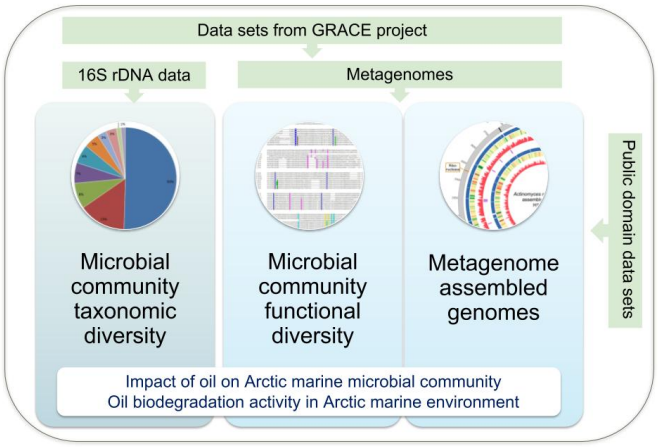
GRACE grant no 679266

# Report on results of omics data meta-analysis

D2.5

WP2: Oil biodegradation and bioremediation



Prepared under contract from the European Commission
Contract n° 679266
Research and Innovation Action
Innovation and Networks Executive Agency
Horizon 2020 BG-2014-2015/BG2015-2

Project acronym: GRACE
Project full title: Integrated oil spill response actions and environmental effects
Start of the project: 01 March 2016
Duration: 42 months
Project coordinator: Finnish Environment Institute (SYKE)
Project website http://www.grace-oil-project.eu

Deliverable title: Report on results of omics data meta-analysis

Deliverable n°: D2.5
Nature of the deliverable: Report
Dissemination level: Public

WP responsible: WP2
Lead beneficiary: UTartu

Due date of deliverable: 31.12.2018
Actual submission date: 7.1.2019

Deliverable status:

| Version | Status | Date | Author | Approved by |
|---|---|---|---|---|
| 1.1 | draft | 21.12.2018 | Jaak Truu, Kristjan Oopkaup, Marika Truu | WP members 2.1.2019 |
| 1.2 | final | 7.1.2019 | Jaak Truu, Kristjan Oopkaup, Marika Truu | Steering group 4.1.2019 |

# Table of Content

## Executive Summary

Advanced DNA sequencing technologies are coupled with the computational challenges to deliver the most relevant biological interpretation of the data collected during the GRACE project. A considerable number of computational tools have been developed to make the most out of the multi-layer omics data sets in microbiology. In order to maximize output from microbial ‘omics’ data sets obtained during GRACE project the integrative knowledge discovery from multiple omics sources is applied. This analysis is based on the integration of high-throughput sequencing datasets (amplicon based and shotgun metagenomes) obtained during GRACE project and relevant public domain data. To analyse and integrate these large 'omics' data sets novel bioinformatics approaches including metagenomic binning for recovery of population genomes of oil-degrading microbial taxa and machine learning methods are applied. Obtained ’omics’ data sets are compared to those found previously in the marine hydrocarbon plumes in order to look for similarities and differences in microbial community composition and metabolic processes in oil-contaminated seawater and sediments due to environmental constraints, remediation strategies and ecosystem type. Finally, based on performed data analysis results the metagenomic prediction platform for inferring oil biodegradation activity under different oil pollution scenarios in Arctic marine environment will be established.

# 1. Introduction to omics data integration

The term *data integration* refers to the situation where, for a given system, multiple sources (and possible types) of data are available and we want to study them integratively to improve knowledge discovery. In order to draw a more comprehensive view of biogeochemical processes performed by marine microbial communities with regard to oil pollution, experimental data made on different layers have to be integrated and analyzed. We define omics *data integration* as the use of multiple sources of information derived from different omics technologies to provide a better understanding of marine microbial community taxonomic and functional structure and its association with oil exposure and oil biodegradation activity. The integration of heterogeneous and large omics data sets creates not only a conceptual challenge but a practical hurdle in the regular analysis of omics data. Integrating several data types — marker gene sequencing and shotgun metagenomics data (and if available metatranscriptomics, metaproteomics, metabolomics data) — for a given study is crucial for a comprehensive understanding of the composition and function of marine microbial communities. In the case of the microbiological omics data sets, there is still no unified definition of omics data integration, nor taxonomy for data-integration methodologies despite some recent efforts on this topic[1,2]. The complexity of marine microbial communities, the technological limits, a large number of biological variables (ie microbial taxa, genes, metabolic pathways) and the relatively low number of biological samples make integrative analyses a challenging issue.

# 2. Resources for omics data integration

## 2.1. Data source discovery

*Data source discovery* is defined as the identification of relevant data sources. In the case of GRACE project, the following omics data are produced – amplicon based sequencing of all samples from lab and field scale experiments, and shotgun DNA sequencing of the subset of these samples. Publicly available data resources consist of 16S rDNA and metagenome data sets deposited to ENA (European Nucleotide Archive) and GenBank, as well as data sets deposited to MG-RAST (Metagenomic Analysis Server) and IMG/M (Integrated Microbial Genomes system).

## 2.2. Analysis methods for omics data integration

### 2.2.1. Data integration of similar and heterogeneous omics data types

The use of specific 16S rRNA gene amplicons as compared to shotgun metagenomics to assess the microbial community structure helps the economy of analysis both in getting better coverage for given sequencing depth and by reduced computational complexity. For many possible 16S rRNA gene amplicon regions there are databases assisting taxonomic and phylogenic classifications of sequences and OTUs produced. However, pooling and analysis of 16S rRNA gene amplicon data sets is a challenge when different 16S rRNA gene regions are amplified in different studies. Among possible data analysis approaches (OTU) clustering independent microbial community analysis could be utilized

By comparison (to 16S rRNA gene amplicon based strategies) in shotgun metagenomic analyses, all DNA in a sample is sequenced and analysed, which while being a more comprehensive approach means that the same sequencing depth yields far lower coverage in such studies. Metagenomes obtained in different studies can be jointly analysed more easily than 16S rRNA gene amplicon based data sets. Another method for analysing metagenome sequencing reads is to assemble the short reads into longer sequences (contigs). These contigs can be further sorted or binned by similarity to assemble partial to full genomes of microorganisms. This allows data exploration beyond taxa and gene annotation, enabling the prediction of multi-gene biosynthetic pathways or even metabolic reconstructions. To assemble partial to full genomes of individual

microorganisms, contigs are sorted (binned) into separate putative genomes with tools such as MaxBin2[3] and CONCOCT[4]. The data processing can be performed employing integrated workflow tools, such as Anvi'o[5] and ATLAS[6], to automate different analysis tasks.
Data integration of heterogeneous omics data types (ie 16S rRNA gene amplicons and metagenomes) is currently an active field of research where biostatisticians are constantly proposing hybrid approaches to improve data utilization and scientific discovery.

### 2.2.2. Software for integrated omics data analysis

#### 2.2.2.1. Web-based solutions and resources

MG-RAST[7] pipeline allows analyzing large shotgun metagenomic data sets as well as amplicon based data sets. In addition, the system maintains a large set of public metagenomes that could be integrated into the analysis. Another similar system, MGnify[8] (former name EBI metagenomics http://www.ebi.ac.uk/metagenomics) provides a free to use platform for the analysis and archiving of sequence data. Marine Metagenomics Portal contains the marine databases; *MarRef, MarDB and MarCat* as well as a pipeline for annotation and analysis of marine metagenomics samples, which provides insight into phylogenetic diversity, metabolic and functional potential of environmental communities[9]. *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 *MarCat* database represents a gene (protein) catalog of uncultivable (and cultivable) marine genes and proteins derived from marine metagenomics samples.

#### 2.2.2.2. Stand-alone software for integrated omics data analysis

**Multivariate exploratory data analysis methods**. mixOmics, an R package dedicated to the multivariate analysis of biological data sets with a specific focus on data exploration, dimension reduction and visualisation. By adopting a systems biology approach, the toolkit provides a wide range of methods that statistically integrate several data sets at once to probe relationships between heterogeneous 'omics' data sets[10].
**Machine learning**, a collection of data-analytical techniques aimed at building predictive models from multi-dimensional datasets, is becoming integral to modern biological research. This method has been used for predictions of metabolic functions in complex microbial communities[11]. The utilization of more sophisticated machine-learning method (deep learning - a neural network that includes multiple hidden layers) is that training a deep neural network requires massive datasets of size often not be attainable in many biological studies[12].

## 3. Omics data integration approach in the GRACE project

In order to maximize output from ‘omics’ data sets obtained during GRACE project the integrative knowledge discovery from multiple omics sources is applied. To analyse and integrate these large 'omics' data sets novel bioinformatics approaches including metagenomic binning for recovery of population genomes of oil-degrading microbial taxa are used. Obtained ’omics’ data sets are compared to those found previously in the marine hydrocarbon plumes in order to look for similarities and differences in microbial community composition and metabolic processes in oil-contaminated seawater and sediments due to environmental constraints, remediation strategies and ecosystem type. Finally, based on performed data analysis results the metagenomic prediction platform for inferring oil biodegradation activity under different oil pollution scenarios will be established.

The analysis is based on the integration of high-throughput sequencing datasets (amplicon based and shotgun metagenomes) obtained during GRACE project and relevant public domain data. During data analysis following sub-tasks will be performed:

1. First, the sample sequencing coverage will be estimated for GRACE and public domain data sets. As an example, sequencing coverage estimates are shown for amplicon based and shotgun metagenome data sets obtained during WP2 lab-scale experiments (Fig. 1). Then these data sets are taxonomically and functionally annotated, and metabolic pathways related to oil biodegradation are described. The obtained sequencing data sets are analyzed by read-based profiling using state-of-the-art tools, such as Kaiju[13], MEGAN[14], and FOAM[15], or by assembly-based analyses, with tools such as metaSPAdes[16], MEGAHIT[17] and MEGAN-LR[18].

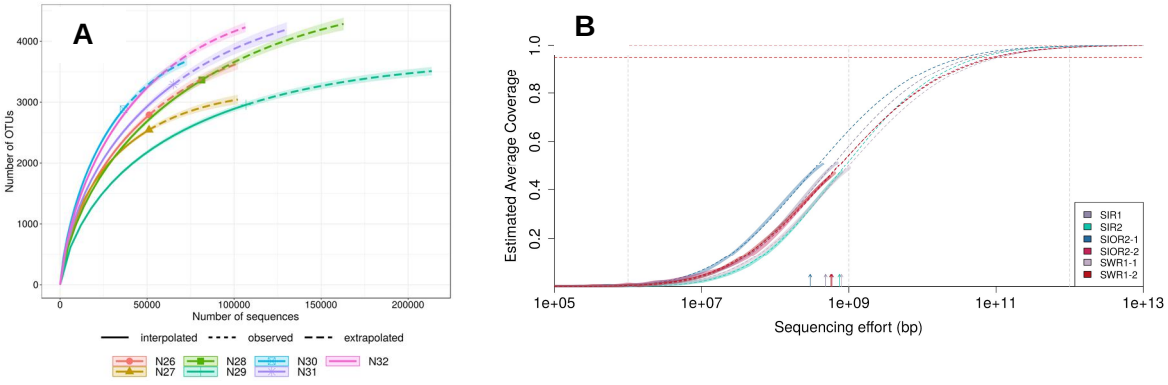



**Figure 1.** Sequencing coverage estimation for samples obtained during oil in ice/seawater mesocosm experiments. A. Bacterial community richness estimation using iNEXT software (http://chao.stat.nthu.edu.tw/wordpress/software_download/inext-online/). Estimated coverage values are in the range 0.97 - 0.99. B. Meganome sequencing coverage estimation using Nonpareil software[19]. Estimated coverage values are in the range 0.45 - 0.51. *Sample code abbreviations: Plot A – samples from 8 month long mesocosm experiment: N26, N27, N28 – seawater control; N29, N30 – seawater with oil; N31, N32 - seawater with oil with NPK amendment; Plot B: SWR1-1 and SWR1-2 – collected seawater, SIR1 and SIR2 sea-ice control in mesocosms, SIOR1-1 and SIOR1-2 – sea-ice with oil in mesocosms.*

2. Recovery of individual genomes (metagenome assembled genomes - MAGs) from obtained metagenomics datasets using binning of assembled contigs to species-level groups both from single metagenomes and related multiple metagenomes from GRACE project experiments. This approach enables to better understand the role of uncultivated microbial species in oil biodegradation in Arctic marine environment.

3. Information about microbial community taxonomic composition and metabolic markers together with abiotic factors will be related to oil biodegradation kinetic parameters and oil remediation strategies using different modelling approaches (structural equation modelling, network analysis,

random forest analysis). The general approach for omics data integration in the GRACE project is outlined in Figure 2.

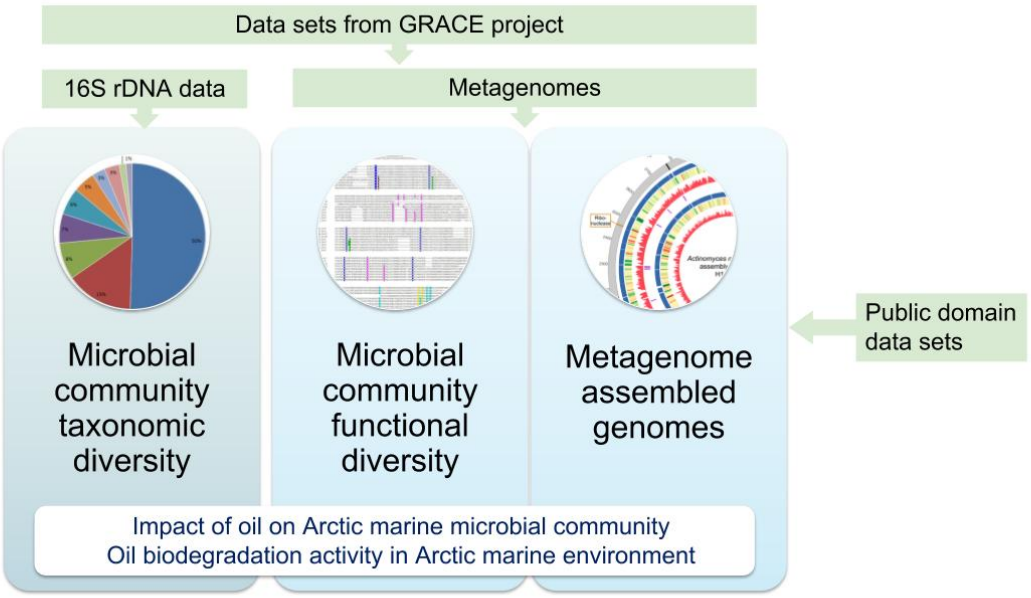



**Figure 2.** Overview of the omics data integration approach in the GRACE project.

## References:


1. Gomez-Cabrero, D. *et al.* Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* **8,** I1 (2014).
2. Huang, S., Chaudhary, K. & Garmire, L. X. More is better: Recent progress in multi-omics data integration methods. *Front. Genet.* **8,** 84 (2017).
3. Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2,** 26 (2014).
4. Alneberg, J. *et al.* CONCOCT: Clustering cONtigs on COverage and ComposiTion. 28 (2013). at <http://arxiv.org/abs/1312.4038>
5. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3,** e1319 (2015).
6. III, R. A. W. *et al.* ATLAS (Automatic Tool for Local Assembly Structures) - a comprehensive infrastructure for assembly, annotation, and genomic binning of metagenomic and metatranscriptomic data. (2017). doi:10.7287/peerj.preprints.2843v1
7. Keegan, K. P., Glass, E. M. & Meyer, F. in *Methods in molecular biology (Clifton, N.J.)* **1399,** 207–233 (2016).
8. Hunter, S. *et al.* EBI metagenomics--a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.* gkt961- (2013). doi:10.1093/nar/gkt961
9. Klemetsen, T. *et al.* The MAR databases: development and implementation of databases specific for marine metagenomics. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx1036
10. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Comput. Biol.* **13,** e1005752 (2017).
11. Langille, M. G. I. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31,** 814–821 (2013).
12. Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. & Collins, J. J. Next-generation machine learning for biological networks. *Cell* **173,** 1581–1592 (2018).
13. Menzel, P. *et al.* Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **7,** 11257 (2016).
14. Huson, D. H. *et al.* MEGAN community edition - Interactive exploration and analysis of large-scale

microbiome sequencing data. *PLOS Comput. Biol.* **12,** e1004957 (2016).
15. Prestat, E. *et al.* FOAM (Functional Ontology Assignments for Metagenomes): a Hidden Markov Model (HMM) database with environmental focus. *Nucleic Acids Res.* gku702- (2014). doi:10.1093/nar/gku702
16. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. metaSPAdes: a new versatile de novo metagenomics assembler. (2016). at <http://arxiv.org/abs/1604.03071>
17. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. 2 (2014). at <http://arxiv.org/abs/1409.7208>
18. Huson, D. *et al.* MEGAN-LR: New algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *bioRxiv* 224535 (2017). doi:10.1101/224535
19. Rodriguez-R, L. M., Gunturu, S., Tiedje, J. M., Cole, J. R. & Konstantinidis, K. T. Nonpareil 3: Fast estimation of metagenomic coverage and sequence diversity. *mSystems* **3,** (2018).