ESCI upf. School of International Studies

Bachelor's Degree in Bioinformatics (UPF-UPC-UB)

Final Grade Project

Genetic diversity limits of a marine microbiome

Sergio Gozalo Miranda

Scientific director: Ramiro Logarés Haurie

[1] CSIC-ICM

Abstract

Our blue planet is a water-dominated habitat with more than 70 % of its surface covered by the ocean and seas. Microorganisms are omnipresent in the oceans and seas and their short generation times and the nearly 4 billion years of evolution of (marine) microorganisms has resulted in an enormous biodiversity.

In recent years there has been progress in marine bioprospecting and is strongly linked to the development of "omics"-based methodologies.The surge of high-throughput sequencing technologies allowed the marine researchers to go one step forward with the study of microbial marine communities, making possible the taxonomic classification of these communities and the discovery of new species. From these results we know that microbial marine communities are made up of thousands of organisms and that those organisms contribute in an unique way because of their genes, so that genes are fundamental for the functionality of the ecosystem. Following that path, the next step is to analyse those genes, to do that, it is mandatory to find the limits, maximum number of genes that can be found in a microbial marine community.

The aim of this project is to find the genetic diversity limits of a marine microbiome, to reach the objective we will use a total of 50 metagenomes obtained from the same microbiome. The main topic will be approached using statistical methods and, to accomplish the goal, we will also compare how the different sample obtention methods, filters and water volume, can affect the diversity and also if the bioinformatic methods applied to the metagenomes lead to different conclusions.

As results we can expose that there are sampling protocols that can affect the diversity, in specific the filters, also, the most important results, are that the function diversity limit is reached while the genetic diversity limit was not reached with the samples of this study.

Supplementary information: Supplementary data are available at GitHub link: https://github.com/SergioGozalo/Practicas/tree/main/Analysis

1. Introduction

We live on a planet where 70% of the surface is covered by water. The marine ecosystem includes the open waters of the ocean and of the seas, the estuaries and other tidal regions, the seafloor and the sub-seafloor, the polar sea ice masses, and brines. Microorganisms are omnipresent in these marine ecosystems. They exist as single organisms or as communities, planktonic or attached to substrates, and exhibiting different types of interactions among themselves and with their abiotic habitat.

Marine microorganisms have short generation times and combined with the billions of years of evolution the result is an enormous biodiversity. The most interesting diversity is in the metabolic pathways that allow the marine microorganisms to be the exclusive drivers of biogeochemical cycling on Earth.

It is known that a considerable number of marine microorganisms remain uncultured so far, and, hence, their potential remains unknown. However, the development of new "omics"-based technologies and methodologies has set a turning point in marine bioprospecting.

Thanks to this interest in marine bioprospecting, several marine microbial diversity studies were carried out, however, those first studies had approaches to the concept of bacterial species that 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. One of the approaches that addressed that issue was the metagenome assembled genome (MAG; Hugerth et al. 2015), based on binning the contigs derived from the co-assembly of multiple metagenomic samples, the other approach was the pan-genome.

These advances motivated the research on marine microbiomes, so circumnavigation initiatives, such as Tara Oceans (2009-2013) and Malaspina (2010-2011) were possible. Using accumulation curves, relative abundances and rarefactions on the samples, the diversity of the marine microbial communities could finally be narrowed down within a margin that can vary, as expected, depending on factors like depth and temperature.

Most metagenomic surveys of the ocean microbiota typically include samples from several locations. Thus, no location is sampled in depth. To address that, in this study, the samples used are from a previous experiment (Mitchell et al. 2018) where there are 50 metagenomes obtained from the same place in the same day with the objective of uncovering the genetic diversity limits of a microbiome.

Determining the limits of gene diversity in a coastal ecosystem can contribute to bioprospecting.There are previous successes in marine bioprospecting such as the anti-cancer drug trabectedin obtained from *Candidatus Endoecteinascidia frumentensis*, a symbiotic gammaproteobacterium in the sea squirt *Ecteinascidia turbinate*. There are also other examples of marine microbiome derived molecules with high utility as industrial products. there are many scientific articles written that address the biotechnological potential of marine microorganisms (Debnath et al. 2007; Kim 2015; Santos-Gandelman et al. 2014)

Knowing that there has been only a superficial work on marine bioprospecting and there have already been discoveries of great impact, it is only natural to keep the research on this field.

2. Objectives

The research goal is to find the genetic diversity limits of a marine coastal microbiome in the Northwestern Mediterranean sea, station SOLA in Banyuls sur Mer, France, test the different sampling protocols, water volumes and filters sizes and compare the data tables obtained from the MAGs with the different databases. This is all to find the limits of microbial diversity and the best bioinformatic methods to obtain it.

3. Methods and materials

3.1. Sample obtention

The sample obtention was performed on the same day using a high volume well pump, in the context of the EMOSE (2017) Inter-Comparison of Marine Plankton Metagenome Analysis Methods (Mitchell et al. 2018). A total of 50 metagenomes were used in this study, the samples from where these metagenomes are derived followed different protocols. See table 1.

3.1.1. Filtration protocols

The different protocols had as objective the focus on different size organisms, in some filters prokaryotes and in others eukaryotes. There were 5 filters: >0.2µm (sterivex), >0.2µm (membrane 142 mm), 0.22-3µm (membrane 142 mm), 3-20µm (membrane 142 mm) and >20µm (membrane 47 mm).

3.1.2. Water volumes

Different water volumes are used to identify if different amounts of filtered water can lead to different results. There were 5 different water volumes: 1L, 10L fractionated as four samples of 2.5L, 100L fractionated as ten samples of 10L, 500L fractionated as 100L and 1000L fractionated as 100L.

Protocol label	Size fraction	Volume 1L	Volume 10L	Volume 100L	Volume 500L	Volume 1000L
S02	>0.2µm (sterivex)	X	X			
S02	>0.2µm (142 mm)		X			
S023	0.22-3µm (142 mm)		X	X	X	X
S320	3-20µm (142 mm)		X	X	X	X
S20	>20µm (47 mm)		X	X	X	X

Table 1.. Sample obtention protocols. This table displays the different sample obtention protocols combination used.

3.2. *Functional tables obtention*

Metagenomes were individually assembled using Megahit and genes were predicted in contigs using Prodigal & metagenemark. Non-redundant genes were annotated with different databases (e.g. KEGG, Pfam, COG). Gene and functional abundance tables were generated after mapping metagenomic reads back to the predicted genes.Then two normalizations were used, on the one hand MetaGS in which the normalization is based on the sequencing effort per sample, on the other hand there is the SCG (single copy gene) normalization in which the normalization is based on the cellular abundance.

The 3 databases used (KEGG, COG and PFAM) and the two normalizations (MetaGS and SCG) resulted in a total of six table

3.3. Metagenome Assembled Genomes

MAGs construction followed an ICM pipeline that coassembles metagenomes using MegaHit, after digital normalization,.Then reads are backmapped to the co-assembly using BWASamtools is used 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 assess contamination and completeness for each bin. Next there is a final check of the bins running CheckM SSU analysis and finally a taxonomy assignment with GTDBTk.

3.4. Computational analysis

The analysis of the metagenomes was performed using R (R Core Team 2020) and Rstudio (RStudio Team 2020). The packages used are:

3.4.1. Vegan

Vegan (Oksanen et al.2020) is a package that provides tools used for descriptive community ecology. The package vegan was mainly used in the diversity analysis and in the comparison of the different function tables. Within all the tools that this package has, in this experiment we used:

3.4.1.1. MetaMDS

This function performs Nonmetric Multidimensional Scaling to try to find a stable solution using random starting points and also standardizes the scaling in the result. Used in the table comparison of the different databases with the dissimilarity parameter set as "Bray-Curtis" to obtain the best result possible. The resultant plot can be found in the supplementary material.

3.4.1.2. Rrarefy

This function generates a random rarefied community data frame of a given size from the original data. The rarefaction is performed without replacement. In the analysis we used this function to remove the effects that could result from different sizes of the genetic function tables.

3.4.1.3. Specaccum

This function generates Species Accumulation Curves to compare the diversity properties of samples. In the experiment this function has been the central axis since it has been the last step with the "exact" method also known as Mao Tau estimate (Colwell et al. 2012).

3.4.1.4. Specslope

This function evaluates the derivative of the species accumulation curve at a given point and gives the rate of increase in the diversity. This function was used to compute the slope of the accumulation curve.

3.4.1.5. Vegdist

This function computes dissimilarity indices. The resultant dissimilarity matrices were used by the other functions to produce the results.

3.4.2. Recluster

Recluster (Dapporto et al. 2020) is a package that provides tools to analyse different aspects of biodiversity using specific algorithms designed for that purpose. The package recluster has been used with the specific intention of comparing the different genetic function tables. The functions used are:

3.4.2.1. Recluster.cons

This function uses a dissimilarity matrix and resamples the order of sites of the matrix to create a series of trees and compute a consensus among all.

This function has been used to perform the clustering, to determine the similarity, between the functional tables and assess the data consistency.

3.4.2.2. Recluster.boot

This function takes as input a tree and a data matrix to perform bootstrapping. The usage of this function in the analysis has been to reassure the consistency of the data

3.5. *Supplementary materials*

Supplementary materials can be found at: https://github.com/SergioGozalo/Practicas/blob/main/Analysis/Supplementary%20materials/Supplementary%20material/Supplementary_materials.pdf

Code can be found at: https://github.com/SergioGozalo/Practicas/tree/main/Analysis

4. Results and discussion

4.1. *Functional tables consistency*

The consistency of the different functional tables was tested using bioinformatic statistical methods in R. The methods used were a stress plot of the dissimilarity matrix, the dissimilarity matrix sites plot and a clustering method that is represented by a bootstrapped tree.

Regarding the stress plots of the different tables we can observe that all of them showed a linear fit R2 between 0.997 and 1, this means that the stress is very low 0.003, the best solutions usually have low stress and this is the case, which is expected since the dissimilarity will increase with the addition of data. See supplementary material

Next, using the dissimilarity matrix obtained from the function vegsit, site plots were generated and we can see that the results are consistent within the expectations. The SCG tables show two clusters and the MetaGS tables follow a similar pattern between them. A similar plot from the databases is expected, this would mean that the sites displayed generate similar clusters. See supplementary material
Finally, the clustering results show similar trees between the SCG tables and between the MetaGS tables. See supplementary material

With these results we can see that the functional tables are very consistent between and within them, regardless of the normalization or database, so the usage in this experiment is correct.

4.2. *Species diversity*

Using the OTU (Operational Taxonomic Unit) as a definition of species a species diversity plot was constructed using accumulation curves.

The results showed that with 50 samples, the genetic diversity limit cannot be said to be reached since the function specslope had a result of 108, that is not close enough to 0 to say that the curve is flat. See Figure 1.

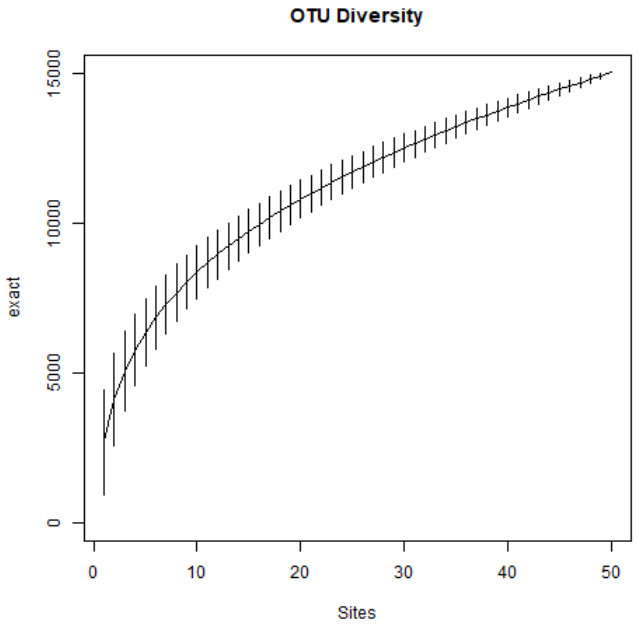

Fig. 1. Species diversity accumulation curve. The number of sites corresponds to the samples, the y axis corresponds to the number of species.

4.3. *Genetic diversity*

Due to limited computational resources, the table used to estimate the genetic diversity was a subsample of the original, containing only 20% of the genes from

the original table. In this case, the plot, as a result of the subsampling, showed a big range of variation, but it can be seen that near the end, the curve is flattened. Again, the limit is not reached but the result is close. See Figure 2.

The number of identified genes was extremely big, in consequence, it is very difficult to reach a limit.

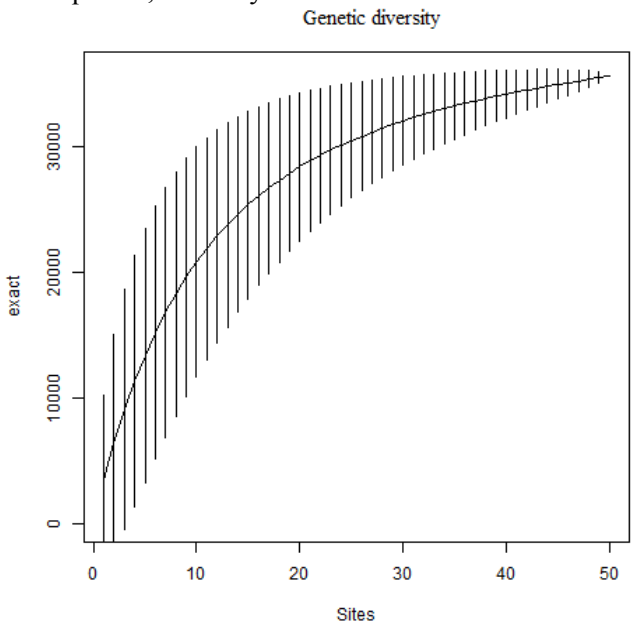

Fig. 2. Genetic diversity accumulation curve. The number of sites corresponds to the samples, the y axis corresponds to the number of genes.

4.4. *Effects of filters*

With the same subsample table used to compute the genetic diversity, we computed 4 different accumulation curves, one for the >0.2µm filters (sterivex and membrane), the next one for the 0.22-3µm filters, another one for the 3-20µm filter and finally the last one for the >20µm filter. The results displayed that the best filters to use are the smaller ones >0.2µm and 0.22-3µm because the slope is smaller, the function specslope returned smaller values for those filters 71 and 309, while the bigger filters had values of 979 and 619, see Figure 3. This can be due to the fact that the smaller filters recover all prokaryotes and that the genetic diversity within prokaryotes is bigger than the diversity within eukaryotes, big filters.

We can say that the diversity is hugely dependent on the filter used, being the smaller filters more effective.

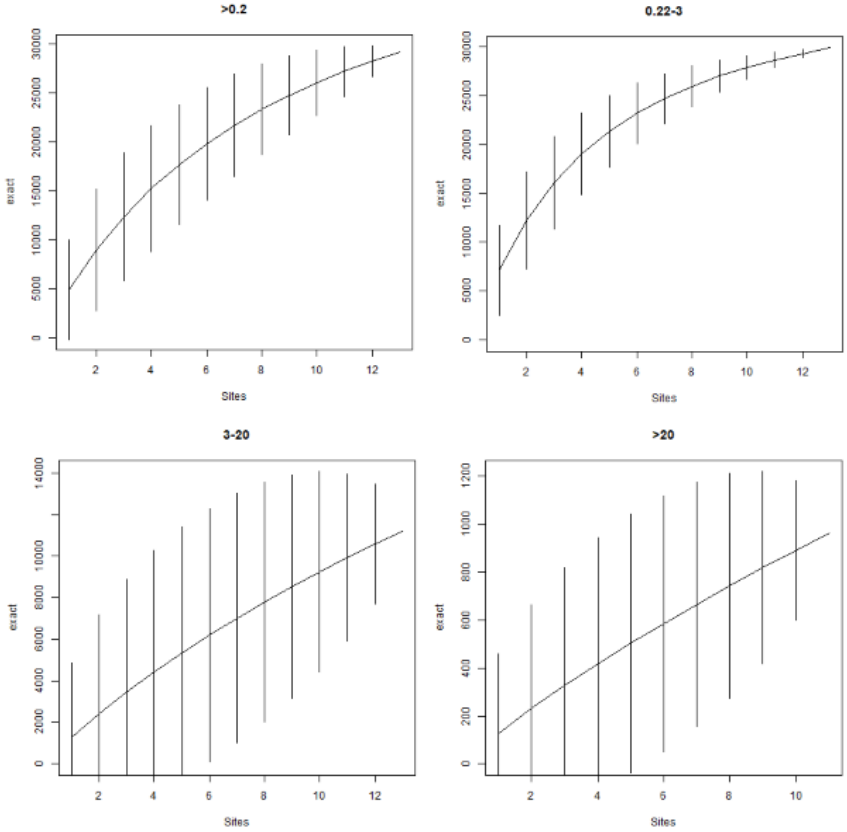

Fig. 3. Filters comparison genetic accumulation curves. The accumulation curves correspond to the filters >0.2µm, 0.22-3µm, 3-20µm and >20µm in that order.

4.5. *Effects of water volumes*

With the same subsample table used to compute the genetic diversity, we computed 3 different accumulation curves, for ten liters, hundred liters and thousand liters, one liter and five hundred liters were ignored due to the lack of samples. The results showed that the volumes of ten and hundred liters were equally effective to capture the genetic diversity, see supplementary material.

Since the available samples with a water volume of thousand liters were less than the other volumes, we decided to run again the accumulation curves of the ten and hundred liters volumes, but this time using only ten samples of those volumes so the comparison would be more fair.. The results showed that the volume of thousand diversity was pretty similar to the other volumes, see supplementary material.

In the case of volumes it can be said that it does not have a major effect in the analysis of the genetic diversity limits.

4.6. *Functional diversity*

It is known that a gene can be involved in different metabolic pathways and, hence, different functions. For this experiment we had different tables, from KEGG, PFAM and COG databases, each one identifying functions using different criteria.

In this case, with accumulation curves it can be seen that with a few samples, the functional diversity limit has been reached. In the case of KEGG, we can see that the limit is reached with 8 samples, regarding COG, it can be said that with one sample the limit has been reached, while with PFAM, we need 3 samples. See Figure 4.

The fact that the functional diversity is easily reached is expected since several genes from different taxa can code for the same function a. In other words, each function contains a plethora of genes from multiple genomes.

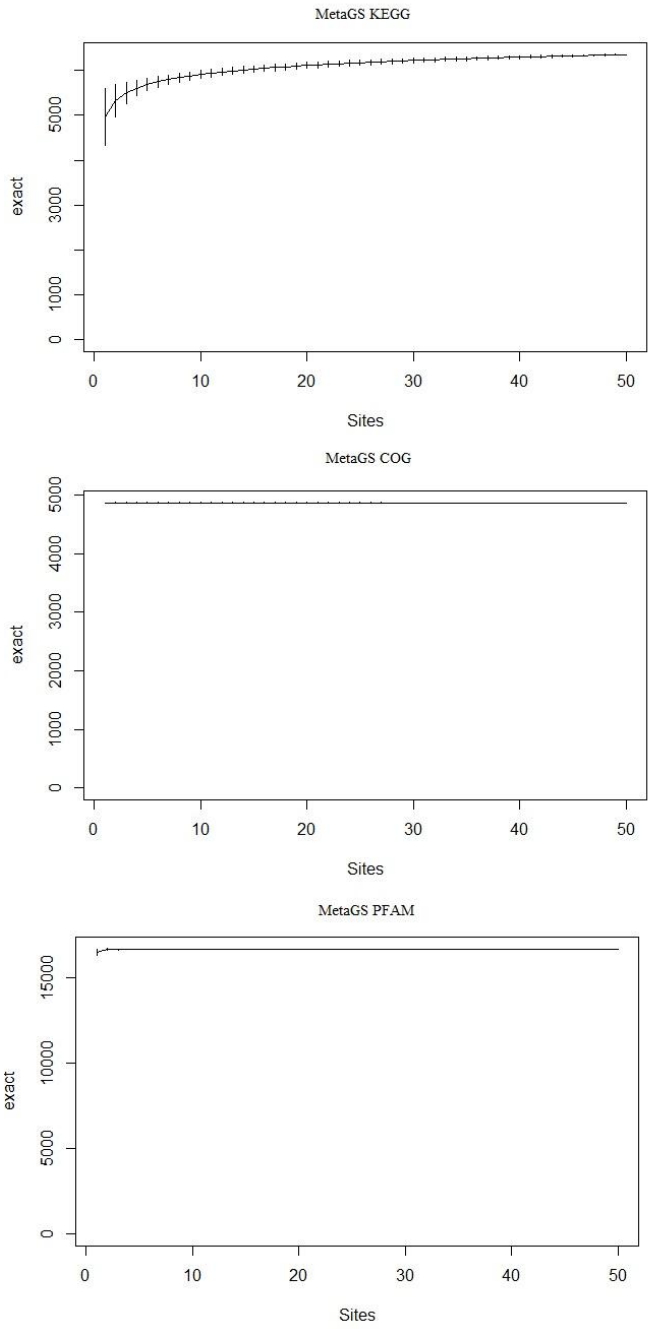

Fig. 4. Functional diversity accumulation curves. The number of sites corresponds to the samples, the y axis corresponds to the number of functions.

4.7. *Comparison between species diversity, genetic diversity and function diversity*

This final comparison of the different diversities studied in this experiment, species, genetic and functional, portrayed a very interesting, yet predecible, result. The bigger the complexity, the harder it is to reach the diversity limit, in this case we

can see that the diversity is easy to reach at the function level, harder at genetic level and hardest at species level. This difference can be explained by the level of complexity of species in respect to genes and functions. See figure 5.

To put some numbers in, when looking at the function level, using KEGG, the diversity limit is easily reached at 8 samples, while when talking about genetic level we cannot say that the limit has been reached since the curve is almost flat, but it looks like it needs a few samples more to reach the limit, this can be due to the subsampling of the table, finally, at the species level, the curve is far from being flat.

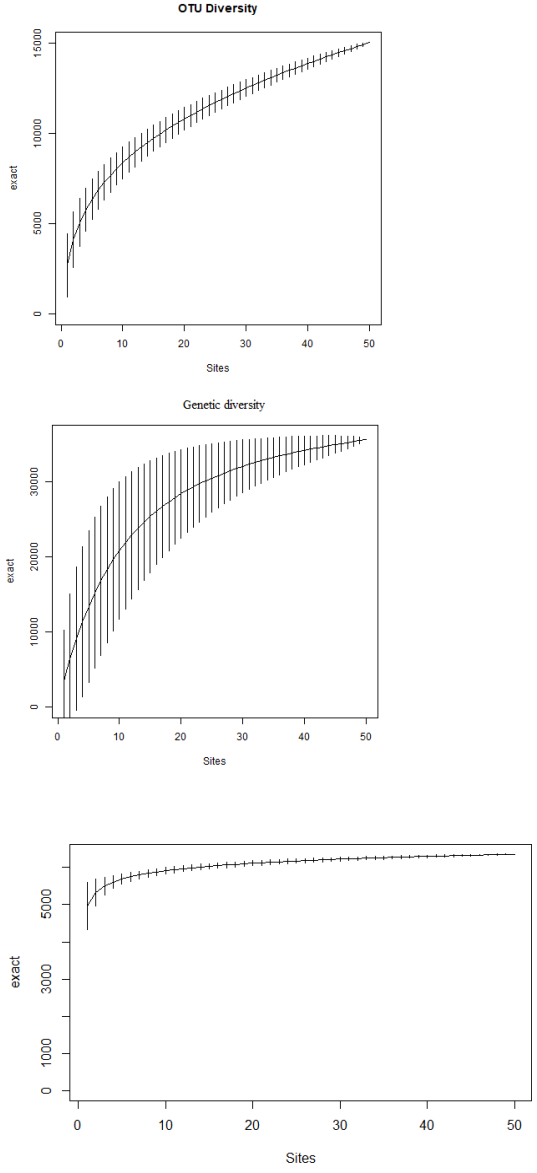

Fig.5. Comparison between species, genetic and function diversity. First species diversity, second genetic diversity and third function diversity.

5. Conclusions

In this work we have explored how the methodologies used to obtain the samples could affect the genetic and species diversity estimated for the ocean microbiota, using accumulation curves to obtain the community diversity and tested how the different levels of diversity are reached. In this case we studied three levels, species using OTUs, genetic using genes and functionality, using the functional tables.

The first step was to analyse if the data that was going to be used to study the functional diversity was suitable for this experiment. Regarding this step, there are some points that need to be mentioned. First, we performed some NMDS tests in all the tables and all of them passed, then we concluded that the most suitable tables are the ones that use the MetaGS normalization, because of the nature of the experiment where we want to compare the samples and a normalization based on the sequencing effort per sample fits better.

Starting with the methodologies, the water volume was expected to be important since the volume is related with the quantity of microbes that are collected, more volume implies more microbes, but that was not the case, the results showed that the amount of biomass collected had no effect on diversity. The results also showed that the filters had a big effect in the diversity, this was expected since the filters select the microbes, the smaller filters allow a larger amount of cells than larger filters .

The most important method in this study was the accumulation curve, all the computations were performed aimed to the final accumulation curves. It has been proved before that accumulation curves are very useful when studying large-scale biological data (Deng C, et al. 2015). As a consequence, this study is centered on them, the final results meet the expectations we had in this method.

The final results of the genetic diversity are expected, we have not reached the limit but we are very close, also, this may have been the result of subsampling the original table, to solve this doubt another analysis needs to be done using a bigger amount of memory from a cluster. The results of the functional analysis were also expected, the limit is easily reached because

of the redundancy of genes that species that share a community have.

As a final conclusion we have to say that this study has thrown some light on which methods are the best to obtain the samples and that the genetic diversity limit may be reached with a few more samples than 50. It is worth mentioning that there are very few environmental studies were 50 or more metagenomes have been sequenced from the same day and place, and that now, thanks to the new technologies, a similar study or even with more metagenomes will cost less, so it is not a dream to say that similar ambiental studies could be done in a variety of places.

6. References

Debnath M, Paul AK, Bisen PS (2007) Natural bioactive compounds and biotechnological potential of marine bacteria. Curr Pharm Biotechnol 8:253–260

Kim SK (2015) Handbook of marine biotechnology. Springer, Heidelberg, p 1512

Santos-Gandelman JF, Giambiagi-deMarval M, Oelemann WM, Laport MS (2014) Biotechnological potential of sponge-associated bacteria. Curr Pharm Biotechnol 15:143–155

Hugerth, Luisa W., John Larsson, Johannes Alneberg, Markus V. Lindh, Catherine Legrand, Jarone Pinhassi, and Anders F. Andersson. 2015. "Metagenome-Assembled Genomes Uncover a Global Brackish Microbiome." Genome Biology 16 (December): 279.

Mitchell, Alex L., Maxim Scheremetjew, Hubert Denise, Simon Potter, Aleksandra Tarkowska, Matloob Qureshi, Gustavo A. Salazar, et al. 2018. "EBI Metagenomics in 2017: Enriching the Analysis of Microbial Communities, from Sequence Reads to Assemblies." Nucleic Acids Research 46 (D1): D726–35.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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

Leonardo Dapporto, Matteo Ramazzotti, Simone Fattorini, Roger Vila, Gerard Talavera and Roger H.L. Dennis (2020). recluster: Ordination Methods for the Analysis of Beta-Diversity Indices. R package version 2.9. https://CRAN.R-project.org/package=recluster

Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.Y., Mao, C.X., Chazdon, R.L. & Longino, J.T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. J. Plant Ecol. 5: 3–21.

Leonardo Dapporto, Matteo Ramazzotti, Simone Fattorini, Roger Vila, Gerard Talavera and Roger H.L. Dennis (2020). recluster: Ordination Methods for the Analysis of Beta-Diversity Indices. R package version 2.9. https://CRAN.R-project.org/package=recluster

Kirchman, David L. 2008. *Microbial Ecology of the Oceans*. Hoboken, N.J: Wiley.

Santos-Júnior, Célio Dias, Hugo Sarmento, Fernando Pellon de Miranda, Flávio Henrique-Silva, and Ramiro Logares. 2020. "Uncovering the Genomic Potential of the Amazon River Microbiome to Degrade Rainforest Organic Matter." *Microbiome* 8 (1): 151.

Dupont, Chris L., Dreux Chappell, Ramiro Logares, and Maria Vila-Costa. 2010. "A Hitchhiker's Guide to the New Molecular Toolbox for Ecologists." https://doi.org/10.4319/ecodas.2010.978-0-9845591-1-4.17.

Logares, Ramiro, Shinichi Sunagawa, Guillem Salazar, Francisco M. Cornejo-Castillo, Isabel Ferrera, Hugo Sarmento, Pascal Hingamp, et al. 2014. "Metagenomic 16S rDNA Illumina Tags Are a Powerful Alternative to Amplicon Sequencing to Explore Diversity and Structure of Microbial Communities." *Environmental Microbiology* 16 (9): 2659–71.

Santos Júnior, Celio & Sarmento, Hugo & Miranda, Fernando & HenriqueSilva, 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.

Locey, Kenneth J., and Jay T. Lennon. 2016. "Scaling Laws Predict Global Microbial Diversity."

Proceedings of the National Academy of Sciences of the United States of America 113 (21): 5970–75.
Lima-Mendez, Gipsi, Karoline Faust, Nicolas Henry, Johan Decelle, Sébastien Colin, Fabrizio Carcillo, Samuel Chaffron, et al. 2015. "Ocean Plankton. Determinants of Community Structure in the Global Plankton Interactome." *Science* 348 (6237): 1262073.
Logares, Ramiro, Ina M. Deutschmann, Pedro C. Junger, Caterina R. Giner, Anders K. Krabberød, Thomas S. B. Schmidt, Laura Rubinat-Ripoll, et al. 2020. "Disentangling the Mechanisms Shaping the Surface Ocean Microbiota." *Microbiome* 8 (1): 55.
Bolhuis, Henk, and Mariana Silvia Cretoiu. 2016. "What Is so Special About Marine Microorganisms? Introduction to the Marine Microbiome—From Diversity to Biotechnological Potential." In *The Marine Microbiome: An Untapped Source of Biodiversity and Biotechnological Potential*, edited by Lucas J. Stal and Mariana Silvia Cretoiu, 3–20. Cham: Springer International Publishing.
Lambert, Stefan, Margot Tragin, Jean-Claude Lozano, Jean-François Ghiglione, Daniel Vaulot, François-Yves Bouget, and Pierre E. Galand. 2019. "Rhythmicity of Coastal Marine Picoeukaryotes, Bacteria and Archaea despite Irregular Environmental Perturbations." *The ISME Journal* 13 (2): 388–401.
Ugland, Karl I., John S. Gray, and Kari E. Ellingsen. 2003. "The Species-Accumulation Curve and Estimation of Species Richness." *The Journal of Animal Ecology* 72 (5): 888–97.
Deng, Chao, Timothy Daley, and Andrew D. Smith. 2015. "Applications of Species Accumulation Curves in Large-Scale Biological Data Analysis." *Quantitative Biology (Beijing, China)* 3 (3): 135–44.