Internship final term report

Genetic diversity limits of a marine microbiome

Student: Sergio Gozalo Miranda

Host institution: CSIC-ICM

Host institution tutor: Ramiro Logares Haurie (ramiro.logares@icm.csic.es)

Start/End: 06/01/2021-07/04/2021.

Abstract

Microbial marine communities are made up of thousands of organisms that contribute with genes that are fundamental for the functionality of the ecosystem. Using marine metagenomics, we will face the challenge of finding the total diversity of genes of a marine community. To solve this challenge and approach this marine genome unknown diversity, this study is going to analyse 50 different metagenomes from the same place obtained within 2 days and, in consequence, try to reach the limits of the marine microbial genetic diversity in this location.

Sample obtention was performed in a marine observatory from the Mediterranean Sea (SOLA) in one day with different protocols, using different filters and water volumes and filter size-fractions. A total of 50 metagenomes were obtained that were analysed with two different bioinformatics pipelines, MGnify (https://www. https://www.ebi.ac.uk/metagenomics/pipelines/2.0) and a custom pipeline from ICM. The results from the different pipelines will be analysed using R software (R- Development-Core-Team 2008) and the package vegan (Oksanen et al. 2008). These results will be used to try to find the best pipeline to assess the genetic diversity and, in consequence, make one of the pipelines the main for the analysis of other marine metagenomes, providing, this way, a standard method. The total number of genes, functions, and taxonomic markers (rRNA-gene) will be analysed to determine whether we have reached the limits of microbial diversity.

Since this project is a basic science project, the immediate contributions are not clear, however, this characteristic also implies that there may be many contributions in the near future, for example, new genes with utility in biotechnology(bioprospection) may be found.

Introduction

This project is the result of the continuation of several marine microbiology projects. One of the first of these projects began with the assumption that the marine biosphere could be formed by two taxa, one oxygenic photoautotroph (e.g., cyanobacterium Prochlorococcus marinus) and one oxygen- respiring chemoorganoheterotroph (e.g., proteobacterium Cand. Pelagibacter ubique). Those two taxa, in theory, could be able to carry on the carbon and nutrient cycles and energy flow. But that is clearly not true, “living beings belong to a multitude of different taxa that interact with each other forming an intricate network” (Sherr and Sherr 1988). Moreover, “the resultant network is woven by all the possible relationships among microorganisms: mutualism, commensalism, parasitism, competition, and predation, generating what has been named: the ecological interactome” (Lima-Mendez et al 2015).

With that information, one will ask himself what mechanisms generate that necessary diversity. If we think about diversity in an evolutionary context, there are two types of constraining mechanisms: the ones that promote evolutionary diversification, that lead to the generation of new taxa, and those pressing for equilibrium that select the most adapted taxa to the ecological setting. The study of those mechanisms has always been challenging when it comes to microbial communities, but the development of high-throughput sequencing (HTS) technologies allowed the study of diversity become easier. From that point, mechanisms promoting the appearance of novel taxa were demonstrated (Rosenzweig et al. 1994) and mechanisms promoting coexistence (Stomp et al. 2004) were also demonstrated.

With the application of HTS for diversity studies in the marine environment a huge improvement was made sequencing up to 107-109 sequences per run, with the possibility of including multiple samples, compared to the Sanger runs that would generate 102 sequences. Also, HTS products did not require PCR amplifications, whereas Sanger products did require it.

And so, the first study of marine microbial diversity using HTS was carried out during the International Census of Marine Microbes 2005-2010. Analysing around 500 samples, the main observations were that the richness was very large and that most of those taxa were found in low abundance. Also, thanks to the contribution of HTS providing many sequences, the definition of microbial species has changed from coherent genomic groups with a DNA-DNA hybridization (DDH) values greater than 70% similarity (Wayne et al. 1987) to use the average nucleotide identity (ANI; Konstantinidis and Tiedje 2005), which several bioinformatic tools such as JSpecies uses. Another approach is to use genome sequences with the microbial species identifier (MiSi; Varghese et al 2015), combining the genome-wide average nucleotide

identity (gANI) with the fraction of orthologous genes as a measure of relatedness.

However, all these approaches were limited to cultured isolated microorganisms, to solve that a new concept for bacterial species was needed to accommodate two characteristics: most of the marine bacterial taxa remains uncultured and the microdiversity within species must be considered. There are two possible solutions that can carry these conditions. Firstly, is the pan-genome approach, the pan-genome includes a core genome containing genes present in all strains of a given species, combined with ANI as a proxy for species (Rodriguez-Valera et al. 2016), this method combined with single-cell genomes (SAGs) can be applied to uncultured microbial taxa. On the other hand, there is the metagenomic assembled genome (MAG; Hugerth et al. 2015), based on binning the contigs derived from the co-assembly of multiple metagenomic samples.

After these approaches were studied, the diversity of the marine microbial communities could finally be narrowed using accumulation curves, for one sample, the relative abundance of taxa and the rarefaction, that compared the diversity of different samples. Using these techniques in combination with samples collected from different circumnavigation initiatives, such as Tara Oceans (2009-2013) and Malaspina (2010-2011) allowed to test and reach the limits of the microbial marine communities, that can vary, as expected, depending on factors like deepness and temperature.

Even with all these techniques there are still blind spots, unknown microbes, named “microbial dark matter” (Marcy et al. 2007), and can be explained by three reasons: 1) The universal considered primers may not hybridize with the rRNA of those microorganisms, this can be solved with metagenomics. 2) Those microbes are so rare that they do not appear in surveys, this can be partially solved by increasing the sequencing effort or studying the diversity under different conditions. 3) We have the ribosomal RNA sequence, but we ignore the responsible microorganism, this problem can be approached by analysing SAGs or reconstructing MAGs.

Even though there is still work to do to construct the list of taxa with abundances of marine microbial communities, it can be said that the taxonomic diversity values have been narrowed down enough to consider this task done, until new technologies allow better studies. So, naturally, with all the information gathered, sequences, the next challenge was raised, finding the total diversity of genes of a marine microbial community. This has been studied but is less known and depends on the techniques used. So, the aim of this project is to help in solving that.

To solve this challenge and approach this marine genome unknown diversity, this study is going to analyse 50 different metagenomes from the same place

obtained within 2 days and, in consequence, try to reach the limits of the marine microbial genetic diversity in this location.

Material and methods

This study will analyse samples obtained in a marine observatory from the Mediterranean Sea (SOLA) in one day. Those samples were collected following different protocols, a combination of filters and water volumes (See Fig 1). The filters used were: >0.2 µm. (Sterivex), >0.2 µm. (142 mm, membrane), 0.22-3 µm. (142 mm, membrane), 3-20 µm. (142 mm, membrane) and >20 µm. (47 mm, membrane). These different filters are used with the objective of enrich different sizes of bacteria and compare Sterivex filters and membrane filters. The water volumes were: 1L, 10L, 100L, 500L and 1000L, the 100L volume was fractioned into samples of 20L, the volumes of 500L and 1000L were fractioned into samples of 100L. This different volumes and fractions had the objective of determining if the volume sampled was a factor to consider for the final experiment, determine the best volume to study prokaryotes or eukaryotes, study saturation and the loss of information due to size-fractionation.

There were two sets of MAGs generated with the raw reads obtained from the samples, one generated using the pipeline MGnify and a custom pipeline created by the CSIC-ICM team. The usage of two pipelines was encouraged with the objective of finding the one that best adapts to the marine metagenomics.

The MGnify pipeline uses SeqPrep to transform the raw reads into the initial reads, the pair-end overlapping reads are merged, then uses Trimmomatic to discard low quality sequences and Biopython to discard short sequences, next using rRNASelector the Prokaryotic rRNA reads are filtered using HMM to identify rRNA sequences, this program outputs two types of reads: on the one hand the ones with rRNA masked that are input to FragGeneScan to predict CDS that will be matched using InterProScan against a subset of databases to obtain a summary of Gene Ontology (GO) and produce a functional analysis. On the other hand, the reads with the rRNA are analysed to detect the 16S rRNA and then using QIIME they are annotated to obtain a taxonomic analysis.

The ICM pipeline generates an assembly using MegaHit, with a digital normalization with the MegaHit co-assembly, then a mapping of the reads with a co-assembly using BWA and Samtools to obtain the indexed sorted bam files needed for binning that is performed in steps with MetaBAT, MaxBin2

and CONCOCT to finally refine it with MetaWRAP, with CheckM implemented to asses contamination and completeness for each bin, next is a final check of the bins running CheckM SSU analysis and finally a taxonomy assignation with GTDBTk.

This project requires huge amounts of statistical analyses and data transformations, hence, we decided to use R software, that can cover these two needs. So far, with the base R and the package vegan we were able to perform all the analysis, that go from the initial exploratory analysis (NMDS, dendrograms, comparisons, etc.) to representations of some results. To reach the main objective of this project we need to implement two more methods: 1) alpha diversity 2) beta and gamma diversity. Alpha diversity that generates the accumulation curve, giving information about how many, in this case, genes are found per reads, so it gives a saturation estimation at which no more genes will be found. Alpha diversity also generates the evenness, or relative abundance of, in this case, genes, informing about the richness and evenness of the sample. Beta and Gamma diversity are used to compare several samples, which is our case with different MAGs, the method used is the rarefaction, that compares the diversity of several samples.

Main dissertation

I started this project by comparing the results obtained by the two different pipelines, on one side the standard metagenomics pipeline provided by MGnify and on the other side the custom pipeline of the CSIC-ICM team. Since the ICM team pipeline selected only the best MAGs, ≥50% completeness and ≤10% of contamination, it had 75 high quality MAGs, so to perform a good comparison the MAGs from the MGnify pipeline were also reduced to the top 75. After that, the comparison was based on the main characteristics of the MAGs: completeness, contamination and N50. Regarding completeness and contamination there were not significant differences, but the comparison of the N50 had a significant difference, ICM N50 fragments were larger, being the mean 25482 bp while the MGnify ones were 19320 bp average. With this data I reached the conclusion that it was better to use the ICM MAGs, since the N50 is larger without having any penalty on completeness or contamination.

Having the MAGs been selected, the next step was to assess the different datasets annotation consistency, the ICM pipeline annotated the genes using three different databases, COG, Pfam and KEGG. With a simple visual analysis of the Bray-Curtis dissimilarity matrix of the different databases against each other in a plot and a statistical analysis, mantel test, this

consistency of results was studied. The results proof that the annotations were consistent, because in both analysis there was a very high correlation.

Next, we performed a function analysis for each database, the objective of that analysis was to find if there were differences between the samples. The results showed that some samples had different functions proportions (See Fig 2). As it can be seen in the results, the samples that used the sampling protocols S20 and S320, meaning that the sampling process can affect the results, this needs to be confirmed by performing the same analysis for all the databases. The next step following this line would be to relate taxonomy with the functions for all the databases.

Finally, the main justification for this project, find the limits of a marine microbial community, would be initially approached using alpha diversity and beta and gamma diversities and represented with the accumulation curve for each sample and the rarefaction for all the samples. The findings from those approaches will determine if the goal has been accomplished or if we need a different implementation of methods. Nevertheless, this project has already given some results about methodologies.

Conclusion

The conclusion I reached so far is that about the project is that the ICM pipeline gave better MAGs, which is important because they are the basis from where all the analysis come, I think that is because in the ICM they are specialised in marine metagenomics and, therefore, are better on this field than the MGnify pipeline that is generally good for all the metagenomics. I was not surprised to see that some sampling protocols affected the results, on the contrary, it was expected because filters are biased towards some microbes based on the size. Finally, I think that this project will yield results that can inspire new projects and expand the line of work that started many years ago, explained in the introduction.

As a personal conclusion about the CI, I have enjoyed it, learned many things about metagenomics, and experienced the real work of a research laboratory group.

Bibliography

- Gasol, J.M., &Kirchman, D.L. (2018). Microbial Ecology of the Oceans (3rd ed.). Wiley-Blackwell

- Santos Júnior, Celio & Sarmento, Hugo & Miranda, Fernando & Henrique-Silva, Flávio & Logares, Ramiro. (2020). Uncovering the genomic potential of the Amazon River microbiome to degrade rainforest organic matter. Microbiome. 8. 10.1186/s40168-020-00930-w.

- Dupont, Chris & Chappell, Dreux & Logares, Ramiro & Vila-Costa, Maria. (2010). A hitchhiker's guide to the new molecular toolbox for ecologists. 10.4319/ecodas.2010.978-0-9845591-1-4.17.

- Logares, Ramiro & Sunagawa, Shinichi & Salazar, Guillem & Cornejo Castillo, Francisco Miguel & Ferrera, Isabel & Sarmento, Hugo & Hingamp, Pascal & Ogata, Hiroyuki & de Vargas, Colomban & Lima-Mendez, Gipsi & Raes, Jeroen & Poulain, Julie & Jaillon, Olivier & Wincker, Patrick & Kandels-Lewis, Stefanie & Karsenti, Eric & Bork, Peer & Acinas, Silvia. (2013). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. Environmental microbiology. 16. 10.1111/1462-2920.12250.

- Locey, Kenneth & Lennon, Jay. (2015). Scaling laws predict global microbial diversity. 10.7287/PEERJ.PREPRINTS.1451V1.

- Mitchell, Alex & Almeida, Alexandre & Beracochea, Martin & Boland, Miguel & Burgin, Josephine & Cochrane, Guy & Crusoe, Michael & Kale, Varsha & Potter, Simon & Richardson, Lorna & Sakharova, Ekaterina & Scheremetjew, Maxim & Korobeynikov, Anton & Kunyavskaya, Olga & Lapidus, Alla & Finn, Robert. (2019). MGnify: the microbiome analysis resource in 2020. Nucleic acids research. 48. 10.1093/nar/gkz1035.

- Sherr, E. and B. Sherr. 1988. Role of microbes in pelagic food webs: A revised concept. Limnol. Oceanogr. 33: 1225–1227.

- Lima-Mendez, G., K. Faust, N. Henry, et al. 2015. Determinants of community structure in the global plankton interactome. Science 348: 1262073.

- Rosenzweig, R. F., R. R. Sharp, D. S. Treves, and J. Adams. 1994. Microbial evolution in a simple unstructured environment: genetic differentiation in Escherichia coli. Genetics 137: 903–917.

- Stomp, M., J. Huisman, F. de Jongh, et al. 2004. Adaptive divergence in pigment composition promotes phytoplankton biodiversity. Nature 432: 104–107

- Wayne, L., D. J. Brenner, R. R. Colwell, et al. 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int. J. Syst. Bacteriol. 37:463–464.

- Konstantinidis, K. T., and J. M. Tiedje. 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. USA 102: 2567–2572.

- Hugerth, L. W., J. Larsson, J. Alneberg, et al. 2015. Metagenome-assembled genomes uncover: A global brackish microbiome. Genome Biol. 16: 279

- Marcy, Y., C. Ouverney, E. M. Bik, et al. 2007. Dissecting biological “dark matter” with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc. Natl. Acad. Sci. USA. 104: 11889–11894.

- RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/

- Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner (2020). vegan: Community Ecology Package. R package version 2.5-7. https://CRAN.R-project.org/package=vegan

Figures

Protocol label	Size fraction	Filtration volume 1L	Filtration volume 10L	Filtration volume 100L	Filtration volume 500L (pooled 100 L replicates)	Filtration volume 1,000L (pooled 100 L replicates)
S02	>0.2 µm (sterivex) ="whole water"	X	X (4x2.5L)			
S02	>0.2 µm (142 mm) ="whole water"		X			
S023+S320+S20	>0.2 µm (142 mm) ="pooled fractions"		X (0.22+3+20)	X (0.22+3+20)	X (0.22+3+20) (5x100L)	X (0.22+3+20) (10x100L)
S023	0.22-3 µm (142 mm)		X	X	X (5x100L)	X (10x100L)
S320	3-20 µm (142 mm)		X	X	X (5x100L)	X (10x100L)
S20	>20 µm (47 mm)		X	X	X (5x100L)	X (10x100L)

Sterivex and membrane filters: Are the results comparable?

Diversity of rare organisms in whole water (focus on prokaryotes?)
How much volume should we filter?
Does it saturate at 10L?

Diversity of rare organisms in size-fractionated water (focus on eukaryotes?)
How much volume should we filter?
Does it saturate at 100L?

Size-fractionation:
Do we gain or lose something by sequencing size fractions?

Figure 1. SOLA Protocol table

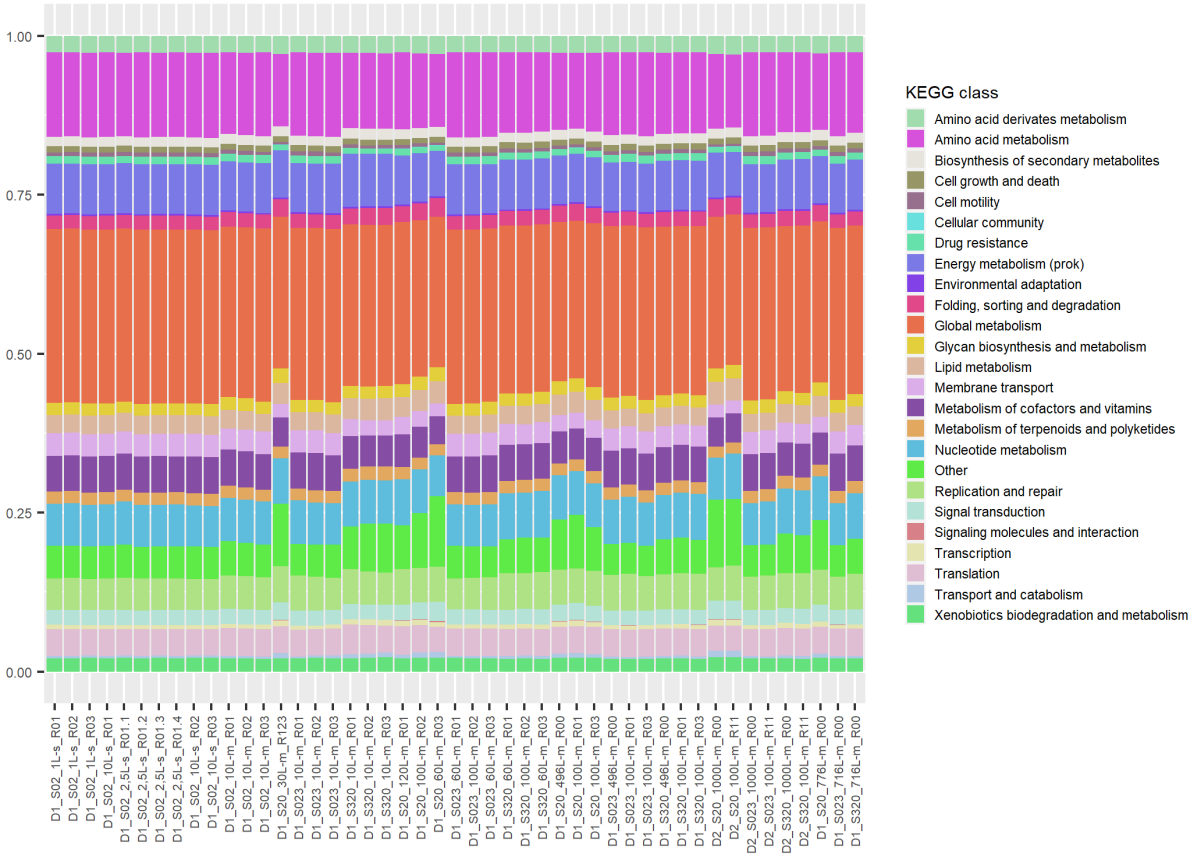

Figure 2. KEGG Function Representation