Convocatoria 2019 «Proyectos de I+D+i»

AVISO IMPORTANTE - La memoria no podrá exceder de 35 páginas. Para rellenar correctamente esta memoria, lea detenidamente las instrucciones disponibles en la web de la convocatoria. Es obligatorio rellenarla en inglés si se solicita más de 100.000,00 €.

IMPORTANT – The research proposal cannot exceed 35 pages. Instructions to fill this document are available in the website. If the project cost exceeds 100.000,00 €, this document must be filled in English.

TÍTULO DEL PROYECTO COORDINADO (ACRÓNIMO): Un enfoque de estudio de la ecología microbiana basado en rasgos y gremios (TRAITS)

***TITLE OF THE COORDINATED PROJECT (ACRONYM):* An approach to study the ecology of marine microbial communities based on traits and guilds (TRAITS)**

DATOS DE LOS SUBPROYECTOS - *SUBPROJECTS DATA*

SUBPROYECTO 1:

IP 1 COORDINADOR 1 (Nombre y apellidos)**:** Javier Tamames de la Huerta

IP 2 COORDINADOR 2 (Nombre y apellidos)**:** Carlos Pedrós-Alió

TÍTULO - *TITLE*: Studying the structure of guilds beyond the species rank. Development of methods for the fast analysis of microbiomes.

SUBPROYECTO 2:

IP 1 (Nombre y apellidos)**:**José Manuel González Hernández

TÍTULO - *TITLE*: Shedding light on marine bacterioplankton and their energy sources from an evolutionary standpoint

SUBPROYECTO 3:

IP 1 (Nombre y apellidos)**:** Sandra Martínez García

TÍTULO - *TITLE*: Towards the identification of functional groups of marine prokaryotes in carbon and nutrient utilization

1. JUSTIFICACIÓN DE LA COORDINACIÓN

Two of the PIs (CP-A and JG) obtained three consecutive Plan Nacional grants between 2005 and 2013. The common theme was to use the power of high throughput sequencing to study the ecology of marine microorganisms and, in particular, of marine Bacteroidetes, used as a model group. As more and more sequences became available, one of the weak points of our team was bioinformatics. So, for the 2013 call, JT (a bioinformatician) joined the group and we have obtained two consecutive Plan Nacional grants. The current one will finish on December 31, 2019.

In 2016 CP-A moved from ICM in Barcelona to CNB in Madrid, and teamed up with JT to create a research group named “Microbiome analysis laboratory” (http://microbiomecnb.com/). Therefore, we organized the current project NOVAMAR in two subprojects, one with JT and CP-A as coPIs at CNB-CSIC (Madrid) and another one with JG as PI at Universidad de La Laguna (Tenerife).

In these two recent projects a weakness of our team was the lack of experimental work. None of the two groups has a wet lab. In the still active project NOVAMAR we solved this issue by collaborating with experimental teams lead by Drs. Fernando Baquero and Teresa Coque at Hospital Ramón y Cajal (Madrid) to analyze human gut microbiome and by Dr. Josefa Antón at Universidad de Alicante to work on solar salterns, an example of extreme environments. Our project provided most of the consumables, sequencing costs, and workforce for such collaborations, while our colleagues provided the lab infrastructure and their knowhow of the respective environments. In the present project we felt the need to incorporate a team that could sample marine environments and carry out mesocosm

experiments with marine microbial communities. Thus, we have incorporated Dr. Sandra Martínez García (SMG) from the Universidad de Vigo to the team.

Since this project builds upon the research of three of the PIs carried out together in the previous and current projects, the need for coordination seems obvious. The organization into three coordinated subprojects is necessary for a more efficient administrative functioning. For example, if we had only one project, computers could not be bought for different institutions (CSIC, Universidad de Vigo, and Universidad de La Laguna), because they would have to be inventoried at the leading institution. This is just an example of the problems that having a single project would create.

The four PIs have complementary backgrounds. CPA is a microbial ecologist with a special interest in diversity. JT is a bioinformatician with interest in metagenomics, JMG is a microbiologist with an interest in genome structure and regulation of environmental bacteria, and SMG is a marine microbiologist with experience in carbon and nutrient fluxes. The added value of joining efforts seems obvious. The three sub-projects have interactions that will require closely working together. We will have Skype meetings frequently, plus at least two meetings in person per year. If we get grants for PhD students, they will spend some time in the other labs to get familiarized with the techniques and have a better scientific education as well as higher profile results.

2. PROPUESTA CIENTÍFICA

C.2.1. General objective, hypothesis and fit within "Estrategia Española de Ciencia y Tecnología y de Innovación and the call "retos investigación".

The final objective of this project is to analyze how the microbial communities are structured by guilds, groups of taxa using the same resources in a similar way. This will require determination of the taxonomic composition of guilds and of their changes in space and time.

Our objective fully agrees with the Societal Challenges of the H2020 program. Our work will be relevant for the contribution of the oceans to global change and for a better planning of the exploitation of marine microbial genetic resources. The project fully agrees with the challenge ("reto") "**Bioeconomía: sostenibilidad de los sistemas de producción primaria y forestales, calidad y seguridad alimantaria, investigación marina y marítima**".

C.2.2. Overview of the project

a) Approach: theory

The common thread through all our recent projects is to use genes to understand the ecology of microorganisms, from genomes to biomes. Our previous project NOVAMAR focused on positive interactions among bacterial taxa that might form modules. These modules would include bacteria able to grow together performing a function that requires the individual members to interact and, presumably, complement each other with their respective genomic suites. In order to keep the problem within doable limits we did a lot of work on communities simpler than the ocean, as solar salterns as an extreme environment with only a few tens of taxa, and the human gut as a community of intermediate complexity with hundreds of taxa. Finally, we also studied marine communities with thousands of taxa.

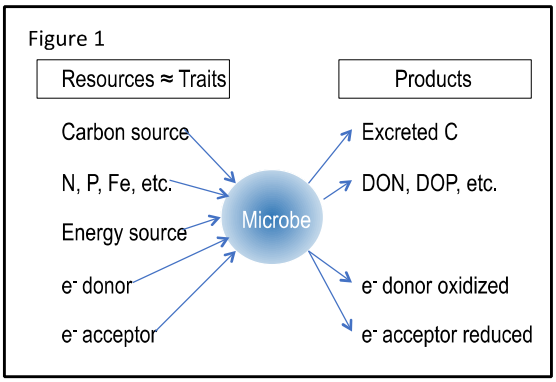
In project NOVAMAR we studied single genes to get insights into the dynamics of the microbiota. Thus, for example, we used the 16S rRNA gene to follow co-variations of bacterial taxa both in time (through Lotka-Volterra generalized models, Jiménez et al., in preparation) and in space (through co-occurrence networks, Puente-Sánchez et al., in preparation). But we also used whole genomes of isolates, single cell amplified genomes (SAGs), and metagenomic amplified genomes (MAGs) to interpret metagenomes and metatranscriptomes, particularly using the marine *Flavobacteriia* as a model group (Royo-Llonch et al., 2017; Cobo-Simón et al. 2017).

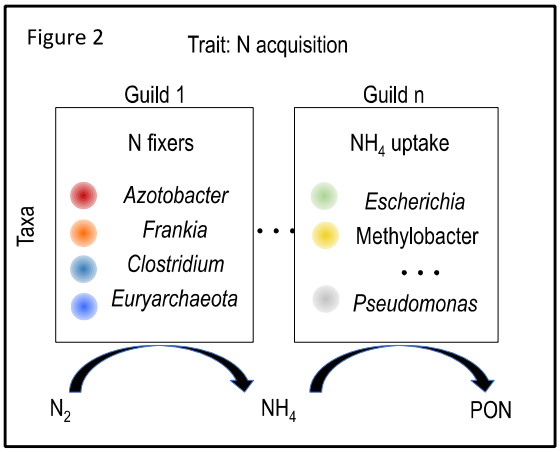
In the present project we will take a different approach. Instead of focusing on groups of taxa collaborating with each other, we will center on groups of taxa defined as **guilds** (see below), whose members presumably compete with each other. Both types of modules are essential for community building. In what follows we will define the concept of guild, we will explain how do we intend to determine guilds from genes, and we will try to clarify how guilds relate to community structure and to ecosystem functions. The whole issue is complex because in using genes to define guilds and guilds to determine community structure and ecological functions, different levels of organization are superimposed.

Root (1967) defined **guilds** as ‘‘a group of species that exploit the same class of environmental resources in a similar way. This term groups together species, without regard to taxonomic position, that overlap significantly in their niche requirements”. As many other ecological concepts, that of “guild” is useful but vague and, moreover, the definition needs to be fine-tuned in order to apply it to microorganisms. Since our project will deal only with microorganisms with a prokaryotic cell structure, from now on we will use “prokaryotes” to include both bacteria and archaea, despite the fact that these two taxa are widely distant phylogenetically.

The basic resources for a prokaryote are shown in Fig. 1. These include carbon and energy sources, inorganic nutrients, and electron donors and acceptors. As a result of the metabolism, these resources are transformed into cell mass, excreted carbon, nitrogen etc., oxidized electron donors, and reduced electron acceptors. There may be additional resources in some cases, for example a given vitamin for a cell that is not able to synthesize it. The ability to acquire each of such resources is a **trait** of the organism. For example, a trait is nitrogen acquisition, and each organism may fulfill this trait through a variety of ways, such as nitrogen fixation or ammonia uptake for example. In order to fulfill this trait, the cell needs certain genes. Since guilds are defined as a group of species exploiting resources in a similar way, the prokaryotes fulfilling the trait nitrogen acquisition through nitrogen fixation constitute a guild (Pedrós-Alió 1989).

In Fig. 2 we show two examples of guilds for the trait nitrogen acquisition. As can be seen, the trait allows for several different guilds. Guild 1 uses N_2 and Guild n uses ammonia. Members of Guild 1 use the same resource in the same way (as expected for the members of a guild) and the same applies to Guild n.





Figures 1-2: Description of traits and guilds

There are several characteristics of this organization that we need to mention. First, the number of taxa in a guild can be very large or very small. Second, the members of the guild do not have to be taxonomically related. For example, members of Guild 1 include members from both domains bacteria (in red tones in the figure) and archaea (in blue tones). And those of Guild n include members from

different bacterial Phyla (in different colors). Again, this was already explicit in Root's definition. Note that a given microorganism may belong to several guilds. For example, most nitrogen fixing prokaryotes can also use ammonia. This will presumably have consequences on the structure and internal competition within a guild that we intend to explore in WP 3.

The ecological literature proposes that a guild with many members will contribute to the stability of the ecosystem (Simberloff and Dayan 1991, Blondel 2003, Koch 2019). Members of a guild are supposed to interact mostly by competing with each other, since they use the same resources in a similar way. As environmental conditions shift, one or another member of the guild will become dominant. The function will be carried out anyway, even though the main players will change. In effect, this is what has been termed "functional redundancy". Unfortunately, this concept has been liberally used in the recent literature with very little precision. For example, it has been used at the level of clusters of orthologous groups (OGs). These groups include genes for very general items such as "amino acid transport and metabolism", "cell wall/membrane/envelope biogenesis" or even "general function prediction". Obviously such functions are going to be present always, since no living being can exist without them. Claiming functional redundancy at this level is a useless tautology. More specific functions have to be considered to analyze whether functional redundancy exist or not (Sunagawa et al. 2015). In the present case of guilds, functional redundancy concerns very specific metabolic reactions such as ammonia uptake or nitrogen fixation. At any rate, this shows that guild structure is an important aspect of communities.

In order to show the intricacies of guild structure, we will use the example in Figure 2. Nitrogen acquisition can be achieved by several different methods, each one constituting a different guild (guilds 1, … , n). Guild 1 obtains nitrogen by fixing atmospheric N_2. A variety of distantly related bacteria and archaea belong to this guild. Other guilds will use nitrate, dissolved organic nitrogen, amino acids or ammonia as resources. We need to state explicitly one final aspect of guild structure: a given taxon may belong to several guilds. For example, as already mentioned, most nitrogen fixers can also use ammonia as a source of nitrogen if present. This aspect was already reviewed by Simberloff and Dayan (1991), who recognized that a given bird species could be a foliage gleaner for the trait food, a tree hole nester for the trait breeding, and an altitude migrant for the trait migration.

There is evidence that in real microbial ecosystems, guilds always have more than one abundant species. A good example is that of purple sulfur bacteria in anaerobic lakes. These bacteria are an extremely clear-cut example of a guild, because they use light as a source of energy, H_2S as a source of electrons, and CO_2 as a source of carbon. Thus, they belong to the same guild in these three traits. Intense competition can be expected among them. Yet, usually there are at least two abundant species, a few low abundant species, and several rare species that may be isolated in pure culture but are hard to detect by molecular methods. Thus, in anoxic Lake Cisó (Girona, Spain) *Thiocystis* sp. and *Amoebobacter* sp. were co-dominant, but a small and a large *Chromatium* species could also be detected in low concentrations, and several other species had been isolated in pure culture such as *Thiocapsa roseopersicina* or *Amoebobacter purpureus* (Casamayor et al. 2000). The two main bacteria became the dominant species at different times of the year, and the function, anoxygenic photosynthesis, was carried out all the time.

The main objective of the present project is to determine the number of taxa and their relative abundances for a number of carefully chosen guilds under different circumstances. Guilds are considered bricks in the construction of communities and we would like to investigate their structure and changes with time. We want to determine the compositions of guilds under different circumstances.

b) Approach: Methodology

We will use genes from genomes and metagenomes to define guilds and explore their composition and variations in time and space. We illustrate the process in Figure 3. First (left hand panel) we will identify marker genes for each guild in each of the chosen traits. For example, *nifH* is a well know marker gene for nitrogen fixation. Next, we will scan the existing databases of genomes for the prokaryotes that have such marker genes. This way we will compile a table of taxa belonging to each guild in each trait.

Next (central panel) we will look for the identified marker genes in metagenomes. Our software SqueezeMeta obtains contigs with annotated genes. As a result we will have marker genes with an assigned taxonomy. With this information we can build a second table (right panel in Figure 3). This table is equivalent to the first one but with taxa and genes derived from the environment, not from genomes.

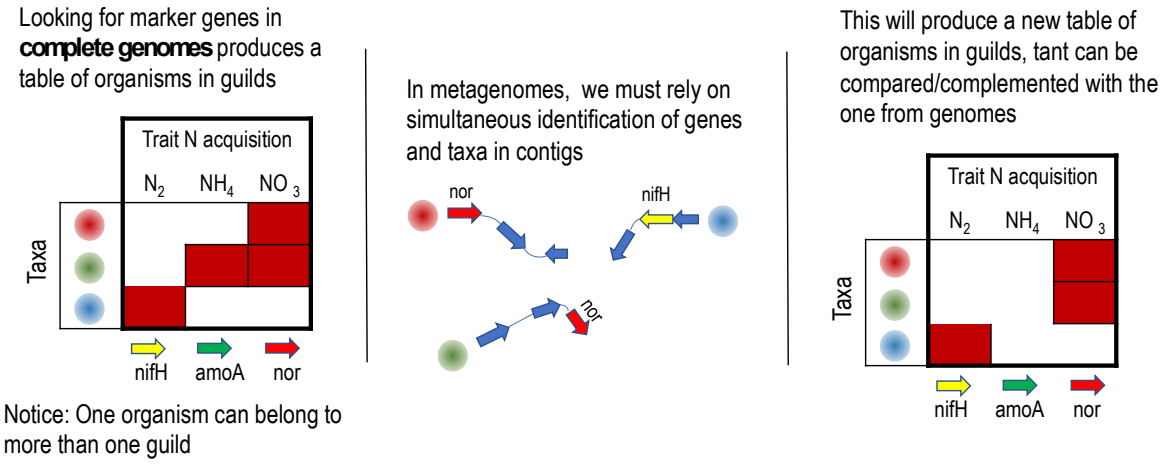


Figure 3: Creation of the tables of guilds in traits

We will then combine both tables for a more accurate picture of the taxa in each guild (Figure 4 left). This table will allow us to look at the composition of each guild in each sample (metagenome, Figure 4 right). And this information will be the material to test our hypotheses about guild composition and variability in space and time.

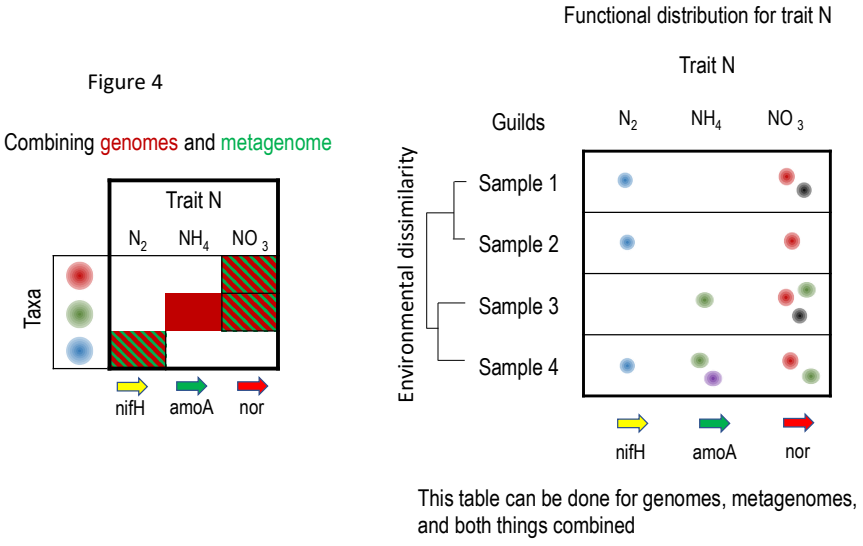


Figure 4: Tables connecting traits, guilds and taxa

c) Summary of WPs

WP1 will optimize a protocol to collect a sample, sequence the metagenome, and process the sequences to provide the results *in situ,* within a day, with a portable lab and computer. This application will be useful both for our project and for the scientific community in general and will be one of the most important deliverables of the current project. WP2 will examine the taxonomic stability of guild composition at the microdiversity level. This WP continues the work already started in our current project, where microdiversity has been analyzed with single nucleotide variants of the 16S rRNA (García-García et al. 2019). The relevant taxonomic level of the members of a guild may range from microbes in different phyla to strains within the same species. Therefore, we will dedicate a work package to clarify the importance of microdiversity in the constitution of such guilds. Finally, in WP3 will try to extract general principles from the structure and variability of the different guilds analyzed in WPs 5-8. The deliverable from this WP will be theoretical and we hope it will be our main contribution to the field of microbial ecology.

WP4 will be devoted to fine tune the methodology for identification of marker genes in metagenomes. This is essential for the remaining WPs. The next four WPs (5-8) will concentrate on the guilds assigned to the traits we have chosen for analysis. We will determine their composition and their changes with time and space in the samples mentioned below. These WPs will analyze guilds using TonB dependent transporters in their outer membrane for uptake of nutrients (WP5), guilds using light as a supplementary energy source (WP6), guilds processing organic carbon (WP7), and guilds for the uptake of inorganic nutrients (specifically N, P and Fe, WP8). The reasons for these choices will be explained under each WP.

d) Methodological developments

We will dedicate considerable effort to methodological developments.

First, we have a work package to bring bioinformatics analysis of metagenomes and metatranscriptomes to the field. This WP1 is based on a deliverable from our previous project: a pipeline that processes raw sequences from metagenomes and generates final tables with assembled, annotated, and taxonomically assigned contigs plus generation of bins. This tool has been named SqueezeMeta (Tamames & Puente-Sánchez 2019) and is freely available at the “resources” tab of our web site (http://microbiomecnb.com/). The development of the *in situ* approach using this pipeline will be very useful for our project but it will also be extremely useful for scientists carrying out sampling in remote places or far away from home labs, such as in oceanographic cruises. We will do a proof of concept using a short cruise and mesocosm experiments in the Ría de Vigo.

Second, in the past two projects, we have suffered the difficulties of using automatically annotated sequences. From a methodological point of view, sequencing is not a major problem anymore. The bottleneck is in annotating the sequences (Fierer et al. 2014). First, this is done by comparison to databases, but many environmental microbes are not in such databases or are poorly represented. As a consequence automatic annotation generates many mistakes. A very frequent mistake is assigning a function of a known ortholog to an unknown paralog. Thus, we will carry out a careful curation of the sequences that will be used for the rest of the project. We propose several strategies for this purpose. We have already used some of them in a careful annotation of the *dmdA* gene (González et al. 2019). This development will be done in WP4 and the deliverable used in WPs 5 to 8.

Finally, in WP2 one aspect will be the development of a strategy to determine the functional content of genomes of strains, not only 16S rDNA genes, as already mentioned above. This WP will also develop an approach to recover taxonomic marker genes (16S rDNA) from metagenomes, the so-called miTags. This approach is described in Box WP2-1.

e) Samples

The present project will use sequences from a variety of samples that are already available and additional ones that will require sequencing. We will concentrate mostly on marine microbial communities, although other systems may be used whenever there is an opportunity or a need.

Day night cycles: We have data from several diel cycles in Antarctica and in the Mediterranean (Projects Pegaso and MiniSUMMER, PI Dr. Rafael Simó, ICM) sequenced by J. M. González, PI of subproject 1.

Time series: We have two time series. One is a collection of samples taken in Cambridge Bay (Arctic Canada) from early March to July of 2014 by Dr. C.J. Mundy and his team from the University of Manitoba. Both metagenomes and metatranscriptomes have been sequenced by the PIs of Subproject 3.

The second time series corresponds to project REMEDIOS (PIs: Beatriz Mouriño and Enrique Nogueira). This project included a short and intense series of 14 days samples every 6 hours, and a long one-year series of weekly sampling at the Ría de Vigo. We will have access through Dr. Sandra Martínez, PI of subproject 3. There are sequenced samples for 16S/18S amplicons and metatranscriptomes. Metagenomes from both REMEDIOS time series will be sequenced in the context of the present project (total of 66 metagenomes)

A third time series corresponds to the project ENVISION (PIs: Eva Teira and Emilio Fernández). This project includes a seasonal sampling of 8 days duration (samples every 24h) in an offshore station in the coastal area of NW Iberian Peninsula. We will have access through Dr. Sandra Martínez, PI of subproject 3. There are sequenced samples for 16S/18S amplicons and metatranscriptomes. Metagenomes from ENVISION time series will be sequenced in the context of the present project (total of 48 metagenomes)

Spatial sampling: We will use publicly available metagenomes and metatranscriptomes from circumnavigation cruises *Tara* Oceans and Malaspina 2010.

Micro- and mesocosms: We will have access to experiments carried out at the U. of Vigo within projects NitrAddex (with nitrogen and phosphorus additions, PI: Sandra Martínez) and ENVISION (with vitamin B12 additions, PIs: Eva Teira and Emilio Fernández). These will be provided by Dr. Sandra Martínez, PI of Subproject 3. From project NitrAddex we will sequence 16S/18S amplicons (20 samples), metatranscriptomes (20 samples) and metagenomes (20 samples) in the context of the present project. From ENVISION experiments there are sequenced samples for 16S/18S amplicons and metatranscriptomes and we will need to sequence metagenomes (21 metagenomes).

Single amplified genomes (SAGs) and metagenome assembled genomes (MAGs): We have about 40 genomes of several SAGs of the *Flavobacteriia, Kordia* and *Polaribacter* obtained in the current project NOVAMAR. These were sequenced by our student Marta Royo for her PhD in the lab of Dr. Silvia G. Acinas (co-director with CPA of Marta's thesis). These sequences are available.

Pure cultures: We have transcriptomes from pure cultures of three different strains of *Polaribacter* grown under light and dark conditions. These experiments were carried out by our PhD student Marta Royo in the laboratory of Dr. Jarone Pinhassi (member of the research team of subproject 3). The sequences are available.

In total we will sequence 168 metagenomes, 33 metatrasncriptomes and 20 16S samples from previous projects.

SUBPROJECT 1.

PIs: Javier Tamames and Carlos Pedrós-Alió, Centro Nacional de Biotecnología, CSIC

In reality, WPs 2 and 3 in this Subproject will use information from Subprojects 2 and 3 and, logically, they should have been presented last. However, the formal requirement that Subproject 1 is the coordinating project has forced us to present subprojects in this way. This Subproject has three very different WPs and, therefore, each WP has its own introduction. WP1will develop a field application useful for all the other WPs. WP2 will explore whether an analysis of guild composition is possible at a more detailed taxonomic level. WP3 will try to find general patterns in the composition and distribution of guilds identified in WPs 5-8 (in Subprojects 2 and 3).

WP1. Development of a system for *in situ* metagenomics analysis for monitoring of marine microbial communities. Responsible PI: Javier Tamames. Also involved: All the team.

State of the art

Metagenomics and metatranscriptomics have emerged as the core tools for analysing microbiomes, because of their potential to obtain a very large amount of information about the composition and functioning of microbial communities. Metagenomics consists in the isolation and sequencing of the environmental DNA contained in a particular sample. As these DNA molecules come from the diverse set of organisms present there, the analysis informs about the characteristics of all of them, allows understanding of the functional capabilities of the original organisms, and provides insights into the functioning of the ecosystem as a whole.

Metagenomics is the central tool for this project. It will allow to determine simultaneously taxonomy and function for the organisms present in the microbiota, and consequently to classify the organisms in guilds for the traits under study. Also, it will be instrumental to determine the patterns of microdiversity in microbial communities that can influence their stability and functioning. Therefore, it would be very important for this and upcoming projects to be able to do metagenomic experiments in the easiest and fastest way possible. This WP proposes a innovative way of doing these complex experiments with minimum equipment and resources.

A metagenomics experiment can be divided in four phases: Sample collection, DNA isolation, DNA sequencing, and sequence analysis (bioinformatics). These steps usually take place in different locations: the sample is collected in the field, it is processed in the wet laboratory (DNA isolation and sequencing), and the results are analysed in the bioinformatics laboratory. This experimental design complicates and delays the attainment of results. This is especially disadvantageous for particular applications where it would be critical to have results directly in the field, and/or as soon as possible. Some examples are the follow-up of the temporal dynamics of processes under exponential growth rates (viral or microbial epidemics) or rapid ecological changes (algal or cyanobacterial blooms). Also for remote, inaccessible locations where it would be important to know quickly if the target microorganism or functional activity is present, to proceed with further sampling.

In this WP we plan to create a portable in-situ metagenomics platform, aiming at doing all steps of the analysis in a single day (get samples in the morning, get results by night). This is a novel proposal for which very few proof-of-concept experiments have been developed. Besides the cited epidemic surveillance in the case of Ebola (Quick et al 2016) and Zika (Theze et al 2018), portable sequencing was also used in Arctic and Antarctic locations (Johnson et al 2017). Most of these studies make use of the minION nanopore sequencing technology, which is portable and can produce up to 30 Gb of DNA sequences in a few hours. But so far, none of these initiatives had the capability of completing the full experiment in a single day, especially because the lack of computational resources and bioinformatic integrated tools for metagenomic analysis. A critical limitation was the need for

powerful, high-throughput computing servers, that needed the transference of high amounts of data via internet to be analysed elsewhere, which can be unavailable at remote locations.

In the last months, we have developed SqueezeMeta, a complete pipeline for metagenomic/metatranscriptomics analysis (Tamames and Puente-Sánchez 2019). This platform runs all the steps of the bioinformatics analysis of metagenomes in a fully automatic way, not requiring human intervention besides the execution of the program. Computing resources are carefully tailored to be able to run in a portable computer, laptop type (for instance, maximum RAM usage is less than 16 Gb), still having a good performance and making it possible to finish the analysis of an average metagenome in just a few hours. Also, the software is totally independent of internet connectivity, because all necessary databases and tools are installed locally. This is, as far as we know, the only software capable to analyze metagenomic data without the need of big computational resources. This opens the way to the development of a fully portable in-situ metagenomic platform,

Therefore, we aim to develop and validate the "lab in a backpack" concept. This is a protocol that will integrate the DNA isolation, sequencing and bioinformatic analysis to be able to fully process a metagenome in less than one day and with few resources (limited power supply and no internet connection). This setup will include:

1) A portable laboratory such as the BentoLab (https://www.bento.bio) for DNA isolation. This device includes a centrifuge and a thermocycler than can be used as a heater for working with a commercial DNA isolation kit such as the DNA microbiome kit from Qiagen (https://www.qiagen.com).

2) A portable sequencing device such as the minION Nanopore from Oxford Nanopore Technologies (https://nanoporetech.com). This sequencer is very small in size, easy to operate, and can be connected directly to the computer, dumping the sequences into the computer´s disk as they are being produced. It weighs under 100 g and plugs into a PC or laptop using a high-speed USB 3.0 cable. Metagenomic libraries can be prepared using commercial kits or using the volTRAX device (https://nanoporetech.com/resource-centre/voltrax-0), an addition to the minION sequencer that allows the automatic preparation of libraries ready to sequence.

3) Our bioinformatics analysis platform SqueezeMeta, capable to analyse one or several metagenomes in a few hours. This system is able to run in a laptop, and it does not need internet connectivity.

4) Power supply can be produced by portable batteries that will be used to power up BentoLab and the laptop (the minION device is powered via connection to the laptop).

We will design the optimal setting of this system by analyzing synthetic microbiomes in the lab, mixing known amounts of known species to produce a mock community. We will replicate all the steps that will be done in the field, identifying the possible deficiencies and carefully monitoring power usage, computational resources, operation time and quality of the obtained results. Once the system is well set up under these controlled conditions, we will perform several experiments in real conditions, in freshwaters (lakes) and marine locations.

This platform will provide the capacity to perform metagenomic experiments at any moment in any place. This will boost the capacity to obtain information about the functioning of the microbial communities addressed in this project, and will be a valuable tool for subsequent ecological and clinical studies involving metagenomics.

Tasks and deliverables

Task 1: Setting up the individual components of the pipeline: BentoLab, DNA extraction kits, VolTRAX (or library preparation kits) and MinION. Deliverable: Complete platform for in-situ metagenomics

Task 2: Sequencing of synthetic microbiomes in controlled (laboratory) conditions. Deliverable: 1-day analysis of synthetic metagenomes

Task 3: Single (1 day) sequencing of an environmental (freshwater) metagenome. Deliverable: 1-day analysis of freshwater metagenome

Task 4. 2-days cruise sequencing and processing marine samples. Deliverable: 1-day analysis of marine metagenomes

WP2. Microdiversity within guilds. Responsible PI: Javier Tamames. Also involved: C. Pedrós-Alió.

In microbial ecology, species have been considered to be the units of diversity. Each species may be formed by several strains. The distribution and abundance of different strains in one or more samples is known as microdiversity, as compared with species diversity that does the same at species level. These strains from the same species share a core genome but also have different genes that make up the accessory genome. The differences in gene content can translate into different functional capabilities, and therefore two individuals from different strains of the same species may carry out slightly different functions in the environment. The differences may also be in adaptation to diverse environmental conditions, such as temperature. This fact challenges the usefulness of species as the most significant unit in microbial ecology.

Variations in the accessory genome are assumed to be responsible for niche differentiation, and consequently allow the discrimination of different ecotypes, pragmatically defined as populations of cells adapted to a given ecological niche. The existence of ecotypes with overlapping ecological functions (but adapted to different niches) has been proposed to confer stability to microbial ecosystems, guaranteeing the long-lasting persistence of bacterial populations (García-García et al. 2019). Indeed, several species have been shown to consist of many different subpopulations that sustain the distribution of the species across broad environmental gradients. It is expected that intra-species diversity also contributes to the preservation of the higher order inter-species interactions that are fundamental for community functioning and stability. Since intra-species differences can be large and produce different patterns of adaptation and functionality, it is sensible to think that, for understanding the dynamics of a microbial community, we should study the associated microdiversity.

The stability of microbial communities has been studied extensively, especially in relation with diversity. Diverse communities will contain functionally equivalent organisms able to respond differentially to the environment, resulting in increased stability against environmental perturbations (the so-called insurance hypothesis). As cited above, it is sensible to think that the most microdiverse species are shielded against environmental variation, as will be the interactions that depend on them. As microbial communities play a relevant role in ecosystem processes, understanding the drivers of microbiome stability is important to predict community response to future disturbances, such as those resulting from global change.

We will link the study of microdiversity with the functional guilds obtained in other WPs, to obtain information on the diversity of taxa linked to the different guilds. Taking into account the functional guilds determined in those WPs, we would study the microdiversity associated to the different functional traits/guilds. We will determine whether particular guilds are more diverse than others and try to link these patterns of diversity with the abundance or importance of these guilds for the ecosystem function. This will integrate trait-based ecology and the consideration of microdiversity in a completely novel way.

The analysis of the 16S rRNA gene has been widely used to define operational taxonomic units (OTUs). Commonly, OTUs have been defined at the 97% similarity level. Recent methodological

developments have allowed the discrimination of Amplicon Sequence Variants (ASVs), sequences that differ in just one nucleotide, overcoming the errors produced in high-throughput sequencing. These ASVs are useful proxies for strain existence, and analyzing them may explain the distribution of ecotypes across environments, unveiling previously overlooked ecological patterns.

Most diversity studies have been done sequencing an amplicon library of some variable regions of the 16S rRNA genes. Even if this molecule is a good marker gene and there is extensive information available for 16S sequences (for instance, we compiled all existing data and 16S sequences in the envDB database, Tamames et al. 2010), amplicon sequencing has the main drawback of producing known and unknown biases in the amplification and sequencing that can produce deviations in the inferred composition.

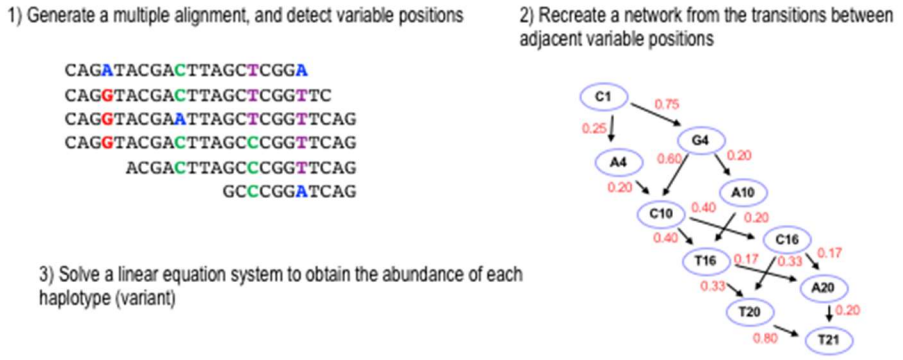
In contrast, metagenomics circumvents these drawbacks since it does not need the amplification step and the estimation of abundance is done via the determination of the coverage of each gene sequence (the number of metagenomic reads that map to that particular gene). Metagenonomics allows the retrieval of the so-called miTags, that is, metagenomic sequences mapping to the 16S rRNA gene, covering it partially. This is optimal because it provides an extensive and mostly unbiased set of 16S sequences that can be used for diversity and microdiversity estimation. But, since the metagenomic sequences are short and can be distributed along all the length of the 16S, only a small subset of these sequences can be used (those belonging to the variable zone used for the analysis). It would be ideal to have a method capable of reconstructing full 16S sequences from these short miTags, so that the whole length of the gene could be used to infer the microdiversity of a particular species.

Accordingly, we will develop a method for reconstructing full-length 16S sequences using paired end metagenomic reads belonging to these genes. A scheme can be seen in Box WP2-1. First, the 16S reads belonging to a particular species are selected., by means of the SqueezeMeta software. The reads for the selected species will be aligned to a set of reference 16S rRNA sequences of the same species, generating a multiple alignment of 16S rRNA reads. Then, the variable positions in the alignment are determined, and we will build a graph of transition between variable positions that can be solved to determine the frequency of all possible haplotypes

---

Box WP2-1: *Description of the procedure to reconstruct full 16S rRNA sequences from miTag sequences.*

Consider that position 4 varies between A and G, and position 10 between A and C. We count the number of reads having each combination (A-A, A-C, G-A and G-C), and get transition probabilities between the two variable positions (simply the percentage of reads of each combination). Once we do this for all pairs of transitions between variable positions, we can obtain the probability of emission of every single full-length variant of the 16S sequence. This is equivalent to building a Hidden Markov Model (HMM) using the variable positions. We can add additional constrains using the secondary structure of the 16S rRNA, in which several bases are paired and therefore the space of possible sequences is limited. The emission probabilities are directly translated into relative abundances of each variant, that are in fact ASVs because they can differ in just a single position.



Using this method, we will be able to reconstruct the ASV abundances for many species using metagenomic data, which will allow analysis of the microdiversity of data sets for which no amplicon data are available, such as the Cambridge Bay Arctic samples (time series) or the samples from the Malaspina expedition (spatial samples). This will be a great methodological advance allowing the inspection of microdiversity for a huge number of previously sequenced metagenomes. Also, it can provide additional, complementary, and possibly more accurate data to the amplicon data series for the REMEDIOS cruises.

The reconstructed 16S variants will be used to study the taxonomic microdiversity of the different samples. Specifically, we will address the following questions:

-Determination of the dominant ASVs for different taxonomic groups, focusing mainly on *Bacteroidetes*, SAR11 (*Candidatus* Pelagibacter), *Roseobacter* and *Alteromonas* which are the most abundant taxa in the oceans.

-Comparison of ASVs of the same species among samples of the same cruise, to determine temporal (in time series) or spatial (in Malaspina samples) dynamics. Linkage of the different ASVs to environmental conditions.

-Comparison of ASV dynamics for the same species in different cruises, trying to determine if the same ASVs are playing dominant roles in different communities.

- Study of the microdiversity associated to the different functional traits/guilds. Link the patterns of microdiversity with the abundance or importance of these guilds for the ecosystem function.

Finally, we will try to determine the functional content of each of these ASVs (or just the dominant ones, if it is not possible to do it for all of them). We will use the software DESMAN (Quince et al 2017) for this purpose. DESMAN works on MAGs such as those obtained by the application of our SqueezeMeta software. DESMAN attempts to separate strains by evaluating the presence of SNPs in the gene sequences of the MAG, that is indicative of the presence of different strains. The software is also capable to infer the gene content of the genome of each of the strains. This will be used for linking the presence of particular accessory genes (those present in the considered strain but not in others) with the abundance of the strain in particular environmental conditions. In this way, we will be able to assess the fitness advantage that certain genes can confer for adapting to different conditions, thus having a much more profound understanding of the functioning of the microbiome. Considering the functional guilds, we could study whether the genes that link the species to a particular guild are conserved in all strains, which will be an indication of the robustness of our guild assignments.

Should the DESMAN software prove not suitable for our purposes, we still can achieve most of the objective above by directly linking the abundance of particular genes to environmental conditions. We

will use simple statistical tests, such as t-test or Welch t-test, to infer the genes of a species that are overrepresented in one environmental condition, indicating that these strains are more likely to contain these genes. This analysis can be done for functional guilds, to determine the importance of guild-marker genes, as stated above.

Tasks and deliverables

Task 1: Determination of microdiversity and dynamics of the community based on amplicon 16S/18S data for the REMEDIOS phytoplankton bloom. Deliverable: Abundance of strains in the REMEDIOS dataset

Task 2: Development of the method for reconstructing 16S rRNA variants from metagenomics miTags. Testing of the method determining microdiversity for the REMEDIOS data and comparison with the results using amplicons (Task 1). Deliverable: Abundance of strains in the REMEDIOS dataset using miTags, and comparison with the previous result.

Task 3: Determination and analysis of microdiversity using the miTags reconstruction method for the temporal and spatial data series of Cambridge Bay (Arctic), Malaspina, and Envision datasets. Deliverable: Abundance of strains in these datasets

Task 4: Resolution of the gene content of the relevant strains in the same samples of task 3. Analysis of the conservation of particular genes in particular environmental conditions (putative adaptive genes). Intra- and inter-species comparison of accessory genes. Analysis of guild-marker genes. Deliverable: Core and accessory genomes for the strains of the most abundant species

WP 3. Structure and function of prokaryotic guilds. Responsible PI: Carlos Pedrós-Alió. Also involved: All PIs

This WP is devoted to find general principles in the way that guilds are organized (structure), how do they function, and how do they contribute to the assemblage of communities. The input to the WP will be the several different guilds studied in WPs 4 to 8. Analyzing and comparing this information, we will try to derive general principles.

The definition of traits and guilds, as this project proposes, provides a functional classification for addressing the functioning of the ecosystems. Our schema is focused on describing the critical elements for the existence and survival of any organism (acquiring energy, acquiring nutrients). Instead of listing all possible metabolic functions for the organism (a classical way to address functional comparison), our proposal focuses in the different ecological strategies to perform critical tasks, which is more sound from the ecological point of view. Besides, our classification schema proposes a well-defined framework for integrating taxa and function from an ecological perspective: traits (critical functions) are performed by guilds (different modes to do it), and guilds are populated by taxa. Then, it is easy to list the possible ecological roles of a taxon (the guilds in which it participates) and to infer its ecological similarity to others (comparing the respective guilds).

The objective of this WP is to illustrate the power of this approach. We will focus in particular questions to explore the structure of guilds and to link them to important ecological questions such as the robustness of microbial communities, or the extent of functional redundancy (how the same function can be performed by different organisms). Specifically, our questions are:

-Do all guilds show a structure of a few abundant and many rare species? Or does this change with different types of resources, environments, or times?
For solving this, we will evaluate measures of diversity such as the Shannon index, using the distribution of abundances of all members of the guild in a given sample. This will inform about the

evenness of the distribution. The comparison between guilds will tell us whether all of them are following similar trends. Comparisons of the same guild in different conditions will inform about the its dynamics and flexibility.

-Do the different members of a guild predominate under different environmental conditions or do a few always dominate the guild?
We will use dissimilarity measures, such as the Bray-Curtis index, to infer the closeness of guild composition when comparing different samples. We will also track in how many instances each member of the guild becomes dominant

-Do guilds with more members provide higher stability to the community?
Here we assume that preserving guild abundance is a measure of community stability. When comparing temporal or spatial series, we will correlate the stability in guild abundance with the number of members of that guild.

-Is functional redundancy a common feature in marine microbial systems? Are there traits/environments in which it is more important?
Functional redundancy will be defined as guilds with more than one member that can replace each other in different samples. We will measure a degree of "competitive exclusion" between members of the guild: in how many instances members of the guild show co-exclusive patterns.

-Can we define "superguilds" that is, groups of organisms sharing the same guild in different traits? E.g. Are PR-containing taxa usually found to degrade polysaccharides and use dissolved organic phosphorus?

To do this, we will compute measures of correlation using guild abundance, to infer the existence of associations between guilds.

How particular members of a guild respond to environmental conditions?
We will use methods of multivariate statistics (such as Principal Component Analysis or Canonical Correspondence Analysis) to study the relationships within guilds, and with the environment. This will inform, for instance, of which environmental conditions are shaping the distribution of particular members of the guild

Tasks and deliverables

Task 1. Analyses regarding the dynamics within guilds (Studying the same guild in different conditions) Deliverable: analysis of guild stability with regard to their composition/diversity

Task 2. Analyses about the similarities between guilds (Comparing different guilds of the same trait in the same conditions). Deliverable: Description of common/specific characteristics of different guilds, with regard to adaptation to a particular condition.

Task 3. How environmental conditions shape guilds (Studying different guilds of the same trait in different conditions). Deliverable: Description of common/specific characteristics of different guilds, with regard to adaptation to diverse conditions

SUBPROJECT 2. Shedding light on marine bacterioplankton and their energy sources from an evolutionary standpoint

PI: José González, Universidad de La Laguna

WP4: Common methodology for metagenomic analyses and careful annotation of marker genes.
Responsible PI: José M. González. Also involved: J. Tamames.

This WP is devoted to the annotation of the key marker genes that will define guilds. These marker genes are responsible of the central function(s) defining the guild. It includes also the metagenomic analysis of the samples. Marker genes will be identified, for each of the guilds, in WPs 5-8. When possible, the identity of these key genes will be determined manually. In other cases, the identification of the genes will be automatic and extracted from the metagenome analysis.

State of the art

Functional diversity is complex due to the intrinsic variability in the protein sequences involved, pathways, and gene organization in operons and levels above operons. In many cases, similar genes are involved in markedly different metabolic pathways (paralogs). In others, enzymes are promiscuous in their substrates with similar 3D structure (Lei et al., 2018; Khersonsky and Tawfik, 2010). In a few cases, orthologs are straightforward to identify and distinguish from paralogs. However, in most cases the study of gene diversity in natural environments is quite difficult and time consuming. Thus, reliable methods to identify the function of any type of gene are still sought for.

Our strategy for annotation will be the careful manual curation of the automatic annotations. One important result of this approach is the identification of novel genes carrying out a given function and genes in novel taxonomic groups of bacteria not previously known to carry out the studied process. In the current project NOVAMAR we developed an approach to this end with the gene *dmdA* (González et al. 2019; Hernández et al., in preparation). This gene is involved in the conversion of the algal metabolite DMSP to the climate active gas DMS. Briefly, we started with sequences of this gene in genomes of bacteria in which the process had been experimentally demonstrated to occur (*Pelagibacter* and *Roseobacter* relatives). We used these sequences to "fish out" similar sequences from metagenomes. The novelty of our approach is in establishing the threshold of similarity that separates orthologs (genes with the same function) from paralogs (genes with similar sequence but different functions). This requires a variety of techniques including building of reliable phylogenetic trees, sequence similarity networks (SSN; Gerlt et al. 2015; Shannon et al. 2003), and manual examination of the gene neighborhood (González et al. 2019). This combination of techniques allows distinction of orthologs from paralogs as well as the identification of lateral gene transfer events. Finally, it reveals the existence of previously unknown genes and taxa. These questions will become more clear with examples in the two following WPs.

SSN are used in one of these techniques. Analysis of SSN is a promising tool to organize genes into clusters of known and unknown orthologs. The method groups peptides by sequence similarity after alignment of all possible sequence pairs. Clusters of known functions can be identified, and information on gene neighborhood will help to classify and identify those involved in metabolic processes. The analysis will be focused on those clusters that show either greater changes during phytoplankton blooms, along the seasons, or whose function is obvious from the gene neighborhood. In either case, they will include genes with known and unknown function. The information of representative sequences will provide clues as to the processes that should be involved. The methods to handle the identification of orthologs groups are as described in González et al. (2019).

This approach will be used to define guild marker genes for WPs 5-8, and to construct the tables of guilds in samples (and taxa in guilds). Since there will be a very large number of genes to be considered, manual annotation of all genes will not be possible. Only a handful of genes will be selected for manual annotation, analysis and database construction. For the rest, a simpler approach will be taken.

Automatic annotation will be carried out for the metagenomic sequences. It involves a number of bioinformatic steps, the main ones being: quality filtering of sequences, assembly or co-assembly, gene prediction, taxonomic and functional assignment, abundance estimation and binning. Consequently, metagenomic analysis is complex, time consuming, and often reproducibility is

compromised. To solve these problems, the SqueezeMeta software developed by Javier Tamames and his group integrates all the pieces of the analysis in a single platform, thus simplifying the analysis and improving the ability to compare between different metagenomes, since all of them are analyzed with exactly the same methodology. SqueezeMeta obtains taxonomic and functional annotation for all the genes in the metagenome, thus providing functional and taxonomic profiles that can be used for comparing between different metagenomes, and correlating the abundance of genes and functions with factors such as the characteristics of the environment.

The methodology that we will use for analyzing the metagenomes will be as follows: The metagenomic sequences obtained in the sequencing step will be co-assembled. Co-assembly implies treating metagenomic datasets from similar environments as one. This way, sequence reads from the same organisms in different samples are pooled, thus facilitating to obtain the most complete possible assembly with longer contigs. After co-assembly, there comes gene prediction and abundance estimation of the genes in each of the samples. The end result will be tables containing the origin, function and abundance of each of the genes detected in the metagenomes of all samples.

Binning follows metagenome analysis. SqueezeMeta recognizes and separate contigs coming from the same genome. Often, these bins are complete enough to correspond to most of the genome of a particular species. Separating genomes provides many advantages since it is possible to infer the metabolic repertoire of the particular species, obtaining a more precise view of their possible roles in the environment. It also makes it possible to produce metabolic models for these organisms, thus facilitating the inference of the response of the cell to different media conditions.

Binning is done by methods based on the differential coverage of contigs in samples: contigs that belong to the same genome must be equally abundant in different samples, that is, their abundances co-vary. Following these co-variation profiles and enriching it with compositional profiles (contigs from the same organism are expected to have similar compositions), it is possible to separate contigs from the same origin.

Tasks and deliverables

Task 1: Construction of databases of manually annotated, key marker genes for each of the guilds. A set of sequences contains the assembled ones from previous metagenome sequences (*Tara* Oceans, Malaspina) and genomes of marine bacteria from our studies and others. Second set contains assembled metagenome sequences from the studies included in this proposal. Deliverable: gene diversity framework for use in the rest of WPs.

Task 2. Metagenome analysis, automatic annotation of genes in the metagenome sequences obtained in this project. Deliverable: Tables containing the origin, function and abundance of each of the genes detected in the metagenomes of all samples.

Task 3. Binning of the metagenomes obtained from our samples. Deliverable: metagenome-assembled genomes (MAGs).

WP5. Guilds with different TonB dependent transporters. Responsible PI: José M. González. Also involved: All the PIs

In this WP we focus on key transporters in the outer membrane. Since transport across the outer membrane is the first step in the utilization of these resources, the objective of this WP is to characterize and organize the genomic information and the identity of the main genes involved in the process of the degradation of complex organic matter.

State of the art

Microorganisms in the ocean thrive in an environment that is mostly depleted in nutrients. They also deal with changes in nutrient concentrations and quality at different spatial and temporal scales. Transport proteins in the membrane play the key role of taking nutrients across into the cytoplasm where the molecules enter metabolic pathways. The first contact with their source of energy, carbon and nutrients is, therefore, through transporters and associated proteins on the membrane. In Gram negative bacteria, however, nutrients need to reach the periplasmatic space before crossing the inner membrane of the cell. This is mostly achieved through two types of proteins: porins and TonB dependent transporters (TBDTs, Fig. WP5-1). Porins are simple tubes that cross the outer membrane. Their size and polarity determine the nutrients that can cross freely. Porins do not need energy but require that the concentration outside of the periplasm be higher than the inside for net entry.

As opposed to porins, TBDTs actively take up nutrients from the outside to the periplasm. For this purpose, TBDTs require energy in the form of a H^+ gradient. Part of the TBDT associated machinery spans the inner cell membrane where it dissipates the H^+ gradient. The energy obtained is translocated to the outer membrane structure through the TonB protein (Fig. WP5-1A). In contrast to porins, the binding of the substrate is usually specific.

TBDTs are expensive to build and to operate and, yet, they are among the most highly expressed genes in both the environment (McCarren et al., 2010; Morris et al., 2010) and cultures (Gómez-Consarnau et al., 2016). This indicates how important they are for the cell to acquire the necessary nutrients. In the marine environment, we find that most TBDT gene copies belong to the *Flavobacteriia* and *Gammaproteobacteria* (data not published). SAR11, the most abundant heterotrophic bacterium in the oceans, has no TBDTs.

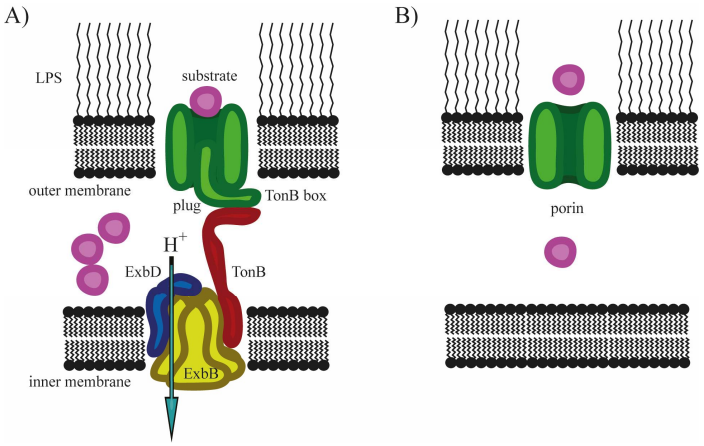


Fig. WP5-1: Components of the TonB system (A) and porin (B). Substrates bind specifically to the plug subunit in the TBDT. The TonB box domain then becomes available to interact with the TonB protein. The contact between the substrate and the plug domain changes the configuration of the plug so that the substrate enters the periplasmic space. Other components of the system participate to allow the system to use energy in the form of a H^+ gradient across the membrane. Porins on panel B allow the substrate to enter the periplasm space passively. In either case, a substrate specific transporter on the inner membrane takes it to the inside of the cell.

Altogether, it seems obvious that possession of TBDTs allows particular strategies to the bacteria that have them. In effect, these transporters allow delineating guilds. A case that we will analyze in detail is that of bacteria using TBDTs to capture and take up polysaccharides, a guild where *Flavobacteriia* are important members. *Flavobacteriia* degrade complex carbohydrates thanks to the

expression of genes organized in polysaccharide utilization loci (PULs). The system is an adaptation to a competitive environment where polysaccharides are an important source of carbon and energy. The recognition, attachment and degradation of complex molecules is a coordinated process that avoids the leakage of enzymes or their products.

We contemplate two scenarios that would explain the diversity of TBDT genetic repertoire in marine bacteria based on what has been published, including articles from our group: (a) Each of the TBDT is designed for a specific substrate or substrates with very similar structures. Then, it might be the case that the most highly expressed TBDT genes encode a transporter that is taking up some nutrient that is necessary for growth. (b) TDBTs might be taking up a broad range of substrates, either the product of the degradation of polymers or oligomers dissolved in the medium. In this case, it would be expected that highly expressed TBDT genes encode promiscuous transporters. Considering the wide range of substrates to grow on and the changing conditions bacterioplankton could encounter, it would have to rely on porins when the concentration outside of the cell is high enough, TBDTs designed for specific growth factors no matter the conditions, those involved in the degradation of polymers and uptake of the products, and TBDTs specific for substrates at low concentration.

Preliminary data in our group suggests that some TBDT genes are shared across *Flavobacteriia* genomes, and at least in some cases, are also highly expressed in the environment and lab cultures. We hypothesize that some of the TBDTs are able to take up a broad range of substrates. The most active TBDTs might be involved in the degradation of complex organic matter and would be less specific than those involved in the uptake of, for example, vitamins or Fe chelates. Considering that bacterioplankton members need to allocate resources effectively, it is intriguing that they spend a great deal of energy for the synthesis of some of these TBDTs. The reason might be a promiscuous system to take up nutrients. Work by Reintjes et al. using microscopy (Reintjes et al., 2017) suggests such a widespread mechanism in marine Bacteroidetes. Such a system would be advantageous as DOM is likely to be a complex mixture, where each component is at low concentrations (Hertkorn et al., 2013; Hansman et al., 2015; Zark et al., 2017). The alternative would be that bacterioplankton would have to depend on growth only during the right conditions when the concentration of some substrates are high enough, stochastic variations of the expression of the right genes (Mitchell et al., 2009), or constitutive expression of transporter genes, which would be highly inefficient. In contrast, a promiscuous TBDT system would guarantee growth substrates in the periplasm at higher concentrations than outside the cell.

Regarding TDBTs designed for key growth factors, our group described a TBDT involved in the uptake of vitamin B1 (Gómez-Consarnau et al., 2016) and found it to be widespread in marine *Flavobacteriia* and *Gammaproteobacteria*, such as *Alteromonas*, that dominate the bacterioplankton (Alonso et al., submitted). Other TBDTs are specific for iron uptake, usually iron bound by siderophores. A preliminary examination of metagenomes from the Cambridge Bay time series, suggests that iron is an intensively disputed resource in the Arctic. Some bacteria such as *Gammaproteobacteria* synthesize specific siderophores and siderophore transporters. *Flavobacteriia* in turn do not spend energy in siderophore synthesis, but have the necessary transporters, thus behaving as cheaters. These two examples show the intricacies of microbial ecology and the kinds of issues that we will address here.

We propose to take a detailed look at those TBDT genes that are the most highly expressed, are involved in key metabolic processes (polysaccharide breakdown, vitamin or Fe uptake) and/or are representative of abundant *Flavobacteriia*. This requires, as mentioned in WP4, a very careful curation of the genes of study. We make use of the information we have gained in model organisms. We have chosen the genus *Polaribacter* for three reasons. First, *Polaribacter* is one of the most abundant genera of *Bacteroidetes* found in the marine environment. This genus is particularly abundant in polar and coastal regions (Pommier et al., 2005; Díez-Vives et al., 2014; Nedashkovskaya et al., 2013;

Malmstrom et al., 2007). The second reason is that we have manually annotated the genome of a *Polaribacter* strain (González et al., 2008) and therefore we know very well these genomes and have the expertise that is required. Third, we have transcriptomes of growing cultures of different *Polaribacter* strains. Expression level will be included to identify the function of the genes. Lastly, currently, our group is involved in the analysis of a number of *Polaribacter* SAGs from the Arctic. This is an opportunity to apply our knowledge to a region where *Polaribacter* is especially abundant.

The hypothesis we will test in this WP is that a set of TBDTs are flexible in their substrate uptake. If this is the case then the diversity of these genes and their operons will be small and common across *Flavobacteriia* genomes. Members of this guild that start the degradation of complex DOM should share similar genes. The alternative will be TBDT genes specifically designed for each of the substrates and therefore a much higher diversity of uptake gene repertoire.

Tasks and deliverables

Task 1: Determination of diversity of TBDTs in representative sequences. Organization into TBDT families. Deliverable: identification of key set of genes involved in the degradation of complex DOM.

Task 2: Analysis of the level of expression of TBDT families in single genomes as a means to identify and classify TDBT families. Deliverable: identification of key marker genes with a general role in the degradation of complex DOM.

Task 3: Diversity of TBDT families of interest in the natural environment. Deliverable: identification of set of genes involved in the initial degradation of complex DOM that identifies this guild.

Task 4: Analysis of patterns of expression across *Flavobacteriia* taxa, environmental conditions, and fluctuations in diel cycles. Conservation of key genes under particular environmental conditions. Deliverable: dynamics of guilds involved in the degradation of complex DOM under fluctuating conditions.

WP6. Photoheterotrophic guilds. Responsible PI: José M. González. Also involved: All PIs.

We will focus on light utilization mechanisms in heterotrophs. Having proteins to be able to use light as an energy source should have an impact on the carbon-cycling in the marine environment. We will analyze proteorhodopsins (PR) and bacteriochlorophyll *a* (Bchla) genes that are highly expressed in metatranscriptomes. The objective will be to organize the genetic information and follow the expression pattern in natural settings. We hypothesize that (a) light harvesting mechanisms are involved in the initial utilization of DOM, degradation of complex organic matter and uptake of substrates at low concentration necessary for growth, and (b) different taxa make use of light in different ways.

State of the art

Sunlight is the principal source of energy in the biosphere. It comes as no surprise that organisms have evolved means to harvest it. Chlorophyll, or Bchla, and carotenoid molecules absorb light in the photosynthetic complexes of plants and bacteria. Then they convey the energy to reaction centers for energy transformation. Compared to these photosynthetic complexes, PRs are much simpler, composed of a single protein with retinal as the light absorbing molecule. In heterotrophic prokaryotes with any of these light absorbing mechanisms, i.e., Bchla - or PR-based systems, energy from sunlight is converted into a H^+ gradient. When exposed to light, this energy can then be dissipated in key cell processes such ATP generation, motility, or substrate uptake into the cell. Heterotrophs able to convert

light into useful energy have an advantage over those that only fend off its damaging effect (Teira et al., 2019). Thus, an array of possible functions has been proposed for PR (Fuhrman et al., 2008; Martinez et al., 2007). In the case of Bchla, however, the main finding is an increase in growth rate in the bacterioplankton fraction that harbors the gene (Ferrera et al., 2017; Koblížek et al., 2007).

Both mechanisms for light energy harvesting caused a turning point at the beginning of this century. Aerobic photosynthetic bacteria had been discovered about 40 years ago. However, their importance in the marine environment was realized only in the early 2000s (Kolber et al., 2000, 2001). Other authors, however, consider that their impact is limited (Goericke, 2002). Around the same time, a new and unsuspected mechanism to obtain energy from light, through PR, was discovered in the ocean (Béjà et al., 2000) and later quantified in metagenomes (Venter et al., 2004). The theoretical scenario in which there was a clear separation between primary producers and consumers suddenly needed a revision. Mixotrophy, or the ability to use both light and organics, seems to be the rule rather than the exception. Surprisingly though, there have been only a few demonstrations of the function of both mechanisms in the marine bacterioplankton (Gómez-Consarnau et al., 2007, 2010, 2016; Akram et al., 2013; Palovaara et al., 2014). Most of the progress has been in the diversity of taxa that contain the gene (Pinhassi et al., 2016) or novel versions of the gene through functional metagenomics (Pushkarev et al., 2018). Considering the wide range of taxa that harbor the PR gene and the estimated high fraction of bacterioplankton cells (around 60%), it is expected that this adaptation plays a main role in the oceanic bacterioplankton physiology. Thus, Palovaara et al. (2014) found that the carbon flow within the cell changes when exposed to light. As opposed to PR, the fraction of bacterioplankton cells that harbor the Bchla genes is smaller, less than 10%. Kirchman and Hanson (2013), based on bioenergetics calculations, predicted that bacteria with the Bchla system harvest more light energy than do PR bacteria. However, Gómez-Consarnau et al. (2019), based on the concentration of retinal in sea water, suggested that PR absorbs as much solar energy as chlorophyll *a*. At any rate, both Bchla and PR-based mechanisms should have a significant impact in the carbon flow in the marine environment (Kolber et al., 2001; Fenchel, 2001; Karl, 2002; Copley, 2002; Béjà et al., 2002; Giovannoni et al., 2005).

We hypothesize that TBDTs are fueled by the activity of PR, since it functions as a proton pump in the presence of light. TBDTs are among the cellular processes that consume the H^+ gradient. PR genes in *Flavobacteriia* are also relatively highly expressed and strongly regulated by light (Palovaara et al., 2014; Gómez-Consarnau et al., 2016). Thus, there is a strong link between WPs 5 and 6.

In a study we have submitted for publication (Alonso et al), we analyzed patterns of PR expression in the *Gammaproteobacteria*. We found that there have been many cases of lateral gene transfer between the *Gammaproteobacteria*, and that strains of *Vibrio* are no exceptions (Akram et al., 2013; Gómez-Consarnau et al., 2010). Fig. WP6-1 shows the diversity of rhodopsin peptides expressed by metatranscriptome analysis in samples taken through two consecutive years on the coast of northern Spain. Within each taxonomic group, we find novel subgroups with no representative sequences in the databases.

We have shown that PR expression does not vary over short periods of time (Palovaara et al., 2014). However, in that study (as in many other studies) gene expression was quantified in bulk or in wide phylogenetic taxa that group too much diversity. We hypothesize, however, that the members of the community respond differently and that there are taxa that respond to light because they gain energy from it. Our goal is to determine those microbial taxa that respond to light over short periods of time. A diel cycle should be the ideal setup since the community does not change substantially during the time the experiment lasts. The response of the bacterioplankton should be tightly regulated and in concert with other genes involved in dissipating the H^+ gradient. The alternative hypothesis would be that bacterioplankton do not gain substantial energy to grow more efficiently than those members of

the community that do not contain the PR gene. If this were the case, then PR should have a more specific function such as the uptake of nutrients at transient low concentration.

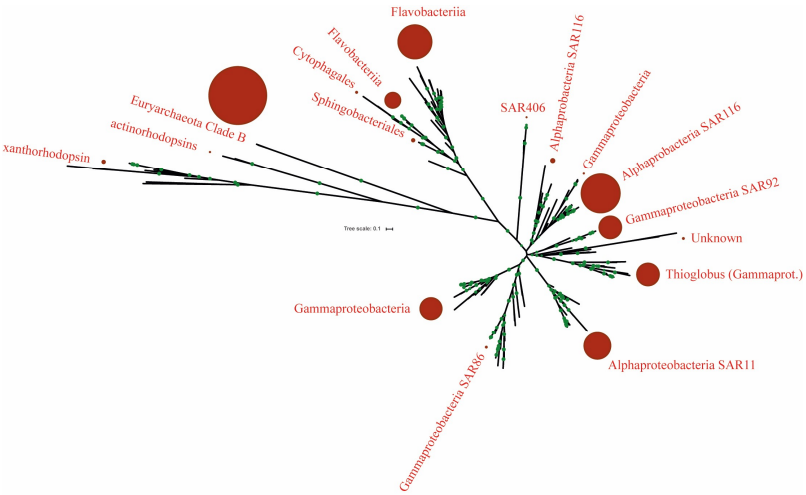


Fig. WP6-1. Diversity of rhodopsin genes expressed in coastal samples taken in northern Spain during a two-year sampling period. The tree is based on peptides in the database that represent the diversity of sequence reads in the metatranscriptomes. The size of the circles is proportional to the number of sequence reads in each group. Notice the number of clusters that belong to the alphaproteobacterial group SAR116 and Gammaproteobacteria*. Unpublished data.*

Tasks and deliverables

Task 1: Diversity analysis of the guilds containing genes encoding PR, PufM, PufL and BchX. Study of the diversity of photoheterotrophs in the samples. Deliverables: information on the diversity of photoheterotrophs in the environment and abundance rank of each taxonomic group.

Task 2: Analysis of interaction between the photoheterophic guilds and other guilds, such as those involved in the degradation of complex DOM. Deliverables: Abundance rank for each of relevant taxonomic groups and dynamics during bloom conditions.

Task 3. Analysis of expression of the genes in diel cycle experiments. Fine-tuned analysis of main taxa involved. Identification of taxa that respond differently during the cycles. Deliverables: dynamics of photoheterotrophic guilds in changing conditions.

SUBPROJECT 3. Towards the identification of functional groups of marine prokaryotes in carbon and nutrient utilization

PI: Sandra Martínez-García (Universidad de Vigo)

Marine prokaryotic structure and function are known to be strongly controlled by resource availability (Martinez-García et al. 2010, 2015, Teira et al. 2011, 2016). Some prokaryotes are now recognized as cosmopolitan and abundant throughout the global ocean like SAR11 and *Prochlorococcus,* while others are known to dominate microbial communities under specific conditions, like in the case of *Roseobacter* that grows associated with phytoplankton blooms. Changes in biotic and abiotic environmental conditions affecting nutrient availability may occur along different temporal or spatial scales and, therefore, variability in the composition of prokaryotic communities

may be found at both scales. If resource availability structures marine microbial communities (as is the case in plants and animals), it is of extreme importance to decipher the ecological mechanisms controlling this organization. Thus, if one is to study the functional roles of different taxa in nutrient cycling in marine microbial communities, it is mandatory to use not only taxonomic but also functional information. In a recent work, Haggerty and Dinsdale (2017) showed, based on metagenomic analyses, that patterns of biogeography in bacterial communities differed depending on the description of taxa or function. This result is important if we aim at studying functional redundancy of nutrient utilization in marine microbial communities: if taxonomy and function do not vary in conjunction, one could conclude that the same taxa have different roles in different environments and/or different taxa have the same function in different environments.

The definition of guilds and traits for prokaryotic nutrient incorporation in different ecological settings will have important implications for understanding marine ecosystem function. Information derived from this project may also help current marine ecosystem models to improve the information included on the functional role in nutrient cycling of different taxa currently agglutinated in the "black box" approach or simply represented as a remineralization constant linear rate (Ward et al 2012) that oversimplify the function of microbial communities.

In the present project we propose the use of information on specific functional genes (metagenomics and metatranscriptomics) related to nutrient utilization in different ecological settings. The following two WPs will therefore focus on four ecologically relevant traits for ecosystem function: Organic Carbon source (WP7) and Nitrogen, Phosphorus, and Iron sources (WP8).

WP7 Identification of different prokaryotic guilds associated to different carbon sources.
Responsible PI: Sandra Martínez-García**.** Also involved: Jarone Pinhassi and all PIs

State of the art

Marine bacteria and archaea process an important part of the organic carbon fixed by autotrophic organisms in the ocean and it is expected that the availability and quality of this organic carbon vary along different temporal and spatial scales (Alonso-Sáez et al. 2007). Dissolved organic matter (DOM) released by phytoplankton consists of a variety of known and unknown components and differs in molecular weight and quality depending on the taxonomic composition and the physiological status of the phytoplankton cell it is derived from. Accordingly, DOM influences the composition of marine bacterial assemblages (Pinhassi et al. 2004). The quantity (Sarmento et al. 2016) and differential composition of phytoplankton-derived organic matter affects heterotrophic prokaryotes since the latter are known to exhibit different potentials to utilize (and eventually remineralize) different compounds (Alonso-Sáez and Gasol 2007). Accordingly, the bacterial assemblages associated with different phytoplankton species differ substantially (Pinhassi 2004). Several studies in the last decade have proven that different prokaryotes are present not only in different phytoplankton blooms, but also in the different stages of the same phytoplankton bloom (from initiation to decay) as the composition and lability of organic matter changes through its development (Teeling et al 2012). For example, *Flavobacteriia* have been found to be more abundant during the bloom decay phase (Pinhassi et al. 2004), as they have been proposed to be specialized in the conversion of complex High Molecular Weight (HMW) organic matter to simpler Low Molecular Weight (LMW) organic matter (González et al. 2011). Similarly, in coastal areas, allochthonous inputs of organic matter of different composition and origin (e.g. atmospheric, riverine) are known to affect the structure and function of bacterioplankton communities (Martinez-García et al 2015, Teira et al 2016).

Although different heterotrophic bacteria and archaea may be adapted to use different portions of the DOM pool (or use particular DOM compounds with different efficiency; Alonso-Sáez et al 2009,

Gómez-Consarnau et al. 2012), few studies have systematically investigated similarities or differences in C substrate use. Experimental work with model bacteria has uncovered mechanisms by which members of the Roseobacter clade degrade marine snow and use polyamines, carbon monoxide, and aromatic compounds. Some prokaryotes (like the abundant Bacteroidetes) are specialists in degradation of polymeric organic matter probably coming from phytoplankton blooms in the ocean (Fernández-Gómez et al. 2013, Mestre et al. 2017). In general, genomes of bacteria specialized in the degradation of polymers usually include a high abundance of TonB-dependent receptors specialized in the transport of oligosaccharides from the outer membrane into the periplasmic space and in the binding of polysaccharides to the outer cell membrane. Marine prokaryotes also have different strategies to specifically use particulate organic matter, for example, through exoenzymes or surface adhesion and gliding. This complex organic matter includes HMW compounds of different nature and specific metabolic pathways are used by different prokaryotes to utilize them. Thus, the genomic basis for the degradation of ***lipids*** (fatty acid metabolism), ***proteins*** (peptidases) and ***polysaccharides*** (glycoside hydrolases) have already been identified in *Bacteroidetes*, one of the major taxa of marine heterotrophic bacterioplankton frequently found on macroscopic organic matter particles and associated to phytoplankton blooms (Fernández-Gómez et al. 2013) (Table WP7-1). On the other hand, some prokaryotes specialize in the utilization of simpler LMW organic carbon compounds (Alonso-Sáez et al. 2012) that are usually acquired through ABC-transporters. For example, SAR11 clade bacteria contribute to carbon cycling in oligotrophic environments thanks to their ability to use a broad range of ***sugars***, ***amino acids***, ***carboxylic acids***, and ***osmolytes*** at low ambient concentrations (Giovannoni et al. 2005). ***C1 compounds*** are also potentially relevant for marine bacteria, since a member of the OM43 clade, a betaproteobacterial clade that is abundant in productive coastal environments, is an obligate methylotroph and the SAR11 clade representative *Candidatus* Pelagibater ubique is known to carry out demethylation and oxidation of different C1 compounds (Table WP7-1).

Trait	Guild	Marker genes
C source	Lipids	K01897, K00249, K01692, K00074, K00632, K00626, K00252
	Proteins	VanY, Peptidase_M16_C, Peptidase_S41, Peptidase_M3
	Polysaccharides	Glyco_hydro_43, Alpha-amylase, SLT, Glucosaminidase
	Sugars	C134_0264-C134_0271, C134_0769-0772, C134_1290-1293
	Amino acids	C134_0805-0807, C134_0953-0957
	Carboxylic acids	C134_0316
	Osmolytes	C134_0797-799, C134_1297-1299; 1301-
	C1 compounds	Fe-ADH (methanol), glnT, glxBCD, soxBDAG (methylamine), bhmT, sardh, dmgdh (GBT), gdvT, dmdA (AMTs), fdhF, fdsB, fdhD, mobA, moeA, fhs, metF, folD (TFH-linked oxidation, gfa, GD-FALDH, FGH (GHS-dependent pathway)

Table WP7-1. Marker genes for each of the different guilds described for the trait "Carbon source"

An effort to organize and group this functional and taxonomic information, in order to identify the main guilds related to the trait "Organic Carbon Source" in different ecological settings, remains to be done.

Thus, we hypothesize that within the trait "Organic Carbon Source", different guilds (examples marked in bold in the text above) may be present in different ecosystems (e.g. productive coastal waters versus oligotrophic open ocean environments) or at different times (e.g. before, during or after a phytoplankton bloom). Different guilds (i.e. carbon sources) will be identified attending to published literature. However, and as usually happens when one is operating with "meta-omics" databases, we also acknowledge a certain degree of discovery-driven results that will emerge from the analysis of the broad set of data we will have available.

Some example of guilds may be:

a) A "C1 user Guild" that is hypothesized to be found in ecosystems of different productivity and composed by different prokaryotes in each environment: for example SAR11 and OM43 clades will be included in the "C1 users guild" in ecosystems with low and high productivity, respectively. We hypothesize that the "C1 users Guild" will be characterized by the following

marker genes depending on the C1-compound used or the metabolic pathway utilized: Fe-ADH, glnT, bhmT, gdvT, fdhF and gfa, among others (Table WP4A).

b) A "Polysaccharides Decomposer Guild" that is hypothesized to be found during phytoplankton blooms when polymeric organic matter is highly available. We hypothesize that the "Polysaccharides Decomposers Guild" will be characterized by the following marker gene domains for glycoside hydrolases: Glyco_hydro_43, Alpha-amylase, SLT, Glucosaminidase. This guild is hypothesized to be largely composed by *Bacteroidetes* among other taxa.

Task 1: Exhaustive review and analysis of the previous literature on different organic C-sources for marine prokaryotes.

The relevant literature will be examined, including studies performed in different environments and also experiments using model systems. The deliverable of this task will be a list with organic compounds experimentally demonstrated to be utilized by marine prokaryotes as carbon source.

Task 2: Performance of perturbation experiments and sequencing of samples.

In order to complete the dataset already available and described in the general introduction to the project, we identify the need for information on the functional short-term responses of prokaryotic communities to organic matter derived from different phytoplankton species. The deliverable of this task will be a metagenomic and metatranscriptomic database (13 samples each) that will be included in the analysis described above and that will be used in WP 5-8.

Natural prokaryotic seawater cultures from the coastal system of NW Spain will be subjected to DOM inputs obtained from concentrated phytoplankton cells. Samples will be taken on board RV Kraken from a central station in the Ría de Vigo (St2) (Fig. WP7-1). The experimental design will include three treatments so organic matter from representatives of different photoautotrophic functional groups will be included:

1) **Control** (no addition made), 2) **Diatom** concentrate addition (e.g. *Chaetoceros* spp), 3) **Dinoflagellate** concentrate addition (e.g. *Dinophysis* spp) 4) **Cyanobacteria** concentrate addition (e.g. *Synechococcus* spp.)

Seawater cultures will be prepared by gentle filtration of natural seawater through a 0.8µm pore-size filter in order to remove phytoplankton and eukaryote grazers of prokaryotes. Experiments will last 48h and will be incubated in a mesocosm setup available at ECIMAT in which irradiance and temperature will be kept similar to the in situ ones (Fig. WP7-1). Samples will be taken every 24h. For these experiments, we will use 10 L UVR transparent bags, which will be filled with 9 L of 0.8µm-filtered surface seawater from the Ría de Vigo. Phytoplankton cultures will be grown up to mid exponential phase when biomass will be harvested by gentle filtration, sonicated, and homogenized. Phytoplankton cultures will be provided by the Unidad de Cultivos Marinos de la Estación de Ciencias Marinas de Toralla (https://cim.uvigo.gal/en/ecimat/) (Fig. WP7-1). Bacterioplankton abundance and functional profiles will be assessed by flow cytometry, metagenomic and metatranscriptomics, respectively. Inorganic and organic nutrient concentrations will be measured.
These experiments will be made in parallel with field sampling designed as proof-of-concept for MinIon device (WP1) so the same prokaryotic assemblages will be studied in the field (*in situ* conditions) and under controlled conditions in the lab (microcosm experiments explained here). During a 3-day field sampling we will collect surface samples in the Ría de Vigo. First day we will sample 3 different stations: an inner station heavily affected by river run-off (St.1), a central station situated in the middle of the Ría de Vigo (intermediate conditions, St. 2), and an outer station close to the Cies Islands where off-shore influence and the upwelling of nutrient-rich deep water is more important (St.3). Real-time analysis of

metagenomic data with MinIon as explained in WP1 will enable comparisons of the functional profile of natural bacterial assemblages exposed to distinct allochthonous nutrient inputs on day 1, and will allow us to decide which community is the most interesting to sample the following days. Phytoplankton and bacterioplankton abundance (clorophyll a concentration and flow cyrometry measurements, respectively) and inorganic and organic nutrient concentrations will be also measured.

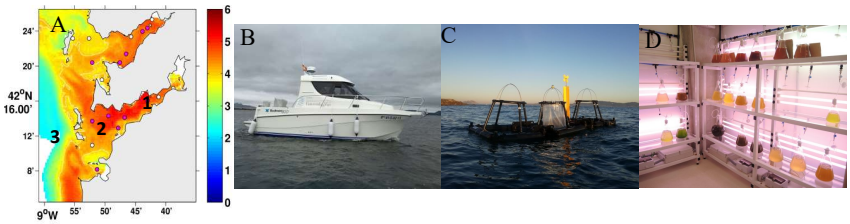


Figure WP7-1. A. Surface Clorophyll a map of the Ría de Vigo showing the sampling stations. B. RV Kraken. C. Floating structure and mesocosms. B. Phytoplankton cultures available at ECIMAT.

Further analytical details:

-Inorganic nutrients concentration, dissolved organic carbon and nitrogen concentration, bacterial biomass and chlorophyll a concentration will be determined following the methods in Martínez-García et al (2010).

-Metagenomic and metatranscriptomic analyses will be performed following the methodology in Bunse et al 2016 and Tamames & Puente-Sánchez 2019, respectively.

Task 3: Identification of marker genes for guilds associated with the different metabolic pathways to utilize each of the C-containing compounds in the list from Task 2.

One or two genes per process will be chosen as "marker genes" attending to their relative importance in the metabolic pathways or the abundance in which they are found in available databases. To determine the genes that will act as markers of particular guilds, we will create a table containing the presence or absence of each gene of the selected processes in the available genomes of marine microorganisms, particularly in the MAR databases. These genomes will be annotated manually in the different guilds. Using a Bayesian approach, we will then calculate a score associated to each gene, indicating the likelihood of the association of the gene and the guild. This will serve to compute the probability of a new genome to belong to a particular guild. The methodology for this task is described in WP4.

The deliverable of this task will be a table in which each compound from the list obtained in Task 2 corresponds with one or several marker genes defining the different guilds on C-sources.

Task 4: Search for different guilds using different C-sources in different available datasets

We will look for different prokaryotes containing the marker genes in their genomes in the different samples and will describe the taxonomic composition of the different guilds on C-sources associated with those marker genes. We will follow the guilds and their corresponding marker genes in the different metagenomic datatsets we have available, including transects in space, monitoring of algal blooms, day-night cycles, and seasonal studies. We will look at the presence/absence of the different guilds in the different samples as well as the taxonomic composition of each guild in the different samples.

The deliverable of this task will be a list containing the guilds on C-sources present in each of the samples analyzed and their taxonomic composition.

Task 5. Relationships between the taxonomic composition of the different guilds using different C-

sources and the characteristics of different ecological settings.

We will look for correlations between the presence/absence of the different guilds on C-sources as well as the taxonomic composition of each guild in the different samples with biotic and abiotic environmental conditions of each specific ecological setting.

The deliverable of this task will be the discussion on the relationship between the presence and composition of different guilds and the different ecological settings. We will explore the differences between ecological settings (e.g. productive/unproductive, different moments of a bloom, day/night, winter/summer) on the composition (number and diversity of members) of guilds related to the trait under study: "C source". We will discuss on the posible functional redundancy found in different environments (i.e. the same guild is composed by different taxa in different ecological settings). Bioinformatic and statistical methodologies for the different questions addressed in this task are explained in detailed in WP3 (e.g. diversity and dissimilarity measures, correlations, co-existence patterns, PCAs). An interesting discussion included in this task will be to explore the coincidences or differences between the information extracted from metagenomic (potential functional profile of an organism) and metatranscriptomic (actual functional expression) databases.

WP8. Identification of prokaryotic guilds associated with different Nitrogen, Phosphorus and Iron sources. Responsible PI: Sandra Martínez-García. Also involved: Jarone Pinhassi and all PIs

State of the art

The diversity and activities of marine microbes have a profound influence on global nutrient cycles of Nitrogen (N), Phosphorus (P), and Iron (Fe), among other nutrients. Conversely, the availability of resources for microbial growth (i.e. the distribution of organic and inorganic forms of elements like N, P and Fe changes markedly at a variety of spatial and temporal scales in the ocean (Moore et al. 2013) and is likely to affect prokaryotic acquisition of resources. The study of different marine prokaryotes has prompted identification of different nutrient sources and mechanisms used by marine prokaryotes to acquire nutrients. Also, metagenomic and metatranscriptomic analyses are contributing with relevant knowledge to the understanding of the uptake and processing of nutrients by complex prokaryotic communities and its genetic basis (Teeling et al. 2012, Satinsky et al. 2014, Ferrera et al. 2015, , Bunse et al. 2016). These studies describe important differences in nutritional strategies between taxonomic classes in the expression of transport systems for nutrient uptake by prokaryotes in different naturally occurring ecological settings. The main strategies found in the surface prokaryotic communities for N, P and Fe utilization are described below.

Nitrogen (N) is available to microorganisms in several inorganic and organic forms (Kirchman 2000) (Table WP8-1). ***Ammonia*** is the most important inorganic form of nitrogen for prokaryotes, and they can account for a large fraction of total ammonia uptake in both pelagic and benthic marine environments. Metabolic pathways for ammonia assimilation as well as ammonia transport and their genomic basis at high and low ammonia concentrations have been now clearly identified in the environment (Zher et al 2002). On the other hand, it is known that some, but not all, prokaryotes are capable of growth on ***nitrate and nitrite*** as a sole N source and the genes encoding these metabolic pathways have also been described for both nitrate and nitrite assimilation and reduction (Satinsky et al 2014). α-Proteobacteria (mostly Roseobacter-like groups), γ-Proteobacteria and Bacteroidetes are generally found to be the dominant bacteria utilizing nitrate in the ocean. ***N_2 fixation*** is the dominant mechanism for introduction of new N into the biosphere. N_2-fixing microorganisms in the ocean are usually identified through the presence of the *nifH* gene which encodes the Fe protein component of nitrogenase, although other genes related to nitrogenase activity have been now identified. *NifH* gene

clone libraries show that non-cyanobacterial diazotrophs (i.e. heterotrophic prokaryotes) are widespread in lakes, estuaries and marine waters, where α-, β- and γ-proteobacteria N_2-fixers are particularly prevalent in marine waters (Rienmann et al 2010). Heterotrophic bacteria are also capable of utilizing dissolved organic nitrogen compounds (Varela et al 2006) including mainly ***amino acids***, ***urea***, ***amines*** and ***humic substances***. In heterotrophic bacteria, evolutionary adaptations to N limitation could involve preferential use of certain amino acids to minimize N content in proteins along with changes in genome and cell size. Roseobacter clade members also have genes involved in utilization of methylated amines, thus indicating that methylated amines could be important in N and C cycling. In contrast, *Cand.* P. ubique exhibits a PII independent response to nitrogen limitation and devotes more resources to the assimilation of nitrogen-rich organic compounds in response to N stress. On the other hand, deamination of glutamate and aspartate and the subsequent liberation of ammonia have been suggested to be used for nitrogen remineralization by *Flavobacteriia* in phytoplankton blooms. This shows that there are significant linkages between the metabolisms of nitrogen and carbon.

As a resource for microorganisms, phosphorus (P) is most frequently considered in the form of ***inorganic phosphate***, and the genomic features related to the uptake of inorganic phosphorus differ mainly depending on its availability. While oligotrophic bacteria like SAR11 at low nutrient extremes are known to count with high affinity transport systems for P, Roseobacter clade members from areas with relatively higher P concentrations exhibit low affinity phosphate systems (Newton et al 2010). It is now known that diverse taxa differentially express homologues of genes for ***polyphosphate*** (salts or esters of polymeric oxyanions) degradation under different conditions (e.g. sulfur-oxidizing microorganisms). However, P can also be bound in ***organic compounds*** (notably nucleic acids). In this regard, Sebastián and Ammerman (2009) suggested that different phosphatases (enzymes that cleave a phosphoric acid monoester, e.g. PhoA and PhoX) are relevant in different ecological settings. PhoA may be important during algal bloom episodes when *Bacteroidetes* are abundant. However, PhoX appears to be much more widespread and more related to γ- and α-proteobacteria, bacterial groups with probably lower P-demands than *Bacteroidetes*. Sebastián and Ammerman (2009) also showed that the number of phosphatase transcripts decreased drastically when phytoplankton bloom crash, probably related to a release of nutrients as the phytoplankton cells break open. The abundance and use of alternative P sources is increasingly recognized. Vila-Costa et al. (2019) showed that alkaline phosphatase activities significantly decreased when organophosphate triesters were available, indicating a relief on P stress, consistent with the role of organophosphate triesters as potential P sources for marine prokaryotes. Interestingly, analysis of metagenomic data-bases for phosphonate utilization genes has shown that they are widespread and abundant among marine prokaryotes, suggesting that phosphonates metabolism is likely to play an important role in P-depleted surface waters, as well as in the more P-rich deep-water column. On the other hand, ***phosphite*** may be an important and previously unrecognized source of P for marine prokaryotes, since the genes for phosphite utilization have been found in some marine bacteria (Martínez et al. 2012) (Table WP8-1).

Trait	Guild	Marker genes
N source	Ammonium	glnA
	Nitrate	nasFEDCBA
	Nitrite	nir
	N_2 fixation	nifH, nifD, nifE, nifB, nifN, nifX)
	Amino acids	slc17A
	Urea	UrtA, ureA, ureB, ureC, ureD, ureE, ureF, ureG, ureH
	Amines	tmm, gmaS
	Humic substances	vanA, vanB, bphA1
P source	Phosphate	pitA (low affinity), pstA/pstB/pstC/pstS (high affinity)
	Polyphosphate	ppk1, ppk2 and exopolyphosphatase
	Organic compounds	PhoA and PhoX
	Phosphonates	phnW, phnX, phnCDE, phnG/phnH/phnM, phnF, phnN, phnO and phnQ
	Phosphite	ptxA, ptxB, ptxC, ptxD
Fe source	Inorganic Fe+2	EfeBOU, FeoB, YfeABCD
	Inorganic Fe+3	YfuABC, FbpABC, FutABC
	Organically complexed Fe	PchABCDEFGHI, PvdADEFHIJLMNOPQ, PvsABCDE
	Heme groups	hmuR, hmuY, hmuT, hmuU, hmuV, hemS

Table WP8-1. Marker genes for each of the different guilds described for the traits "Nitrogen source", Phosphorus source" and "Iron source".

A comprehensive characterization of the main guilds related to nutrient (N, P, Fe) utilization in marine microbial food webs from different ecological settings has not been done.

Iron (Fe) is known to be essential for cell metabolism being necessary for key processes like photosynthesis, respiration, and nitrogen fixation. Since the bioavailability of Fe is very low in many oceanic regions, marine prokaryotes have developed different strategies for iron uptake. (Table WP8-1). Inorganic iron uptake (reduced $\boldsymbol{Fe^{+2}}$ and oxidized $\boldsymbol{Fe^{+3}}$ more important in coastal waters and in open ocean environments, respectively) has been described in Actinobacteria, α-, β- and γ-proteobacteria and Bacteroidetes among others (Toulza et al. 2012). In well oxygenated surface waters inorganic Fe concentrations are very low and marine bacteria mostly utilize ***organically complexed Fe***. In this regard, different bacteria among α-, γ-proteobacteria and firmicutes have evolved the capacity to synthetize small high-affinity iron-chelating compounds: siderophores. Other bacteria however, are able to take up naturally-occurring siderophores like some representatives of γ-proteobacteria and Bacteroidetes (Toulza et al. 2012). Different kinds of siderophores such as petrobactin, ochrobactins, synechobactins, ferrioxamines, amphibactins, alterobactin or dicitrate have been described to be important in different ecological settings in the marine environment (Boiteau et al 2016). Other bacteria, known to be associated to particles like γ- and α-proteobacteria and including Roseobacter and Bacteroidetes, have evolved the capacity to acquire Fe from ***heme groups*** or heme-containing proteins, for which they utilize hemophores or simple direct uptake (Hopkinson & Barbeau 2012). Finally, the storage of Fe inside the cell has been also observed as a relatively common strategy in Bacteroidetes, Actinobacteria, α- and γ-proteobacteria among others as the concentration of Fe increases (Toulza et al. 2012).

Thus, we hypothesize that within the traits "Nitrogen Sources", "Phosphorus Sources" and "Iron Sources" different guilds (marked in bold in the text above) may be present in different ecosystems (e.g. nutrient-rich waters versus oligotrophic open ocean environments) or at different times (e.g. winter mixing, summer stratification). We can provide some example guilds but, as explained in WP4, apart form the guilds highlighted in the text we acknowledge the very probable possibility that more guilds will emerge when analysing the different datasets.

Some examples of guilds may be:

a) An "Amine utilization Guild" that is hypothesized to be found during phytoplanktom blooms. The "Amine utilization Guild" is hypothesized to be characterized by the following marker genes: *tmm*, *gmaS* and to be composed by the Roseobacter genus among others.

b) A "Dissolved Organic Phosphorus (DOP) utilization Guild" that is hypothesized to be present during phytoplanktom blooms and characterized by the phosphatases marker genes PhoA and

PhoX. The "DOP utilization Gild" is hypothesized to be composed by Bacteroidetes in nutrient-rich envioronments and γ- and α-proteobacteria in environments with lower P availability.

c) An "Organically complexed Fe utilization Guild" that is hypothesized to be present in Fe-limited environments and characterized by the presence of siderophore synthesis marker genes like PchA, PvdA and PvsA. The "Organically complexed Fe utilization Guild" is hypothesized to be composed by γ-proteobacteria and Bacteroidetes among others

WP8 is organized in the same tasks as those already explained for WP4, with the difference that in WP5 the traits under study will N, P and Fe sources. Briefly the tasks in WP5 will be:

Task 1: Review and analysis of the previous literature on different N, P, Fe sources for marine prokaryotes.

Deliverable: a list with possible compounds used by marine prokaryotes to obtain N, P and Fe.

Task 2: Identification of marker genes for guilds associated with the different metabolic pathways utilized by prokaryotes to utilize each of the N-, P- and Fe-containing compounds in the list from Task 5.1.

Deliverable: a table in which each compound from the list obtained in Task 1 corresponds with one or several marker genes defining the different guilds on N-, P- and Fe-sources.

Task 3: Search for different guilds using different N-, P- and Fe-sources in different available datasets

Deliverable: a list containing the N-, P- and Fe-guilds and their taxonomic composition present in each of the samples analyzed.

Task 4 Relate the taxonomic composition of the different guilds using different N-, P- and Fe -sources with the characteristics of different ecological settings.

Deliverable: discussion on the relationship between the presence and composition of different N-, P- and Fe-guilds and the different ecological settings

REFERENCES:

An * indicates the paper includes one of the PIs as author.

*Akram et al. 2013. Environmental Microbiology 15: 1400-1415.
Alonso-Sáez, L 2009. Journal of plankton research, 31(11), 1373-1383.
Alonso-Sáez, L. 2012. Limnology and Oceanography, 57(3), 798-808.
Alonso-Sáez, L., & Gasol, J. M. 2007. Appl. Environ. Microbiol., 73(11), 3528-3535.
Alonso-Sáez, L., 2007. Aquatic Microbial Ecology, 46(1), 43-53.
Baltar, F. 2016. Frontiers in Microbiology, 7, 1670.
Béjà et al. 2000. Science 289: 1902-1906.
Béjà. 2002. Nature 415:630-633.
Blondel J. Oikos. 2003;100(2):223-231.
Boiteau, R. M., 2016. Proc Nat Acad Sci USA, 113(50), 14237-14242.
*Bunse, C. et al 2016. Nature Climate Change, 6(5), 483.
*Casamayor, E.O., et al. 2000. Appl. Environ. Microbiol. 66(2): 499-508.
Cho and Salyers. 2001. J. Bacteriol. 183: 7224-7230.
*Cobo-Simón, M., Tamames, J. 2017. BMC Genomics, 18 (1), art. no. 499.
Copley. 2002. All at sea. Nature 415:572-574.
*Díez-Vives, C. et al. 2014. Systematic and Applied Microbiology 37: 68-78.
*Díez-Vives, C. et al. 2019. Molecular Ecology. 00:1–14. doi.org/10.1111/mec.15068
Fenchel. 2001. Marine bugs and carbon flow. Science 292: 2444-2445.
*Fernández-Gómez, B. et al. 2013. The ISME journal, 7(5), 1026.

Ferrera et al. 2017. ISME J. 11:2391-2393.
Ferrera, I. and O. Sánchez 2015. Current Opinion in Microbiology, 25, 33-39.
Fuhrman et al. 2008. Nature Rev. Microbiol. 6:488-494.
*García-García N, et al. 2019. ISME J. doi:10.1038/s41396-019-0487-8.
Gerlt et al. 2015. Biochim Biophys Acta. 1854:1019-1037.
Giovannoni, S. J 2005. Science, 309(5738), 1242-1245.
Goericke. 2002. Limnol. Oceanogr. 47:290-295.
*Gómez-Consarnau et al. 2007. Nature 445: 110-213.
*Gómez-Consarnau et al. 2010. PLoS Biol. 8: e1000358.
*Gómez-Consarnau et al. 2016. ISME J. 10: 1102-1112.
Gómez-Consarnau et al. 2019. Science Advances 5: eaaw8855.
Gómez-Consarnau, L. et al 2012. Environmental Microbiology, 14(9), 2361-2378.
*González et al. 2008. Proc Nat Acad Sci USA 105: 8724-8729.
*González et al. 2011. Appl. Environ. Microbiol., 77(24), 8676-8686.
*González et al. 2019. ISME Journal, doi.org/10.1038/s41396-019-0347-6
Haggerty, J. M., & Dinsdale, E. A. 2017. Global Ecology and Biogeography, 26(2), 177-190.
Hansman et al. 2015. Mar. Chem. 177: 288-297.
Hertkorn et al. 2013. Biogeosciences 10: 1583-1624.
Hopkinson, B. M., & Barbeau, K. A. 2012. Environmental Microbiology, 14(1), 114-128.
*Jiménez et al. (in preparation).
Johnson SS, et al. 2017. Real-Time J Biomol Tech. Apr;28(1):2-7
Karl. 2002. Nature 415:590-591.
Khersonsky and Tawfik. 2010. Annual Review of Biochemistry 79:471-505.
Kirchman and Hanson. 2013. Environ Microbiol Rep. 5: 188-199.
Koblížek et al. 2007. Environmental Microbiology 9:2401-2406.
Koch BAE, et al. 2019. Acta Biol. Colomb., 24(2):224-231
Kolber et al. 2000. Nature 407:177-179.
Kolber et al. 2001. Science 292:2492-2495.
Lei et al. 2018. Biochemistry 57:3364-3377.
Malmstrom et al. 2007. Aquatic Microbial Ecology 47:45-55.
Martinez et al. 2007. Proc Nat Acad Sci USA 104: 5590-5595.
*Martínez-García, S. 2010. Marine Ecology Progress Series, 416, 17-33.
*Martínez-García, S. 2012. Marine Ecology Progress Series, 462, 9-19.
*Martínez-García, S. 2015. Estuarine, Coastal and Shelf Science, 153, 18-28.
McCarren et al. 2010. PNAS 107: 16420-16427.
Mestre, M., et al. The ISME journal, 11(4), 999.
Mitchell et al. 2009. Nature 460:220-224.
Moore, C. M. 2013. Nature Geoscience, 6(9), 701.
Morris et al. 2010. ISME J. 4: 673-685.
Nedashkovskaya et al. 2013. Current Microbiology 66:16-21.
Newton, R. J. 2010. The ISME Journal, 4(6), 784.
*Palovaara et al. 2014. PNAS 111: E3650-E3658.
*Pedrós-Alió, C. 1989. Chapter 8, En U. Sommer (ed.) Plankton Ecology:Succession in Plankton Communities, Science Tech./Springer Verlag, pp. 297-335.
*Pignatelli M, Moya A, Tamames J. 2009. Environ Microbiol Rep., Jun;1(3):191-7.
*Pinhassi et al. 2016. Microbiol. Mol. Biol. Rev. 80: 929-954.
*Pinhassi, J. 2004. Appl. Environ. Microbiol., 70(11), 6753-6766.
Pommier et al. 2005. Aquatic Microbial Ecology 41:79-89.
Prieto, A. 2015. Journal of Plankton Research, 38(1), 55-63.
*Puente-Sánchez et al. (in preparation).
Pushkarev et al., 2018. Nature 558:595-599.
Quick J, et al. 2016. Nature, 530(7589):228-232.
Quince C, et al. 2017. Genome Biol. 18(1):181. doi: 10.1186/s13059-017-1309-9.
Reintjes et al. 2017. ISME J. 11:1640-1650.
Riemann, L. 2010. Aquatic Microbial Ecology, 61(3), 235-247.
Root, R. B. 1967. Ecol. Monogr. 37: 317–350
*Royo-Llonch, M., et al. 2017. Frontiers in Microbiology, 8: article 1317: 1-14.
Sarmento, H. 2016. The ISME journal, 10(11), 2582.
Satinsky, B. M. 2014. Proc Nat Acad Sci USA, 111(30), 11085-11090.

Sebastian, M., & Ammerman, J. W. (2009). The ISME journal, 3(5), 563.
Shannon et al. 2003. Genome Res. 2003;13:2498-2504.
Simberloff, D. and Dayan, T. 1991. Annu. Rev. Ecol. Syst. 22: 115–143.
Sun, J. (2011). PloS one, 6(8), e23973.
Sunagawa, S. et al. (2015). Science 348 (6237).
*Tamames J, et al. 2019. BMC Genomics 2019, in press.
*Tamames J, Puente-Sánchez F. 2019. Front Microbiol. 9:3349.
Teeling, H. et al. 2012. Science, 336(6081), 608-611.
*Teira et al. 2019. Environmental Microbiology. In press.
*Teira, E. et al 2011. Marine Ecology Progress Series, 426, 87-104.
*Teira, E. et al 2016. Marine Ecology Progress Series, 542, 39-50.
Thézé J, et al. 2018. Cell Host Microbe. 23(6):855-864.e7.
Toulza, E., 2012. PLoS One, 7(2), e30931.
Van Gemerden, H. 1974. Microbial Ecology, Vol. 1, 104-119
Varela, M. M. (2006).. Microbial Ecology, 51(4), 487-500.
Venter et al. 2004. Science 304: 66-74.
Vila-Costa, M., et al. (2019). Scientific Reports, 9(1), 233.
Zark et al. 2017. Mar. Chem. 191: 9-15.
Zehr, J. P., & Ward, B. B. (2002). Appl. Environ. Microbiol., 68(3), 1015-1024.

C.2.4. Infrastructure and equipment available and justification of the requested personnel.

a) Subproject 1.

At the Microbiome Analysis Laboratory, CNB-CSIC we have 4 work stations HP Pavillion Elite, 8 to 16 Gb RAM each, Intel i5/i7; one NRENDER R08-E5 server with 24 cores and 256 Gb RAM; a backup and storage system Synology DS414j

b) Subproject 2.

We have a working station (Satélite i7, 4 GHZ, 32 GB, 2xSSD, 480 GB) in our department; one Dell PowerEdge R710, 16 GB RAM, with up to 16 processing units; and one Dell PowerEdge R540, 32 GB RAM, with up to 40 processing units. External storage capacity is also available

c) Subproject 3

GOB is currently formed by 1 Professor of Ecology, 4 Senior Lecturers, 3 postdoctoral researchers, 9 post-graduate students and 2 technicians. The group has the infrastructure needed for the development of the activities considered in TRAITS. The GOB laboratory at the University of Vigo is fully equipped to carry out measurements of chlorophyll a concentration, primary production, bacterial production, and DNA and RNA filtration. We have several vacuum pumps and filtration systems, 2 peristaltic pumps and filtration units for DNA and RNA filtration, a hybridization oven, 2 thermocyclers, 1 epifluorescence microscope, 1 inverted microscope, 2 fluorometers, 1 FRR fluorometer, 2 refrigerated microcentrifuges, 1 microcentrifuge for DNA and RNA extraction, 2 UVR air-clean hoods, an HPLC chromatograph with UV-vis detector system, a cool dark room, several fridges, three -20ºC freezers and one -80 º freezer. GOB also has access to the large outdoors meso-/microcosms infrastructure, the R/V Kraken and the subaquatic unit all of them at the Marine Biological Station of the University of Vigo (ECIMAT)

C,2.5. Chronogram

Quarters	1			4				8				12
WP1. MinIon in-situ sequencing												
Task 1. Setting up the individual components of the pipeline	■	■	■	■								
Task 2. Sequencing of synthetic microbiomes					■	■						
Task 3. 1 day sequencing of an environmental metagenome							■					
Task 4. 2-days cruise sequencing of marine samples								■				
WP2. Microdiversity												
Task 1. Microdiversity based on amplicons, REMEDIOS	■	■										
Task 2. Methods for reconstructing 16S rRNA from miTags			■	■	■	■						
Task 3. Microdivesity based on reconstructed 16S rRNAs							■	■				
Task 4. Resolution of the gene content of the relevant strains									■	■	■	■
WP3. General Analysis of Guilds												
Task 1. Dynamics within guilds					■	■	■	■	■	■	■	■
Task 2. Similarities between guilds					■	■	■	■	■	■	■	■
Task 3. Similarity between environmental conditions					■	■	■	■	■	■	■	■
WP4. Methodological Approach												
Task 1. Construction of database of key marker genes	■	■	■	■	■	■	■	■				
Task 2. Metagenome analysis, automatic annotation of genes	■	■	■	■	■	■						
Task 3. Binning of the metagenomes					■	■	■	■				
WP5. TonB dependent transporters												
Task 1: Determination of TBDTs in representative sequences	■	■	■	■	■	■	■	■				
Task 2: Analysis of level of expression of TBDT families					■	■	■	■	■	■		
Task 3: Diversity of TBDT families in the natural									■	■	■	■
Task 4: Analysis of expression						■	■	■	■			
WP6. Photoheterotrophic guilds												
Task 1. Construction of peptide databases	■	■	■	■	■	■						
Task 2. Analysis of gene expression						■	■	■	■	■	■	■
WP7. Identification of C-sources guilds												
Task 1. Production of a list of C-containing compounds	■	■	■	■								
Task 2. Perturbation experiments and sequencing of samples	■	■	■									
Task 3. Identification of marker genes for C-guilds	■	■	■	■	■							
Task 4. Search for C-guilds in available datasets			■	■	■	■	■	■	■			
Task 5. C-guilds in different ecological settings					■	■	■	■	■	■	■	■
WP8. N-, P- and Fe-sources guilds												
Task 1. List of N-, P-, and Fe-containing compounds	■	■	■	■								
Task 2. Identification of marker genes for N-, P-, and Fe-	■	■	■	■	■							
Task 3. Search for N-, P-, and Fe-guilds in available datasets			■	■	■	■	■	■				
Task 4. N-, P-, and Fe-guilds in different settings					■	■	■	■	■	■	■	■
Coordination meetings	■		■		■		■		■		■	
Dissemination of results							■	■	■	■	■	■

3. IMPACTO ESPERADO DE LOS RESULTADOS - *EXPECTED RESULTS IMPACT*

C.3.1. Scientific-technical, social and/or economic impacts

The objectives defined in TRAITS tackle scientific questions at the frontier of knowledge in marine microbial ecology. Bacterial activities and adaptations to the environment have been explored in the last century. Yet, renewed interest among the research community can be appreciated from rapidly accumulating sound publications over the last 30 years. To gain understanding on bacterioplankton and its role on the cycling of elements, "omic" technologies are essential to improve our understanding of bacterioplankton ecology and evolution in the changing environment. TRAITS will improve our ability to predict how marine microbes will respond to future oceanic scenarios, thereby better constraining and modelling oceanic nutrient fluxes and biogeochemical cycles. We expect that TRAITS will generate cutting-edge results that will be published in high-impact, specialized international scientific journals. We also expect to publish the most relevant results in top generalistic scientific journals.

We hope that the results achieved here will be similar to those obtained in previous projects (see "spreading plan"). In the past we have published substantially, sometimes in high impact journals, and we have a respectable international visibility (see CVs). Our expectation is to have at least a similar impact in the scientific fields touched by the project. Our best contribution to society is the training of first class technicians and PhD students that will be available in the job market (let's hope that it gets better). Finally, we actively participate in many outreach activities for the general public and we act as advisers on scientific issues to many TV, radio, internet and printed media, an extremely important function of scientists and a very effective way of transferring knowledge to the general public.

C.3.2. Spreading plan

Our spreading plan is based on four legs: scientific, outreach, teaching, and business.
a) Scientific spreading plan is based on scientific publications and presentations at meetings.
The 4 PIs have published a substantial amount papers in the last few years, an average above 10 papers per year. We have published in high impact multidisciplinary journals such as Nature, Science or PNAS. Moreover, we have published several review papers in the most important journals in our fields, such as Annual Review of Marine Science, Trends in Microbiology, Nature Reviews in Microbiology, and Current Opinion in Microbiology. Members of the team have made presentations at many international symposia during the last ten years and, in many cases they have been asked to give plenary or invited lectures such as at the International Society for Microbial Ecology (Brazil 1995, Mexico 2004, Washington 2010, Copenhagen 2012, Korea 2014), International Union of Microbiological Societies (IUMS, Montreal, Canada 2014), American Society for Microbiology (Salt Lake City, Utah, 2002, Toronto, Canada, 2007), Sociedad Chilena de Microbiología, Sociedad Española de Microbiología, Spanish Society for Biotechnology (BIOTECH, Madrid, Spain 2014), Federation of European Microbiology Societies (Valencia, Spain 2017), Gordon Research Conferences (Maine, USA 2010, Ventura, California, USA 2011, Il Giocco, Italy 2012, Girona, Spain 2016). Since our record has been sustained at a similar level for over a decade, we see no reason why it should not be comparable during the next few years.
b) Outreach plan. We think that outreach is an essential part of our activity, perhaps as important as the science itself. Increasing the scientific awareness of the general public is a necessary investment on the wellness of society. In addition, it has a payback to development of the scientific activity, since an educated society will put science in value and will demand more scientific progress. With this in mind, we want to make a strong effort to communicate our motivations and results to the general public. We will participate actively in different outreach initiatives, such as science fairs like "Semana por la Ciencia" and "La noche de los investigadores". We have taken part in Radio and TV programs related to science, and will continue to do so. We will take advantage of the digital media to increase the visibility of our research. One example is the interview to JT in Tendencias21 web site

(http://www.tendencias21.net/La-biologia-de-sistemas-nueva-frontera_a41507.html). We will use several digital platforms to inform of our project. The webpage of our group (**http://microbiomecnb.com/**) and the CNB webpage (http://www.cnb.csic.es/index.php/es/investigacion/departamentos-de-investigacion/biologia-desistemas/modelado-de-comunidades-microbianas) will be updated regularly with the results of our research. Also we will work closely with the Oficina de Comunicación at CNB to produce regular press reports.

To reach a wider audience, we plan to use social networks. Especially Facebook and Twitter are excellent resources for increasing the visibility of our project. Through the follow-up of our colleagues we expect to reach a wider dissemination between their connections. Also, we plan to regularly contribute to well-know scientific blogs such as the CSIC blog “ciencia para llevar” (http://blogs.20minutos.es/ciencia-para-llevar-csic), “la ciencia y sus demonios” (https://lacienciaysusdemonios.com/), and perhaps the renowned blog “Small Things Considered” (http://schaechter.asmblog.org), since in the past we have been in contact with its author Dr. Moselio Schaechter. We also regularly write papers for outreach magazines. The outreach books "La Vida al Límite" and “Bajo la piel del océano”, by CP-A, have been published by CSIC-La Catarata and Plataforma Editorial respectively.

c) Teaching. Two of us (JMG and SMG) are University professors and have their regular teaching responsibilities, including general Microbiology and Marine Microbiology. We also teach graduate courses (both national and international, such as the prestigious Auguron Institute courses on Microbial Oceanography at the U. Hawaii, and ECODIM in Chile). Also, one of us (JT) is regularly teaching several courses on metagenomics and high throughput sequencing analysis, especially focused on microbial ecology applications.

d) Business. As explained in the following section our research approach is basic, but we engage companies in our projects and actively carry out contacts through the Technology Transfer Office at CNB (http://www.cnb.csic.es/index.php/es/transferencia-de-tecnologia).

C.3.3. Transfer of research results.
This is a basic science project and there is no foreseeable immediate transfer or commercial exploitation of our results. However, we have several entities interested in our project. In particular we have projects with Globachem, a Belgian company and with Total, the French oil company for different developments

4. CAPACIDAD FORMATIVA - *TRAINING CAPACITY*

C.4.1. Planned training plan
CNB is one of the largest CSIC institutes and a “Severo Ochoa center of excellence” since 2015. The facilities are state-of-the-art and the students are exposed to a wide range of scientists and fellow students in different areas of biology. The number of seminars, for example, runs close to 200 per year. CNB has created a Training Advisory Committee, whose role is to assist the CNB Director’s team in the design, implementation and follow up of training of Ph.D. students. Its members will serve as first-line contact persons for PhD students who need assistance. In fact, one of us (JT) is a member of this committee (http://www.cnb.csic.es/index.php/es/empleo-formacion/trainingadvisory- comitee). Students, for example, receive a training course on laboratory protocols, safety issues and so on. Other activities for students include a welcome event, a predoctoral scientific workshop and workshops on career orientation and writing of scientific papers. Students enroll in one of the Ph. D. programs at the Autonomous University of Madrid (UAM), one of the four best ranking Spanish universities. Some programs are Molecular Biosciences, Microbiology or Ecology. CNB is conveniently located within the UAM campus, which facilitates interactions with different faculties. We encourage students to

participate in seminars and attend scientific meetings. We also have a biweekly meeting where they have to present their work and their plans in English. This constant attention is a guarantee that no time will be wasted in cul-de-sac experiments and that they will acquire the abilities to present their work orally.

In the University of Vigo, the candidate will be trained in the GOB group (currently formed by 1 Professor of Ecology, 4 Senior Lecturers, 3 postdoctoral researchers, 9 post-graduate students and 2 technicians) and the DO*MAR doctorate program (http://campusdomar.es/en/doctorado/). This program was created under the umbrella of the Campus of International Excellence "Campus do Mar" the "DO*MAR" program The "Marine Science, Technology and Management Doctorate Program (DO*MAR), is an international, integrated, interdisciplinary programme of excellence where all topics related to marine and maritime issues are subjected to be addressed from a high-quality research approach.

C.4.2. Ph.D. thesis completed or in progress (last 10 years)

Natalia García. Second year. Doctorado en Microbiología. UAM. Directed by Javier Tamames.

Diego Jiménez. Second year. Doctorado en Microbiología. UAM. Directed by Javier Tamames.

Marta Cobo Simón. Fourth year. Doctorado en Biociencias Moleculares, U. Autónoma de Madrid. Directed by Javier Tamames.

Marta Royo Llonch. Fourth year. Doctorado en Microbiología, U. Autònoma de Barcelona. Codirected by C. Pedrós-Alió and Silvia G. Acinas (Instituto de Ciencias del Mar, CSIC).

Roy Mackenzie Calderón. 2014. Doctorado en Microbiología, U. Autònoma de Barcelona. Co-directed by C. Pedrós-Alió and Beatriz Díez (Pontificia Universidad Católica de Chile).

Montserrat Coll Lladó. 2013. Doctora en Ciencias del Mar, U. Politècnica de Catalunya. Co-directed by C. Pedrós-Alió and Silvia G. Acinas (Instituto de Ciencias del Mar, CSIC).

Beatriz Fernández Gómez. 2012. Mayo. Doctorada en Ciencias del Mar, U. de Las Palmas de Gran Canaria. Co-directed by C. Pedrós-Alió and José M. González.

Lorena Escudero. 2009. Doctora en Microbiología, U. Autònoma de Barcelona. Co-directed by C. Pedrós-Alió and Cecilia Demergasso (Universidad Católica del Norte, Chile).

Erick Delgadillo. First year. Doctorado en Ciencias del Mar. U. de Vigo. Codirected by: S. Martínez-García and E. Teira.

Cecilia Costas. First year. Doctorado en Ciencias del Mar. U. de Vigo. Codirected by: S. Martínez-García and E. Teira

Clara Pérez-Martínez. Third year. Doctorado en Microbiología. Linnaeus University. Codirected by: Jarone Pinhassi and S. Martínez-García.

Benjamin Pontiller. Third year. Doctorado en Microbiología. Linnaeus University. Codirected by: Jarone Pinhassi and S. Martínez-García.

C.4.3. Brief description of the professional or scientific career of the past students.

Out of the eight PhD students who defended their theses before 2002, five have permanent research positions (three at ICM and two at other institutions) and three went into teaching at High Schools. Furthermore, the four most senior ones have become independent and well known scientists in the field of microbial ecology. Of the four most recent graduates, one has got a permanent position as senior researcher at the Universidad Católica del Norte in Antofagasta, Chile; two have postdoctoral positions at the Pontificia Universidad Católica in Chile and the fourth is looking for a postdoc.