BIOGEOGRAPHY OF BIOSYNTHETIC GENE CLUSTERS IN THE MARINE ENVIRONMENT

E. Pereira [1]*, P. L. Buttigieg [2], M. Medema [3], M. Yeong [3], R. Kottmann [1], A. Fernandez-Guerra [4] and F. O. Glöckner [5]
[1] Max Planck Institute for Marine Microbiology - epereira@mpi-bremen.de
[2] Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research
[3] Wageningen University
[4] University of Oxford
[5] Jacobs University Bremen

Abstract
Marine microbial communities produce a high diversity of natural products encoded by Biosynthetic Gene Clusters (BGCs). So far most of this metabolic diversity remains unknown, due to technical limitations in studying microorganisms. Metagenomics represents a powerful approach to access the metabolic potential of microbial communities. In this work, we used the metagenomic data from the TARA Oceans and Ocean Sampling Day sequencing campaigns, in order to study the biogeography of BGCs in marine environments. These datasets offer a valuable opportunity to explore the unknown world of BGCs of marine microbial communities, and the underlying mechanisms that structure the composition of these genes in the environment. It represents the first biogeographic analysis of BGCs on marine environments on a global scale.

Keywords: Biogeography, Mediterranean Sea

The secondary metabolism (SM) of microorganisms comprises a large diversity of functions and compounds. SM plays an important role in the ecology and physiology of microbes. Many aspects of their lifestyle are affected by their SM, including nutrient scavenging, chemical synthesis and environmental sensing. Additionally, SM produces a variety of products that are highly valuable for industrial and medical applications, like vitamins, antibiotics, bioplastics among others. The genes involved in these metabolic pathways are commonly organized in Biosynthetic Gene Clusters (BGCs): physically clustered genes that encode the biosynthetic enzymes for a pathway.
Metagenomic data provides an opportunity to access this reservoir of microbial functional diversity. Although some of the biggest metagenomic sequencing efforts are from the ocean (e.g. Global Ocean Sampling Expedition [1]; TARA Oceans [2] and Ocean Sampling Day [3]), to our knowledge, there are no comprehensive surveys exploring BGCs composition in this environment.

In this work, we study the biogeography of BGCs in marine environments. That is to say, study their patterns in space, in time and along environmental gradients, and understand the processes generating and maintaining such distribution patterns. In turn, these analyses will allow us to explore the unknown universe of BGCs in the ocean and contribute to the discovery of new bioactive compounds. All data were taken from the TARA Oceans and Ocean Sampling Day campaigns. The majority of the samples from these data sets correspond to open ocean and coastal environments, respectively. In total, we analyzed 392 metagenomic samples. Based on 66,531,749 contigs from the assembly of these metagenomes, we identified 13,137 BGCs, belonging to 49 different classes. For this task, we first used the UproC tool [4] to extract contigs with signatures proteins or protein domains of BGCs. The resulting subset of contigs was classified with antiSMASH [5]. Additionally, the corresponding 16S ribosomal DNA amplicon sequences from these samples were extracted using SortMeRNA [6], refined and classified by the SILVAngs pipeline [7]. Finally, the BGC and taxonomic annotations were integrated with the environmental data. Applying a series of biogeographic exploratory analyses, we aim to study the association between the BGCs and taxonomic distributions, and the environmental parameters.

Those biogeographic analyses will contribute to the understanding of the factors determining the diversity and patterns of BGCs as well as their taxonomic origin, i. e. exploring the differences between coastal and open ocean environments (Fig. 1). Furthermore, this study helps to understand the ecological and evolutionary processes that shape the relationship between marine microbial communities and the environment. Last but no least, the biogeographic analyses of BGCs can be useful for identifying hotspots of biosynthetic diversity and guide the discovery of novel bioactive natural products in the future.

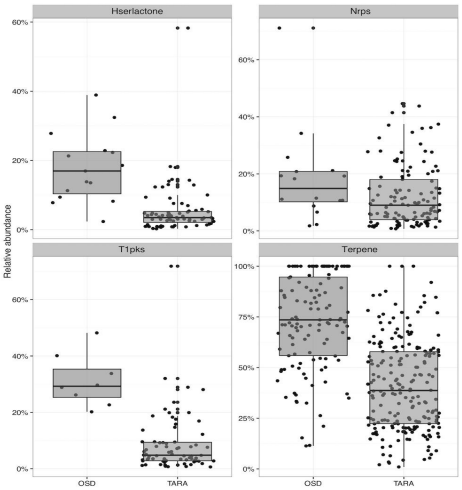

Fig. 1. Boxplot of the BGC distributions of the OSD and TARA datasets of four different classes. Hserlaconte: Homoserine Lactone Synthase, Nrps: Nonribosomal Peptide Synthetase, T1pks: Type I Polyketide Synthase and Trepene: Terpene Synthase.

References
1 - Rusch DB et al. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5 (3):e77
2 - Karsenti E et al. (2011) A holistic approach to marine eco-systems biology. PLoS Biol 9(10):e1001177
3 - Kopf A et al. (2015) The ocean sampling day consortium. GigaScience 4:27
4 - Meinicke P. (2014) UProC: tools for ultra-fast protein domain classification. Bioinformatics 31(9):1382–8
5 - Blin K et al. (2013) antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res 41(Web Server issue):W204–12
6 - Kopylova E, Noé L, Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28 (24):3211–7
7 - Quast C et al. (2013) The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res 41 (D1):D590–6