Marine

Microbial Biodiversity,

Bioinformatics & Biotechnology

Grant agreement n°287589

Acronym: Micro B3

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

Deliverable 5.12

Community Annotation Module Documentation

Version: 1.0

Circulated to: Renzo Kottmann, Mesude Bicak and OSD Core Team

Approved by: Prof. Frank Oliver Glöckner, 05/01/2016

Expected Submission Date: 30/06/2015

Actual Submission Date: 08/01/2016

Lead Party for Deliverable: CNRS
Mail: pascal.hingamp@gmail.com Tel.: +33(0)491 82 54 35

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

Originally the community annotation task 5.5 was focussed on single gene annotation of metagenomic data (See D5.1 and D5.2). However, OSD community demanded supporting and organizing a much wider spectrum of analyses approaches on the OSD metagenomes and amplicon based data sets. This led to the foundation of the "OSD Analysis Consortium" with more than 130 members. The 3rd Micro B3 training workshop was instrumental to kick-off the open collaborating on the analysis of the OSD 2014 data. This report describes the organization of the OSD Analysis consortium and collaboration tools, the OSD data, and gives an overview on the various analyses.

Table of Contents

1. Introduction

Building on the community spirit of Ocean Sampling Day mega-sequencing campaigns, the OSD Analysis Consortium consists of 130+ experts, with an aim to collectively and interactively analyse OSD 2014 data.

2. OSD Analysis Consortium

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 pipeline devised by the Micro B3 Information System[1] as well as the submission of all OSD raw sequences and metadata to relevant databases and their public distribution. In early 2015, at the 3rd Micro B3 training workshop at the EBI, an OSD Jamboree was held where 30 OSD participants came together to work on OSD 2014 data together with experts from OACG. During this week long Jamboree, we had the opportunity to observe the community spirit further with the extraordinary enthusiasm and interest of OSD participants in not only participating at the OSD campaign by collecting samples but also in collective analysis of OSD 2014 data. With 30 experts in the room, there was so much expertise, varied perspectives and different ideas of looking at the OSD data.

Consequently, OSD Core Team decided to take yet another initiative and transform the OACG into an OSD Analysis Consortium for open, collaborative analysis of OSD 2014 data. As a first step, all OSD Jamboree participants were invited to come on board, which was enthusiastically accepted by all participants at the EBI. OSD Core Team then circulated an open invitation to the OSD community via the osd-all mailing list to join this effort, inviting all interested individual or groups to send a short proposal outlining their analysis ideas.

45 proposals were received. Their titles are listed below.

	OSD Proposal Title
1	Global biogeography and endemicity
2	Biogeography of marine viruses
3	EBI Metagenomic analysis using UniPept
4	Photosynthetic phytoplankton biogeography
5	N2-fixing Cyanobacteria-Eukaryotes symbioses
6	Difference between Micro-Hitchhikers and Micro-Colonizers
7	Abundance and diversity of photoheterorophic bacteria
8	Abundance and distribution of heterotrophic bacteria and protist along OSD
9	Patterns of community structure in marine-coastal microbes: a global snapshot
10	Genetic diversity of functional genes involved in light acquisition and nutrient metabolism in marine heterotrophic bacteria
11	Controls on the Distribution of Polyphosphate Metabolism Genes
12	Comparing Microbial diversity in African Mediterranean and Atlantic marine ecosystems

[1] See Deliverable 5.8, 5.88, and 5.13

13	Contig binning from OSD Metagenomes
14	Picocyanobacteria distribution patterns
15	Anthropogenic-induced alteration of microbial plankton communities in the coastal ocean: a global, simultaneous study based on amplicon and metagenomic 16S rRNA gene sequences
16	Identification and Analyses of over-represented IPR codes
17	A Unipept based 16S-rRNA independent strategy to analyze the Oceans biodiversity
18	Horizontal gene transfer potential & bacterial immunity potential
19	Biogeography and environmental controls of the oceanic nitrogen cycle
20	Global biodiversity (taxonomic and functional) patterns of oceanic microbial communities: challenging the latitudinal gradient hypothesis
21	Harnessing the metabolic potential encrypted in marine microbial dark matter
22	Metabolic pathway (KEGG) profiling of all OSD samples
23	Diversity & relative abundance of eukaryotic ISIP genes
24	16S rRNA taxonomic analysis
25	Natural bacterial communities as bioindicators of chronic oil pollution in coastal areas
26	Negative counterfactuals and value of OSD standardisation effort
27	Suggestion of different topics and help in analysis, filling gaps
28	Comparative Entropy-based 16S and 18S Analysis of Global Richness
29	Cold adaptation in marine microbes
30	Gene function mining in OSD metagenomic data: an integrative approach for discoveries
31	Bioprospection of heavy metal resistance and tolerance genes in the marine microbiome for bioremediation purposes
32	OSD gene functions provide insights into novel global biogeochemical cycle connectivity
33	Distribution of iron uptake and iron metabolism systems in the OSD dataset
34	Distribution and environmental controls on Nitrogen biogeochemical functions
35	Comparative Metagenomics to Indicate Sites Under Anthropogenic Pressure: BTEX Example
36	C1 cycling
37	Screening OSD data for putative pathogen, fecal indicator, virulence, & antibiotic-resistance markers
38	Comparison of Taxonomic Diversity and putative pathogen sequences at Coral-Associated OSD Sites
39	Track novelty in eukariotic biodiversity
40	Worldwide distribution of Harmful Microalgae across all OSD samples
41	Surveying Multicellular Animals using Marine Environmental DNA
42	Metagenomic 18S rRNA sequences as a tool for assessing changes in phytoplankton assemblage structure driven by human pressures in the Mediterranean Sea
43	Species interactions across the world’s oceans
44	Multidimensional comparison of marine metagenomics and Setting up initial oceans microbial health index
45	Characterization of the eukaryotic microbiome by 18SrRNA metabarcoding data analysis and assessment of the relative resolution of V4 and V9 regions

As detailed in D2.10 an OSD Paper Taskforce was put together which reviewed all the proposals and grouped them into three main categories, assigning a task leader for each category:

1. **Diversity**
   (Using OTU-based metrics and alternatives such as MED, UniPept etc.)

   Task Leader: Linda Amaral-Zettler

2. **Insights metabolic functions (with focus on human impact) and their role in the ecosystems from Metagenomes**

   Task Leaders: Francesca Malfatti & Chris Bowler

3. **Towards an understanding of broad-scale ecological patterns**

   Task Leaders: Daniele Iudicone & Francesca Malfatti

Shortly after the assessment of all proposals, an abstract and a paper outline were put together by the OSD Paper Taskforce under the leadership of Dr Francesca Malfatti. This was circulated to OSD Analysis Consortium and received positive feedback with only minor comments. This is now a working Google Doc and can be accessed via: https://docs.google.com/document/d/1sW1lkn5A1iyAdM2sOw31kxrcLADFhkwRf1NIMp3GDKI/edit?usp=drive_web

OSD Analysis Consortium now consists of 130+ experts, led by Dr Mesude Bicak and Dr Francesca Malfatti (National Institute of Oceanography and Experimental Geophysics, Italy).

3. Collaboration tools

The main tool for communication is the OSD analysis mailing list (osd-analysis@microb3.org). However, for documentation, source code sharing and file sharing the consortium needed ways to exchange these information 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".

3.1 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 listed in Table 1.

Table 1: Overview of main wiki pages for documentation and discussion

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

3.2 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.

4. OSD 2014 data

OSD 2014 data can be grouped into three categories: 1. sequence data, 2. environmental data, and 3. ancillary data as detailed below.

4.1 Sequence data

OSD Bremen Team was in charge of releasing all OSD 2014 sequences. This involved pre-processing, quality checking and generation of "raw" and "workable datasets" which were versions of 16S, 18S and metagenomic sequences after having gone through processing according to agreed-upon quality standards. They were then submitted to European Nucleotide Archive (ENA) as agreed within Micro B3. All submission links, as well as relevant detailed documentation are provided via the "Overview of OSD 2014 data analysis" page on GitHub https://github.com/MicroB3-IS/osd-analysis/wiki/Guide-to-OSD-2014-data.

4.2 Environmental data

This is the metadata provided by OSD Site Coordinators along with their collected samples. First version of OSD 2014 environmental data was released by OSD Bremen Team in time for the OSD Jamboree at the EBI. This then went through several iterations as OSD Analysis Consortium members identified discrepant and wrong information during their analyses. OSD site coordinators were contacted individually and asked to check and provide correct information accordingly. Final version of OSD 2014 environmental is publicly available via GitHub along with detailed documentation.

4.3 Ancillary data

OSD 2014 ancillary data are extrapolated values based on the latitude and longitude of OSD 2014 sampling stations, from relevant public environmental databases. This was performed by Dr Shruti Malviya with support from Dr Daniele Iudicone, Dr Francesca Malfatti, Professor Chris Bowler and Dr Mesude Bicak.

Since the majority of the OSD Site Coordinators were unable to provide us with an extensive list of environmental metadata, OSD Analysis Consortium collectively decided to extend the set of "overall environmental data", in order to have the ability to perform in-depth analysis. An "Ancillary Data Request Form" was put together and circulated among OSD Analysis Consortium to allow participants to facilitate their requests for additional data https://github.com/MicroB3-IS/osd-analysis/wiki/Requests-for-ancillary-data.

Ancillary data was retrieved from multiple datasets, including Ben Halpern's dataset [1], which is the most relevant dataset given the nature of the OSD Sites being "coastal" with around 75% of sampling sites being within 10km from the coast. This effort increased the value of OSD 2014 dataset tremendously and has been highly appreciated by OSD Analysis Consortium participants. Final version of OSD 2014 ancillary data is publicly available via GitHub.

5. Latest Progress and Future Plans

OSD Analysis Consortium participants have been provided with two deadlines. One deadline at the end of July 2015 before the summer break to send short Interim reports, in order to enable OSD Paper Taskforce to further check on their progress and needs. 15 interim reports were received. Rest of the proposal PIs were contacted individually for a short update by email.

Second deadline was on December 4th 2015 to send final reports on their analyses results and discussion. 17 final reports were received. 10 proposal PIs informed they are running behind the deadline and will provide their reports after Christmas break.

Dr Malfatti and Dr Bicak are currently in progress of reviewing all received final reports, contacting proposal PIs for further information or suggestions to aid with their analyses accordingly, while collating methods, results and findings with an aim to prepare a manuscript. Dr Malfatti and Bicak will circulate a draft among OSD Paper Taskforce after Christmas break for their comments and feedback.

Despite Micro B3 officially ends in December 2015, the OSD Analysis Core Group will continue working beyond that at least until the first OSD analysis paper is published. A first draft of the paper is targeted for circulation among the OSD Analysis Consortium by late January 2016.

6. References

[1] Halpern BS, Frazier M, Potapenko J, Casey K, Koenig K, Longo C, Lowndes JS, Rockwood RC, Selig ER, Selkoe KA, Walbridge S. 2015. Spatial and temporal changes in cumulative human impacts on the world’s ocean. Nature Communications. 6:7615. DOI: 10.1038/ncomms8615