add new data (augment) and then analyse the relevant data to gain new insights for publication (act).

In the context of Micro B3, the information system delivers an integrated set of software components for complex scientific discovery workflows of large-scale sampling campaigns such as OSD and MyOSD 2014 and 2015. Therefore the remainder of this documented is structured according to the steps of the data workflow as depicted in Figure 2.

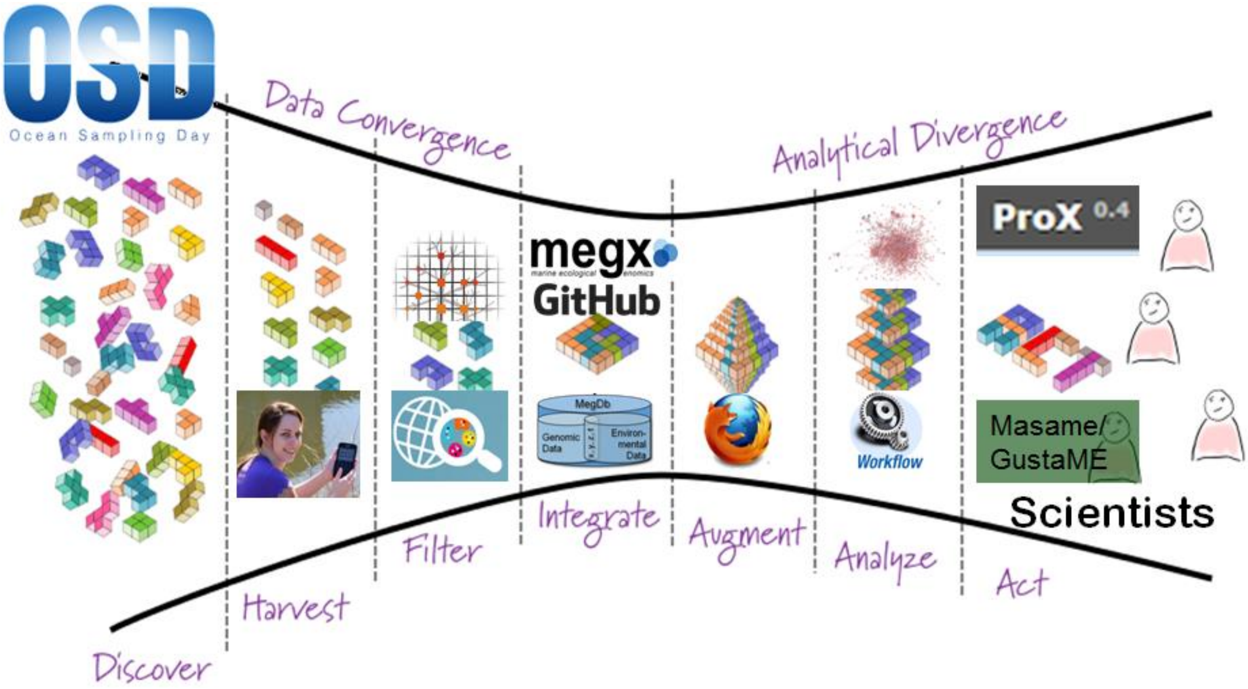

Figure 2: Scientific discovery workflow including icons of some of the software components of Micro B3-IS.

Technically, a component is a piece of software, which can also be installed and used on its own with only few dependencies instead of the whole information system. Throughout the following sections each component will be summarized in a "Component Description Table" with the structure as described in Table 1 below.

Table 1: Structure and explanation of the Component Description Table

Name	Name of the component
Description	Short description of the purpose and scope of the component
Maturity Status	The level of how far the component is developed: • Early prototype (some first functionality implemented)

	• Prototype (basic functionality implemented) • Demonstrator (main functionality implemented) • Production (it is in full use)
Owner(s)	Who has the Intellectual Property Rights on the component
Contact:	The main contact person for further information
Source Code:	Link to the source code of the component if available
License:	License under which the component is or will be released
Further links and documentation	Links to further documentation and other related information

Discover

The discovery phase in the context of Micro B3's Ocean Sampling Days and MyOSDs is defined as the organization of these events and the actual sampling done by the participants. For this step, the main relevant aspects for the design of the Micro B3-IS are which data will be generated when by whom and when will which material and data get deposited. Already since the OSD pilots, which started end 2012, an OSD Sites Registry was established to collect data about which marine station intends to sample at which geographic sites and who are the responsible institutes and contact persons (deliverable 5.7). This allowed keeping an overview and managing the communication with all OSD participants. Only late and after lessons learned from OSD pilot events the logistics of OSD were fixed as shown in a simplified scheme in Figure 3. The OSD participants send the samples to a central place for DNA Extraction (in both years samples were sent to MPI-MM and DNA was extracted at AWI). DNA was extracted from most filters for sequencing at a central sequencing facility that sent the raw sequence data on hard drives per mail delivery. In parallel, one filter of each sample, which was excluded from DNA extraction, was labeled with barcodes of the Smithsonian National Museum of Natural History/NMNH Biorepository for long-term bio-archiving. In parallel, all contextual (meta)data of each sample gets registered in the OSD Registry (see Table 2) by OSD participants. After curation of the contextual data in the OSD Registry, the data gets submitted to various archives and the Smithsonian. Overall, the sampling protocol, a description of which measurements should be reported, and the logistics are well defined in the OSD Handbook (deliverable 4.3).

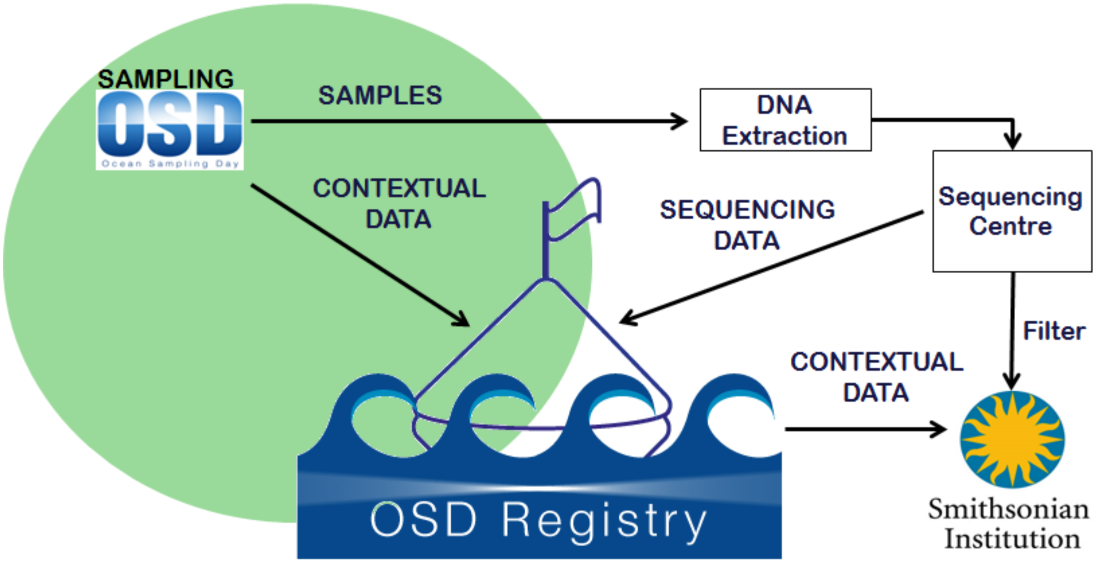

Figure 3: Logistic aspect of OSD sampling, DNA extraction, bio-archiving and sequencing.

Harvest/Gathering

Ocean Sampling Day

This step is mainly concerned with "early consistent data acquisition" of OSD meta- and sequence data. The scientific participants first had to send the data on manually written log-sheets together with the samples and second fill-out the online sample registration web-form. Both were designed according to the "Marine Microbial Biodiversity, Bioinformatics and Biotechnology" (M2B3) standard [1]. This "double bookkeeping" was originally decided as a means to have an analog back-up of contextual data in case of a disaster with the online sample registration form and database. However, it turned out that this strategy was instead instrumental for the manual curation and quality check of the contextual data. It allowed discovering inconsistencies between the different records by checking the online submitted data against the log-sheets and to communicate with the participants to correct the data.

In 2014, the citizens were just asked to submit data via the OSD Smartphone App. In 2015 the MyOSD campaign included microbial sampling. Therefore, the same double bookkeeping strategy was used as well: citizens were asked to submit data via the app as well as filling-out a MyOSD log-sheet and return it back together with the samples.

The OSD Registry as a whole has a landing page for people interested (http://mb3is.megx.net/osd-registry) and serves the overview of scientific sites that are participating (http://mb3is.megx.net/osd-registry/list). The data submitted via the OSD Smartphone App is displayed at http://mb3is.megx.net/osd-app/samples.

Minimum Information about a Biosynthetic Gene cluster (MIBiG)

A wide variety of enzymatic pathways that produce specialized metabolites in bacteria, fungi and plants are known to be encoded in biosynthetic gene clusters. Information about these clusters, pathways and metabolites is currently dispersed throughout the literature, making it difficult to exploit. To facilitate consistent and systematic deposition and retrieval of data on biosynthetic gene clusters, therefore the Minimum Information about a Biosynthetic Gene cluster (MIBiG) data standard was proposed [2]. For submission of new MIBiG-compliant data by scientists in the field, an interactive online submission form (available from http://mibig.secondarymetabolites.org) was prepared, which was extensively tested through the community annotation effort. The web service and database for storing the MIBiG data was developed and is hosted as part of Micro B3-IS. This demonstrates the modularity and extensibility of Micro B3-IS.

Table 2: OSD Registry

Name	OSD Registry
Description	Web pages and online web-forms for the submission and display of OSD and MyOSD contextual data.
Maturity Status	Production
Owner(s)	OXFORD, MPI-MM, Interworks
Contact:	Renzo Kottmann
Source Code:	https://github.com/MicroB3-IS/megxnet/tree/master/net.megx.osd.registry
License:	Apache 2
Further links and documentation	Available at: http://mb3is.megx.net/osd-registry

Table 3: OSD Smartphone App

Name	OSD Smartphone App
Description	The OSD App is specially designed for people who like to participate in the Ocean Sampling Day events With this App everyone can enter measured data, observations and photos of the oceans and seas directly into their phones and send it to the OSD App Server to make it publicly available for everyone.
Maturity Status	Production
Owner(s)	Interworks, MPI-MM
Contact:	Aleksandar Memca
Source Code:	Available upon request
License:	Not determined
Further links and documentation	Tutorial video: https://youtu.be/1lhDdPbzuTs Google Play Store: https://play.google.com/store/apps/details?id=com.iw.esa&hl=en

	Apple Store: https://itunes.apple.com/us/app/osd-citizen/id834353532?mt=8

Table 4: Minimum Information about a Biosynthetic Gene cluster (MIBiG) Service

Name	**Minimum Information about a Biosynthetic Gene cluster (MIBiG) Service**
Description	Web Service and storage of MIBiG data from MIBiG-compliant data submissions by scientists.
Maturity Status	Production
Owner(s)	University Wageningen, MPI
Contact:	Marnix M. Medema, Renzo Kottmann
Source Code:	https://github.com/MicroB3-IS/megxnet/tree/master/net.megx.mibig
License:	Apache 2
Further links and documentation	Submission form: http://mibig.secondarymetabolites.org

Filter

The purpose of this step is to filter from the harvested set of data the subset which is relevant for integration and further analysis. In the context Micro-B3-IS this entails the pre-processing of the sequence data and the part of the bioinformatics analysis where the main output are taxonomic assignments and gene function lists (see deliverables 5.8 and 5.88). Only these bioinformatics analysis combined with the curation and checking of the metadata allow determining which sequence data from which samples qualify for full integration and further analysis. Moreover, detailed bioinformatics processing of the sequences derived from marine samples and isolated marine organisms is required to turn high volumes of data into useful, scientifically meaningful information, ready to be interpreted by the marine science community.

Figure 4 gives an overview of which filtering and sequencing feeds which bioinformatics pipeline and the major types of outputs. The orange left part was exploited by OSD/MyOSD. They all have a common pre-processing pipeline upstream. This pipeline takes the sequences from the sequencing machine as input and performs common tasks such as adapter and primer clipping, trimming, merging and length filtering with strict quality parameters. The goal of formalizing this pipeline is to provide all OSD participants (and the whole scientific community) with a single, quality-controlled dataset in order to ensure comparability and repeatability of analysis results (see Table 5 for references to detailed documentation).

The metagenome sequencing from the prokaryote fraction was analyzed by the EBI Metagenomics pipeline [3][4] and Metagenomic Traits pipeline (MG Traits in Figure 4). All amplicon data from the prokaryotic (16S and 18S) and eukaryotic fraction (18S) were analyzed by the SILVAngs pipeline [5]. Detailed results of the OSD 2014 analysis are documented on the OSD Community Analysis Collaboration Pages at https://github.com/MicroB3-IS/osd-analysis/wiki.

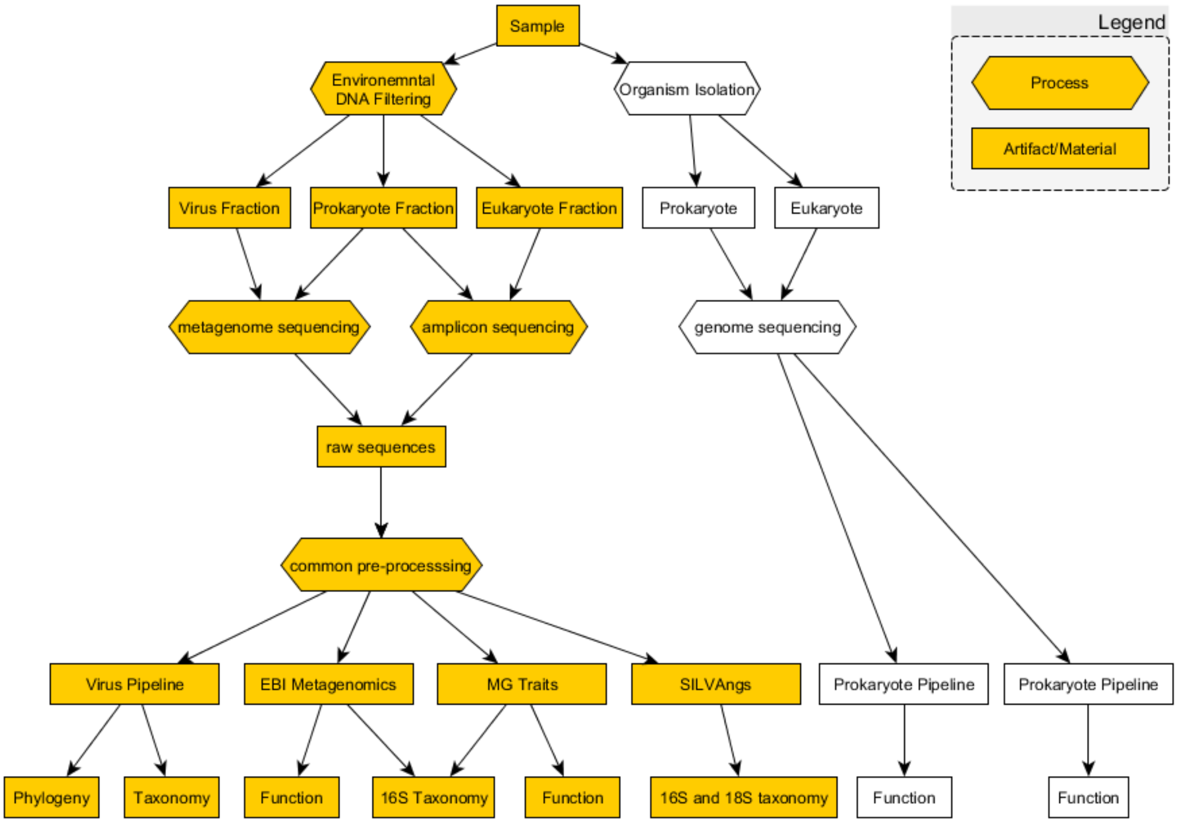

Figure 4: Overview of sampling and sequencing and which pipeline produces which main type of output. Rectangles are either material or digital artifacts and hexagons indicate processes. All boxes in orange indicate which process and artifacts where actually used for OSD and MyOSD.

Table 5: OSD Pre-Processing Pipeline

Name	**OSD Pre-Processing Pipeline**
Description	Creates a subset of the of the original raw sequence data which fulfils defined sequence quality criteria, suitable for further analyses.
Maturity Status	Production
Owner(s)	MPI-MM, Jacobs
Contact:	Antonio Fernandez-Guerra
Source Code:	https://colab.mpi-bremen.de/micro-b3/svn/analysis-scripts/trunk/osd-analysis/osd-pre-processing/
License:	Apache 2
Further links and documentation	Online Documentation: https://github.com/MicroB3-IS/osd-analysis/wiki/Sequence-Data-Pre-Processing

Table 6: EBI Metagenomics Portal

Name	**EBI Metagenomics Portal**
Description	The EBI Metagenomics service is an automated pipeline for the analysis and archiving of metagenomic data that aims to provide insights into the phylogenetic diversity as well as the functional and metabolic potential of a sample.
Maturity Status	Production
Owner(s)	EMBL-EBI
Contact:	Rob Finn
Source Code:	Not applicable
License:	Not applicable
Further links and documentation	https://www.ebi.ac.uk/metagenomics and https://www.ebi.ac.uk/metagenomics/about

Table 7: MB3-Virus pipeline

Name	**MB3-Virus pipeline**
Description	The first virus pipeline module developed is designed to identify sequences of likely viral origin in environmental shotgun ('metagenomic') sequence data.
Maturity Status	Demonstrator
Owner(s)	IGS (CNRS partner 6)
Contact:	Pascal Hingamp
Source Code:	Not applicable
License:	Not determined
Further links and documentation	Deliverable 5.8

Table 8: Metagenomic Traits Pipeline

Name	**Metagenomic Trait Pipeline**
Description	The metagenomic trait pipeline calculates a number of ecologically interesting traits of bacterial communities as observed by high-throughput metagenomic DNA sequencing.
Maturity Status	Demonstrator

Owner(s)	MPI-MM
Contact:	Antonio Fernandez-Guerra
Source Code:	Web page: https://github.com/MicroB3-IS/megxnet/tree/master/net.megx.mg-traits
License:	Apache 2
Further links and documentation	Overview:https://colab.mpi-bremen.de/wiki/display/microb3/Metagenomic+Traits Web service Documentation https://www.biodiversitycatalogue.org/services/64

Table 9

Name	**Eukaryotes**
Description	An annotation pipeline able to structurally annotate any marine eukaryotic genome, especially protists.
Maturity Status	Demonstrator
Owner(s)	Genoscope
Contact:	Olivier Jaillon
Source Code:	Not applicable
License:	Not applicable
Further links and documentation	Deliverable 5.8 and 5.88

Table 10: SILVAngs

Name	**SILVAngs**
Description	SILVAngs (SILVA Next Generation Sequencing) provides fast and accurate taxonomic classification of next generation sequencing amplicon data.
Maturity Status	Production
Owner(s)	JacobsUni, MPI-MM, Ribocon
Contact:	Frank Oliver Glöckner: ngs-contact@arb-silva.de
Source Code:	Not applicable
License:	https://www.arb-silva.de/silva-license-information/
Further links and	https://www.arb-silva.de/ngs and SilvaNGS user guide at https://www.arb-

documentation	silva.de/ngs/service/file/?file=SILVAngs_User_Guide_15_12_15.pdf

Table 11: **MB3-Prokaryotic pipeline**

Name	**MB3-Prokaryotic Pipeline**
Description	The MB3-Prokaryotic pipeline is an in-house hosted pipeline for the analysis and automatic annotation of assembled bacterial and archaeal genomes.
Maturity Status	Demonstrator
Owner(s)	MPI-MM
Contact:	Renzo Kottmann
Source Code:	Not applicable
License:	Free to use
Further links and documentation	Deliverable 5.8

Integrate

The data of the OSD scientific discovery workflow requires to integrate contextual data about the environment as well as sequence data from initial sampling up to web based end-user access. This integration is also required to be able to generate various data products derived from the data by the various analysis pipelines (see above).

MegDb

The Microbial Ecological Genomics Database (MegDb) is underlying most of the Micro B3 Information System components. Currently, MegDb hosts data for components such as user management, Ocean Sampling Day data, Metagenomic Traits, environmental data layers, sequence data, PubMap, ProX, data on Biosynthetic Gene Clusters [2] among others. The aim of MegDb is to guaranty consistent storage of all data needed for general marine microbial ecosystems research and biotechnology as well of for the special use case Ocean Sampling Day. The technical design and implementation draws from conclusions and requirements of WP 3-7 and implements the M2B3 standard [1]. Although, currently MegDb is physically a single database for the management of data for and from many Micro B3-IS components, it is logically designed in a way that physical separation in different databases is possible at any given time.

PostBIS: efficient storage of sequence data

PostgreSQL Bioinformatics Information System (PostBIS) was developed by MPIMM. Main development was by Michael Schneider in his master Thesis “Efficient Representation of Biological Sequences in a Relational Database Management System” in collaboration with Prof. Dr. Stefan Kurtz University of Hamburg [6].

PostBIS is an open-source PostgreSQL extension project to facilitate sequence-based Bioinformatics in PostgreSQL. It offers:

- Highly efficient specialized sequence data types incl.
    - nucleotide sequences
    - protein sequences
    - alignments
- additional domain-specific functions
- indexing of sequences

The storage requirements for sequence data using PostBIS are around 2 bits per nucleotide base which is around 25% of PostgreSQL native text data type and around 5 bits per amino acid compared to around 9 bits for PostgreSQL’s native text data type (Figure 5).

The loading of complete sequence databases using PostBIS is also significantly faster in all tested cases (Figure 6). All in all, query Access is around 1000x faster compared to using PostgreSQL native text data type.

PostBIS is used in production by MegDb since more than two years and no issues were encountered.

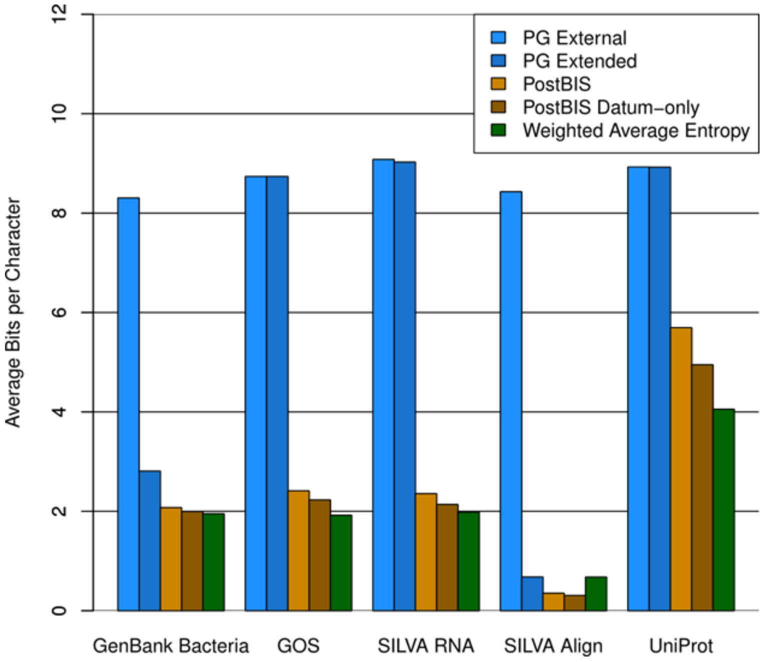

Figure 5: See the PostBIS space-efficiency (brownish) compared to PostgreSQL's built-in compression (blue). The graphic shows results for complete genomes (GenBank Bacteria), short reads (GOS), RNA sequences (SILVA RNA), aligned RNA sequences (SILVA Align) and amino acid sequences (UniProt).

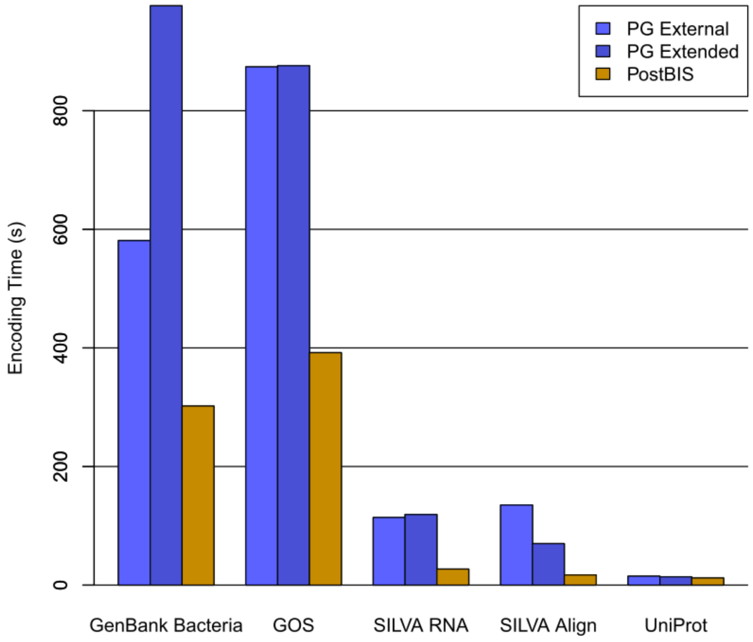

Figure 6: Encoding time (s) of complete sequence databases without and with PostBIS. Tested data are: GenBank Bacteria, Global Ocean Survey metagenomes (GOS), SILVA RNA unaligned and aligned and UniProt protein database.

Table 12

Name	PostBis
Description	PostgreSQL extension for efficient storage molecular sequence data
Maturity Status	Production
Owner(s)	MPI-MM
Contact:	Renzo Kottmann
Source Code:	https://colab.mpi-bremen.de/postbis/svn/trunk/
License:	PostgreSQL License
Further links and documentation	https://colab.mpi-bremen.de/wiki/display/pbis/PostBIS

Table 13: MegDb

Name	MegDb
Description	Microbial Ecological (Meta)Genomics Database (MegDb) for the integration of environmental and molecular sequence data
Maturity Status	Production
Owner(s)	MPI-MM
Contact:	Renzo Kottmann
Source Code:	Upon request
License:	Apache 2
Further links and documentation	http://resources.megx.net/megdb-doc/index.html

Augment

PubMap

The World Map of Publications (PubMap) system was developed to crowdsource the creation of a georeferenced bibliography (see deliverable 5.11 for details). This includes locations worldwide and it is not limited to certain journals or fields of study. PubMap has two main components, the PubMap Curator (PMC) a bookmarklet for manual georeferenced annotation of publications and the PubMap Browser (PMB) a web application to retrieve those georeferenced publications. Each publication should be annotated depending on the origin of the study material not on the location of the university or institute at which the study was conducted. For example, if a publication discusses a water sample which was taken in the Pacific Ocean but the measurements and analyses were done at Harvard University, the publication should be georeferenced with the location of the Pacific Ocean and not of Harvard. The idea is to create a database which can be filled with information by volunteers and the stored information is openly available and can be retrieved by everyone. PubMap is based on Chon CMS and hence integrated in the Micro B3 Information System (Micro B3-IS). PubMap provides an easy and straightforward crowdsourcing approach to annotate scientific publications with geographic coordinates and place names. It is designed such that users do not have to spend much time and effort on the annotation. This should help to increase the motivation and willingness of users to curate several publications. PubMap is in beta version to see if further improvements can be made. At the moment, PubMap only works with abstract websites of PubMed, which restricts the publications variety. There are millions of publications out there, alone PubMed comprises more than 24 million citations at this date, and it is said that scientific output doubles every nine years. Accordingly, it will not be possible to georeference them all manually. To overcome the highly time demanding work of georeferencing publications by hand text mining would be the most straightforward alternative. Although, text mining can save a lot of manual work, first a good algorithm needs to be developed and properly tested. Beforehand though, a gold standard corpus of manually annotated publications has to be developed. The data collected by PubMap could function as such a corpus and help to develop a data mining tool to extract geographic information out of scientific publications.

Table 14: PubMap

Name	PubMap
Description	PubMap provides an easy and straightforward crowdsourcing approach to annotate scientific publications with geographic coordinates and place names.
Maturity Status	Demonstrator
Owner(s)	MPI-MM
Contact:	Julia Schnetzer

Source Code:	https://github.com/MicroB3-IS/megxnet/tree/master/net.megx.pubmap
License:	Apache 2
Further links and documentation	Available at http://mb3is.megx.net/pubmap/

Analysis

ProX: Visualizing the Unknowns with Protein Explorer

As a part of task 7.1c a new approach to discover new potential functions with biotechnological applications using metagenomic data was developed. In collaboration with the University of Applied Sciences, Bremen Matthias Stock developed the Protein Explorer (ProX application) [7]. It is entirely written in JavaScript and hosted on megx.net as Chon CMS plugin. Although ProX can be used for any kind of graph, its main purpose is to allow all users to explore known-unknown networks based on PFAMs. Performance measurements and comparisons to other products proofed that no library for exists that is neither powerful enough to render the given graph from in a web browser on commodity hardware nor as feature-complete as comparable applications. It has a built-in node search functionality based on the PFAM name or PFAM accession; it also retrieves metadata from each node from the PFAM web services and the user can create ego networks centered to their desired node. Currently, ProX is only available with special permission and is given on demand. A short video demonstrates the functionality of ProX and can be watched here https://www.dropbox.com/s/8ztqxtqhir726nu/prox.mp4?dl=0.

ProX is used by work package 7 (see deliverable 7.1) to graphically model the associations between the metagenomic known fraction (all open reading frames we can assign a function in terms of protein domains) with the unknown fraction (all potential open reading frames that could encode a protein but without any known function). At the moment we have used the GOS dataset to test the method but we are going to include all metagenomes available from TARA Oceans, Malaspina, and Human Microbiome Project among others to generate environment specific networks.

ProX then allows all users to explore the resulting known-unknown for new targets of biotechnological interest.

GUSTAME and MASAME

The popularity of multivariate analyses is continuing to increase and their application to microbial ecological data has become technically simplified; however, a developed and up-to-date understanding of their properties and limitations is still not widespread in the community. As a result, many microbial ecologists who are not equipped with deep numerical training face a ‘black box’ approach to multivariate analysis and the associated risks of misapplying techniques or misinterpreting results. Reviewers, too, often face uncertainty in evaluating whether researchers have performed appropriate analyses and produced fair interpretations of their results. To support and promote the constantly developing understanding of multivariate analyses in microbial ecology, the GUide to STatistical Analysis in Microbial Ecology (GUSTA ME; http://mb3is.megx.net/gustame) – an online, dynamically updated resource with content tailored to the needs of the microbial ecology community was published together with Multivariate AnalysiS Applications for Microbial Ecology (MASAME) suite [8]. GUSTA ME is an interactive ‘living’ review of multivariate analyses with specific relevance to the microbial ecology community. Its content offers an accessible resource for teaching and reference, while its implementation allows users to quickly locate and focus their efforts on analytical approaches pertinent to their investigations. Selected pages across GUSTA ME include links to interactive analysis applications which allow users to perform the technique or procedure discussed on that page, either on their own data sets (which may be uploaded as comma-separated-value files) or on preloaded example data. Collectively, these applications are referred to as the Multivariate AnalysiS Applications for Microbial Ecology (MASAME) suite (http://mb3is.megx.net/masame/). MASAME applications are accessed through user-friendly web-pages, rendered by the *shiny* package, which call upon numerous functions from well-known packages belonging to the statistical programming environment and language, R. For example, (partial) RDA, (partial) CCA, NMDS, and PCNM methods from the *vegan* package are combined with supporting functions which allow data transformations using standard and ecologically meaningful methods, plotting, and download functionality on a single webpage. Users need not know the R language, as point-and-click interfaces are common to all MASAME applications. Such tools add a practical complement to GUSTA ME's review of multivariate analysis techniques and are easily enhanced to address new needs as they arise.

As it further develops, GUSTA ME has the potential to become a focal repository for accessible analytical knowledge and debate in microbial ecology, wherein methods that have become central to ecology, as well as their criticisms may be easily explored. GUSTA ME and MASAME complement the Micro B3-IS' data management, integration, and processing modules by providing support in the analysis of integrated data; however, both resources may also be used independently.

Table 15

Name	**ProX**
Description	Efficient web-based visualization of complex protein networks
Maturity Status	Demonstrator
Owner(s)	MPI-MM
Contact:	Renzo Kottmann
Source Code:	https://github.com/MicroB3-IS/megxnet/tree/master/net.megx.prox
License:	Apache 2
Further links and documentation	Demonstration Video: https://www.dropbox.com/s/8ztqxtqhir726nu/prox.mp4?dl=0

Table 16

Name	Masame
Description	Multivariate AnalysiS Applications for Microbial Ecology (MASAME). Interactive applications for statistical analysis.
Maturity Status	Production
Owner(s)	AWI
Contact:	Pier Luigi Buttigieg
Source Code:	Not applicable
License:	Not determined
Further links and documentation	http://mb3is.megx.net/masame/

Table 17

Name	Gustame
Description	GUSTA ME is an interactive ‘living’ review of multivariate analyses with specific relevance to the microbial ecology community
Maturity Status	Production
Owner(s)	AWI
Contact:	Pier Luigi Buttigieg
Source Code:	Not applicable
License:	Not determines
Further links and documentation	http://mb3is.megx.net/gustame

Act

The last “Act” step is special because it can mean different things depending on the context. In the typical business case, act means that the analysis reveals new insights which let business act upon e.g. by changing marketing strategy. In a more scientific context of Micro B3, it either simply mean that the data is taken for further investigation out site the current data workflow. Or it means that the result of an analysis reveals new insights about the studied object. Therefore, the last step can be seen as “act of acquiring knowledge”. Recalling the definition of the Information System from the introduction, the Micro B3-IS delivers information, knowledge, and digital products throughout the whole scientific discovery workflow. Act is simply the aspect of making use of Micro B3-IS.

Base Components

Several components and development aspects of the Micro B3-IS are not directly reflected by the whole scientific discovery workflow which were described thus far.

Chon CMS

Many components of the whole scientific discovery workflow are based on the underlying Chon Content Management System for data-centric web application development (see deliverable 5.11 for technical details). Chon CMS itself is based on OSGi standard technology platform. By default it has plugins for simple web page content management and web services. This allows to not only build web applications as typically done by content management systems, but also web services utilizing the same base services as e.g. for accessing data from the underlying database.

Based on these several other components have been developed which enable interoperability, visualization, and creation of web based user interfaces as 3-Tier web application. Namely, components for information security are one important cross-cutting feature which are used by several Micro B3-IS components. For example the OSD Registry, PubMap and ProX use the underlying security mechanisms for authenticating and authorizing their users (see deliverable 5.11).

Table 18

Name	**Chon CMS**
Description	OSGi based web content management system
Maturity Status	Production
Owner(s)	Interworks
Contact:	Renzo Kottmann
Source Code:	https://github.com/MicroB3-IS/choncms

License:	Apache 2
Further links and documentation	https://github.com/MicroB3-IS/choncms

OSD Community Analysis Collaboration Pages and File Repository

An OSD Analysis Core Group (OACG) of 25 experts within Micro B3 was formally established in October 2014 to coordinate the analysis of all OSD data in line with the analysis pipelines devised by the Micro B3 Information System as well as the submission of all OSD raw sequences and metadata to relevant databases and their public distribution. The OACG needed ways to exchange documentation, source code and data files among all participants and in a transparent manner to the public. Hence, Micro B3 setup the "OSD Community Analysis Collaboration Pages" and the "OSD Analysis File Repository".

OSD Community Analysis Collaboration Pages

GitHub is used for source code sharing and Wiki based documentation of the intermediate analysis results. The main entry page is https://github.com/MicroB3-IS/osd-analysis/wiki. Currently there are 5 wiki pages for the main topics of discussion and documentation which are actively edited. An overview is given in https://github.com/MicroB3-IS/osd-analysis/wiki/Guide-to-OSD-2014-data**Error! Reference source not found.**. In addition, all issues and requests for additional data are actively managed by GiHub's issue tracker at https://github.com/MicroB3-IS/osd-analysis/issues. Important links in GitHub are:

- Overview of OSD 2014 data analysis at https://github.com/MicroB3-IS/osd-analysis/wiki/Guide-to-OSD-2014-data
- Details of the curated environmental data of OSD samples at https://github.com/MicroB3-IS/osd-analysis/wiki/OSD-2014-environmental-data-csv-documentation
- Documentation of OSD assemblies from metagenomes at https://github.com/MicroB3-IS/osd-analysis/wiki/OSD-assemblies
- Documentation of OSD Pre-Processing pipeline at https://github.com/MicroB3-IS/osd-analysis/wiki/Sequence-Data-Pre-Processing

OSD Analysis Files Repository

The original sequence and environmental data are archived at ENA and PANGAEA respectively. However, many more kinds of files need to be shared for further analysis. These files are often in the size range of GBs and GitHub can be used only for file sizes in the range of MBs. Moreover, GitHub's policy does not allow for use as a file sharing platform.

Therefore, all files that are not archived and need to be shared are currently hosted at Max Planck Institute Bremen. The main entry point is the http://mb3is.megx.net/osd-files URL, which redirects to a publically shared directory on an OwnCloud instance. This indirection allows changing file location and hosting at any time without the need to change the URL. In fact, all documentation on the community pages uses the main URL as a basis to directly link to the relevant files or sub-directories. This relieves users from the need to understand the underlying directory structure.

Interoperability Aspects

Interoperability is a property of a system to work with other systems. An important goal of Micro B3 was to build on- and work with existing European infrastructures and other database providers such as SeaDataNet, European Nucleotide Archive (ENA), EuroBIS and Pangaea.

Therefore, plans were developed to establish and enhance data flow from Micro B3-IS to and between these existing infrastructures (D3.3 and D3.5). This implies what data gets exchanged how. The published data standard on reporting “Marine Microbial Biodiversity, Bioinformatics and Biotechnology” (M2B3) data was instrumental in defining the implementation independent semantics and format of the relevant data [1]. In other words this standard defines what data needs to be gathered and flow through the scientific discovery workflow of Micro B3-IS.

The general data exchange mechanisms are based on Web Services which are services mainly used for interoperable machine-to-machine communication. Consequently, numerous web services were developed in the context of the Micro B3-IS. Several new services were developed by existing infrastructures and are already described in several deliverables (D3.6, D4.4, D4.7, and D5.7) and as part of the M2B3 standard [1]. Notably, all agreed to use the existing industrial standard “OpenSearch” for data retrieval services and Open Geospatial Consortium standards like Web Map- and Web Feature Services (WMS and WFS) for geospatial mapping of data.

Most of the other new web services were developed in the context of Ocean Sampling Day and are hosted on megx.net (http://mb3is.megx.net). These are implemented following the RESTful architectural style [9].

References

[1] P. ten Hoopen, S. Pesant, R. Kottmann, A. Kopf, M. Bicak, S. Claus, K. Deneudt, C. Borremans, P. Thijsse, S. Dekeyzer, D. M. Schaap, C. Bowler, F. O. Glöckner, and G. Cochrane, “Marine microbial biodiversity, bioinformatics and biotechnology (M2B3) data reporting and service standards,” *Stand. Genomic Sci.*, vol. 10, no. 1, p. 20, May 2015.

[2] M. H. Medema, R. Kottmann, P. Yilmaz, M. Cummings, J. B. Biggins, K. Blin, I. de Bruijn, Y. H. Chooi, J. Claesen, R. C. Coates, P. Cruz-Morales, S. Duddela, S. Düsterhus, D. J. Edwards, D. P. Fewer, N. Garg, C. Geiger, J. P. Gomez-Escribano, A. Greule, M. Hadjithomas, A. S. Haines, E. J. N. Helfrich, M. L. Hillwig, K. Ishida, A. C. Jones, C. S. Jones, K. Jungmann, C. Kegler, H. U. Kim, P. Kötter, D. Krug, J. Masschelein, A. V Melnik, S. M. Mantovani, E. A. Monroe, M. Moore, N. Moss, H.-W. Nützmann, G. Pan, A. Pati, D. Petras, F. J. Reen, F. Rosconi, Z. Rui, Z. Tian, N. J. Tobias, Y. Tsunematsu, P. Wiemann, E. Wyckoff, X. Yan, G. Yim, F. Yu, Y. Xie, B. Aigle, A. K. Apel, C. J. Balibar, E. P. Balskus, F. Barona-Gómez, A. Bechthold, H. B. Bode, R. Borriss, S. F. Brady, A. A. Brakhage, P. Caffrey, Y.-Q. Cheng, J. Clardy, R. J. Cox, R. De Mot, S. Donadio, M. S. Donia, W. A. van der Donk, P. C. Dorrestein, S. Doyle, A. J. M. Driessen, M. Ehling-Schulz, K.-D. Entian, M. A. Fischbach, L. Gerwick, W. H. Gerwick, H. Gross, B. Gust, C. Hertweck, M. Höfte, S. E. Jensen, J. Ju, L. Katz, L. Kaysser, J. L. Klassen, N. P. Keller, J. Kormanec, O. P. Kuipers, T. Kuzuyama, N. C. Kyrpides, H.-J. Kwon, S. Lautru, R. Lavigne, C. Y. Lee, B. Linquan, X. Liu, W. Liu, A. Luzhetskyy, T. Mahmud, Y. Mast, C. Méndez, M. Metsä-Ketelä, J. Micklefield, D. A. Mitchell, B. S. Moore, L. M. Moreira, R. Müller, B. A. Neilan, M. Nett, J. Nielsen, F. O’Gara, H. Oikawa, A. Osbourn, M. S. Osburne, B. Ostash, S. M. Payne, J.-L. Pernodet, M. Petricek, J. Piel, O. Ploux, J. M. Raaijmakers, J. A. Salas, E. K. Schmitt, B. Scott, R. F. Seipke, B. Shen, D. H. Sherman, K. Sivonen, M. J. Smanski, M. Sosio, E. Stegmann, R. D. Süssmuth, K. Tahlan, C. M. Thomas, Y. Tang, A. W. Truman, M. Viaud, J. D. Walton, C. T. Walsh, T. Weber, G. P. van Wezel, B. Wilkinson, J. M. Willey, W. Wohlleben, G. D. Wright, N. Ziemert, C. Zhang, S. B. Zotchev, R. Breitling, E. Takano, and F. O. Glöckner, “Minimum Information about a Biosynthetic Gene cluster,” *Nat. Chem. Biol.*, vol. 11, no. 9, pp. 625–631, Aug. 2015.

[3] S. Hunter, M. Corbett, H. Denise, M. Fraser, A. Gonzalez-Beltran, C. Hunter, P. Jones, R. Leinonen, C. McAnulla, E. Maguire, J. Maslen, A. Mitchell, G. Nuka, A. Oisel, S. Pesseat, R. Radhakrishnan, P. Rocca-Serra, M. Scheremetjew, P. Sterk, D. Vaughan, G. Cochrane, D. Field, and S.-A. Sansone, “EBI metagenomics--a new resource for the analysis and archiving of metagenomic data.,” *Nucleic Acids Res.*, p. gkt961–, Oct. 2013.

[4] A. Mitchell, F. Bucchini, G. Cochrane, H. Denise, P. ten Hoopen, M. Fraser, S. Pesseat, S. Potter, M. Scheremetjew, P. Sterk, and R. D. Finn*, “EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D595–D603, 2016.

[5] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, “The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.,” *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D590–6, Jan. 2013.

[6] M. Schneider, “Efficient Representation of Biological Sequences in a Relational Database Management System,” 2012.

[7] M. Stock, “Efficient Web-Based Visualization of Complex Data Sets in Marine Microbiology,” Hochschule Bremen, 2013.

[8] P. L. Buttigieg and A. Ramette, “A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses.,” *FEMS Microbiol. Ecol.*, vol. 90, no. 3, pp. 543–50, Dec. 2014.

[9] R. T. Fielding, “Architectural Styles and the Design of Network-based Software Architectures,” University of California, 2000.