Improvements in natural product biosynthetic gene clusters research and functional trait-based approaches in metagenomics

by

Emiliano Pereira Flores, MSc.

A Thesis submitted in partial fulfillment of requirements for the degree of Doctor of Philosophy in Bioinformatics

Thesis committee:

Prof. Dr. Frank Oliver Glöckner (Chair)
Max Planck Institute for Marine Microbiology
Jacobs University Bremen

Prof. Dr. Matthias S. Ullrich
Jacobs University Bremen

Prof. Dr. Marc-Thorsten Hütt
Jacobs University Bremen

Asst. Prof. Dr. Marnix Medema
Wageningen University, Netherlands

Prof. Dr. Matias Arim
University of the Republic, Uruguay

Date of Defense: September 3, 2019

Department of Life Science and Chemistry, Jacobs University Bremen
Statutory Declaration

<table>
<thead>
<tr>
<th>Family Name, Given/First Name</th>
<th>Pereira Flores, Emiliano</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matriculationnumber</td>
<td>20331319</td>
</tr>
<tr>
<td>What kind of thesis are you submitting:</td>
<td>PhD-Thesis</td>
</tr>
</tbody>
</table>

**English: Declaration of Authorship**

I hereby declare that the thesis submitted was created and written solely by myself without any external support. Any sources, direct or indirect, are marked as such. I am aware of the fact that the contents of the thesis in digital form may be revised with regard to usage of unauthorized aid as well as whether the whole or parts of it may be identified as plagiarism. I do agree my work to be entered into a database for it to be compared with existing sources, where it will remain in order to enable further comparisons with future theses. This does not grant any rights of reproduction and usage, however.

The Thesis has been written independently and has not been submitted at any other university for the conferral of a PhD degree; neither has the thesis been previously published in full.

**German: Erklärung der Autorenschaft (Urheberschaft)**


Diese Arbeit wurde in der vorliegenden Form weder einer anderen Prüfungsbehörde vorgelegt noch wurde das Gesamtdokument bisher veröffentlicht.

---------------------------------------------------------------
Date, Signature
Abstract

Microorganisms encompass a large metabolic diversity, are key drivers of ecosystem functioning and are fundamental for maintaining all other forms of life on earth. Metagenomics, by allowing access to the genomic material obtained directly from the environment, represents a major field of research in microbiology. The recent advent of high-throughput sequencing technologies has pushed the scale and scope of metagenomic studies. Today more than ever, metagenomics is critical to advance our knowledge of microorganisms. The research work presented in this thesis develops two (interconnected) lines of research in the field of metagenomics. The first of these is the study of natural product Biosynthetic Gene Clusters (BGCs). Microorganisms encode a broad diversity of BGCs responsible for the production of several compounds with valuable industrial applications. BGCs are also important from an ecological perspective, as these participate in interactions between organisms and with the environment. To improve the exploitation of metagenomic data in BGC exploration analyses, we developed the Biosynthetic Gene Cluster Metagenomic Exploration toolbox (BiG-MEx). BiG-MEx can rapidly estimate the BGC domain and chemical class composition of a metagenomic sample, and perform a series of domain diversity analyses. The second research line developed in this thesis is the application of functional trait-based approaches in metagenomics. Functional traits provide valuable information to study microorganisms’ ecology. We developed the ags.sh and acn.sh tools, which rapidly and accurately quantify the average genome size and 16S rRNA gene average copy number from metagenomic data. These are two metagenomic functional traits that can be used to characterize the dominant ecological strategies in a microbial com-
munity. Additionally, we developed the Mg-Traits pipeline and database. The pipeline integrates BiG-MEx domain profiling routine, and the ags.sh and acn.sh tools together with other metagenomic functional traits into a single workflow, and the database provides metagenomic functional trait data computed on publicly available metagenomes. In conclusion, this thesis contributes to the advancement of BGC mining analyses and functional trait-based approaches in metagenomics.
Abbreviations

A: Adenylation
AB: Acidic to Basic
ACN: 16S rRNA gene Average Copy Number
AGS: Average Genome Size
APE: Absolute Percentage Error
AT: Acyltransferase
BGC: Biosynthetic Gene Cluster
C: Condensation
DACC: Data Analysis and Coordination Center
DCM: Deep Chlorophyll Maximum
DOM: Dissolved Organic Matter
FT: Functional Trait
GOS: Global Ocean Sampling
HMP: Human Microbiome Project
KS: Ketosynthase
MAG: Metagenome Assembled Genome
MdAPE: Medians of the Absolute Percentage Error
MES: Mesopelagic
MIX: Subsurface epipelagic mixed layer
MLR: Multiple Linear Regression
NGS: Next Generation Sequencing
NGs: Number of Genomes
NRPS: Non-Ribosomal Peptide Synthase
ODU: Operational Domain Unit
OM-RGC: Ocean Microbial Reference Gene Catalogue
ORF: Open Reading Frame
OTU: Operational Taxonomic Unit
PCA: Principal Component Analysis
PCoA: Principal Coordinates Analysis
PCP: Peptidyl Carrier Protein
PCR: Polymerase Chain Reaction
PD: Phylogenetic Diversity
PERMANOVA: Permutational Multivariate Analysis of Variance
PKS: Polyketide Synthase
PSSM: Position Specific Scoring Matrix
RF: Random Forest
SRF: Surface
T: Acyl carrier protein
# Contents

1 Introduction 3

1.1 Why microorganisms? .................................................. 3
1.2 From early microbiology to metagenomics .......................... 4
1.3 Metagenomic computational challenges and advances .......... 7
1.4 Functional trait-based approaches in metagenomics .......... 11
1.5 Natural product biosynthetic gene clusters ......................... 13
1.6 Research objectives .................................................... 16

2 Results and Discussion 18

2.1 Overview ................................................................. 18
2.2 Mining metagenomes for natural product biosynthetic gene clusters: unlocking new potential with ultrafast techniques ........ 20
2.3 Fast and accurate average genome size and 16S rRNA gene average copy number computation in metagenomic data ............ 56
2.4 Mg-Traits pipeline and database: advancing functional trait-based approaches in metagenomics ............................. 70

3 Final remarks and conclusions 100

3.1 Metagenomic mining of natural product biosynthetic gene clusters ................................................................. 100
3.2 Functional trait-based approaches in metagenomics ............ 103

4 Outlook 106
Chapter 1

Introduction

1.1 Why microorganisms?

Microorganisms are defined as "small living organisms not to be seen with the naked eye" (Singh and Dwivedi 2004). This group includes prokaryotes, viruses, and eukaryotes like algae, protozoa, and fungi. Microorganisms comprise an immense phylogenetic diversity and inhabit every conceivable niche in environments ranging from the animal gut to the abyssal ocean. The wide range of environmental conditions inhabited by microorganisms has led to the development of a great diversity of survival and growth strategies. In addition, microorganisms are critical for maintaining all other forms of life on earth and play a fundamental role in global biogeochemical processes. The human gut microbiome and the microbial carbon pump are two different examples that illustrate the importance of these organisms and why their study is fundamental to understand life on earth. The human gut microbiome is estimated to contain at least a similar number of cells as the human body and is known to be involved in physiological functions that are vital for the host like harvesting energy, regulating the host immunity and protecting against pathogens (Thursby and Juge 2017). On the other hand, the microbial carbon pump describes a global scale pathway, which takes place in the oceans and where microbes “pump” labile dissolved organic matter (DOM) into a pool of refractory DOM. Such process results in long-term carbon sequestration in
the water column and consequently influence the global climate (Jiao et al. 2010). It is important to note that the microbial carbon pump was proposed less than a decade ago, and in fact, the role of microbes in the oceans as major players in global biogeochemical processes started to be acknowledged only in the last few decades (Pomeroy 1974; Legendre et al. 2015). Even today, there is much to learn about essential aspects of microorganisms’ biology, for example, microorganisms’ metabolic capacities, distributional and diversity patterns, and community dynamics.

1.2 From early microbiology to metagenomics

The first study of microorganisms can be traced back to van Leeuwenhoek in the 1670s. Van Leeuwenhoek was able to see microorganisms using self-made microscopes, which had an unprecedented magnifying power. However, it was not until the work of Louis Pasteur in the late 1800s associating the fermentation process with microorganisms, and the subsequent invention of pure culture by Robert Koch, that microbiology started to develop, leading to the golden age of microbiology (Guardino 2005). This period was characterized by major scientific advances, like the germ theory of disease, the identification of different infectious disease agents, staining techniques, and vaccination procedures (Blevins and Bronze 2010). Robert Koch’s pure culture technique established a rigorous approach to study microorganisms and set a standard in microbiology. Today, most of our knowledge about microorganisms’ biochemistry and molecular biology was gained through the application of pure culture techniques. However, this technique suffers from a critical limitation: the majority of microorganisms cannot be grown in pure culture. This problem, commonly referred to as the ‘great plate count anomaly’ (Staley and Konopka 1985), has biased microbiology and especially limited microbial ecology studies.

In the 1900s, seminal scientific studies like the demonstration that DNA constitutes the genetic material, and the discovery of the DNA structure and replication mechanism, had crucial implications in microbiology. Building on these studies, molecular tech-
niques were developed in the last few decades and paved the way for breakthroughs in microbiology. In 1977, Carl Woese showed that the rRNA genes can be used as evolutionary chronometers. Based on the phylogenetic information derived from these genes, Woese redrew the tree of life and discovered a new kind of microbial life: Archaea (Woese 1987). Woese’s work had a highly significant impact on microbial ecology and evolution studies, by allowing an objective taxonomic classification and assessment of phylogenetic relationships. Later on, in 1985 Pace and colleagues analyzed rRNA gene sequences obtained directly from an environmental sample without culturing (Pace et al. 1986). Although initially, this approach was highly cumbersome, the invention of the Polymerase Chain Reaction (PCR) technique facilitated its application. PCR allowed scientists to selectively make copies of an rRNA gene from a mixed DNA sample, accelerating the discovery of new taxa (Amann, Ludwig, and Schleifer 1995). Since these pioneering advances, the ribosomal rRNA genes (i.e., 16S and 23S rRNA) have been established as the phylogenetic markers of choice, and a great number of surveys have been applied to describe the microbial diversity in nature. This approach was first applied by Giovannoni in 1990 to explore the Sargasso Sea (Giovannoni et al. 1990) and is still applied today to survey environments as dissimilar as the air and deep-sea sediments (e.g., Cha et al. (2017) and Zhang et al. (2016), respectively). It has been estimated that from 2008 to 2012 there were about $4 \times 10^4$ prokaryotic species discovered per year (Yarza et al. 2014). However, such rRNA gene surveys provide little information about the metabolic capabilities of microorganisms.

Improvements in sequencing technologies, primarily due to the Human Genome Project research efforts, were one of the next changing course developments in microbiology (Hood and Rowen 2013). In 1995, H. influenzae was the first bacterial species to have its genome sequenced (Fleischmann et al. 1995). In the ensuing years, and particularly after the advent of Next Generation Sequencing (NGS) technologies, sequencing costs dropped radically (Wetterstrand 2015) and complete sequenced genomes started to accumulate rapidly (https://www.ncbi.nlm.nih.gov/genbank/statistics). In addition, advances in sequencing technologies allowed the devel-
opment of arguably one of the most important culture-independent methods: metagenomics. Metagenomics is defined as the analysis of microorganisms based on genomic data (i.e., a pool of microbial genomes from a community) obtained directly from the environment (Handelsman et al. 1998). As such, a metagenomic sample consists of millions of short-reads, for which their taxonomic origin and genome coordinates are unknown. This methodology provides access to the genomic content of microorganisms that cannot be grown in the laboratory. It allows the characterization of the coding potential of environmental samples and sequences not represented in databases (Gilbert and Dupont 2011).

The initial work of Delong’s group in 1996 was key in defining the field of metagenomics. Delong and co-workers constructed an environmental library (fosmid DNA) from coastal marine picoplankton and identified a 38.5kb clone containing a 16S rRNA gene, which indicated that the DNA fragment was derived from an archeon that has never been cultured (Stein et al. 1996). Several other works followed characterizing the microbial communities from different environments (e.g., Béjà et al. 2000; Rondon et al. 2000; Nesbø et al. 2005). However, the first large-scale metagenomic sequencing project was performed in 2004 by Craig Venter and colleagues (Venter et al. 2004). These researchers used Sanger sequencing technology to sequence the marine microbial communities found at the Bermuda Atlantic Time-series Study (BATS) station in the Sargasso Sea, generating more than one billion base pairs. As an extension of this work, Rusch et al. (2007) generated 6.3 billion base pairs using Sanger sequencing technology during the Global Ocean Sampling (GOS) expedition. In this endeavor, Rusch and colleagues collected several samples (initially 44 and later extended to 83 samples) covering a wide range of distinct surface marine environments from across the globe. A relevant example of a more recent large-scale metagenomic project is the TARA Oceans expedition (Karsenti et al. 2011). This dataset, which also consists of metagenomic samples of marine environments obtained across the globe, played a central role in the development of this thesis. TARA Oceans metagenomes were sequenced using NGS techniques, which in comparison to Sanger-sequenced
metagenomes, allowed for an increase of several orders of magnitude in the amount of data generated. In total, the TARA Oceans expedition generated 243 metagenomes which sum more than 7 trillion sequenced base pairs. A large number of publications followed these works, mainly based on the GOS and TARA Oceans datasets, which showed the value of such sequencing efforts to improve our understanding of the marine microbial communities (for a list of publications see https://www.jcvi.org/gos and https://www.embl.de/tara-oceans/start/publications).

1.3 Metagenomic computational challenges and advances

Typically, to characterize a metagenome, its sequences are functionally annotated based on their sequence similarity to genes of known function (e.g., using the Pfam database (Finn et al. 2014)). In addition, phylogenetic marker genes are identified and taxonomically annotated based on their sequence similarity or phylogenetic relationships to taxonomically annotated genes (e.g., annotating 16S rRNA genes based on their similarity to reference genes of the Silva database (Quast et al. 2013)). This data is finally summarized as counts of reads assigned to different functional genes or taxonomic groups (see Fig. 1.3-1) (see Quince et al. (2017) for a review).

The large amounts of data generated by current sequencing technologies, although highly advantageous, pose a challenge to the characterization of metagenomic samples. Although computational power has also increased considerably during the last decades, it cannot cope with the amount of data currently generated (see Fig. 1.3-2). This situation will aggravate even more in the near future, as the volume of metagenomic data continues to increase exponentially (Benson et al. 2017). On top of this, the short-read sequences produced by NGS technologies add another layer of complexity to the metagenomic data analysis. In most cases, short-read sequences can only capture gene fragments, and consequently, are more challenging to annotate (i.e., there is not enough information to determine its function).
Chapter 1. Introduction

Figure 1.3-1: Metagenomic project workflow. A metagenomic project starts by sampling the DNA from a naturally occurring microbial community, which is sequenced after creating a shotgun metagenomic library. Subsequently, the generated metagenomic data is subjected to a quality check (e.g., removing low-quality regions, and reads below a specific length threshold), and functionally and taxonomically annotated. Alternatively, the quality checked metagenome can be assembled, and the gene annotation is performed on the resulting contigs. In this latter case, to estimate the gene abundance (or counts), the metagenomic reads are mapped back on the assembled contigs. Finally, the functional and taxonomic count data of a metagenomic sample is represented as a vector (or matrix for various metagenomic samples), to be used in downstream analyses. The Illumina sequencer MiSeq image is courtesy of Illumina, Inc.

Figure 1.3-2: Computational power vs. sequencing costs. The plot shows a comparison between the computer power increase represented by Moore’s law, which doubles every two years, and the number of sequenced megabases (Mb) per US dollar. The latter outpaced the computer power by several orders of magnitude from the beginning of 2008 onwards, with the advent of NGS technologies. Data from https://www.genome.gov/sequencingcostsdata.
Chapter 1. Introduction

Such massive amounts of metagenomic short-read data have demanded the development of new bioinformatics algorithms and tools to improve and accelerate the characterization of metagenomic data. The Ultrafast Protein Classification (UProC) tool (Meinicke 2015) is an example of the recent bioinformatic advances toward this end. Although other tools developed in the last few years have also contributed to improving the characterization of metagenomes (e.g., DIAMOND (Buchfink, Xie, and Huson 2014), FragGeneScanPlus (Kim et al. 2015), and MMseqs2 (Steinegger and Söding 2017)), UProC played a strategic role in the work of this thesis and we briefly describe its workflow in Box 1. This tool implements a new algorithm termed Mosaic Matching to classify query metagenomic reads into protein domains. This new algorithm makes UProC up to three orders of magnitude faster and considerably more sensitive than profile-based methods when annotating unassembled short-read sequences.

The large amounts of short-read metagenomic data also pose new opportunities. For example, metagenomics has the potential to improve taxonomic surveys. 16S rRNA genes identified in metagenomic data have been shown to be a better alternative to 16S rRNA gene amplicons, given that the latter are subjected to PCR amplification biases. Thanks to the large reduction in sequencing costs, it is now economically possible to perform taxonomic profiling using metagenomic data, which also has the advantage of retrieving functional genes (Logares et al. 2014). In addition, metagenomic data can be used to recover genome sequences from environmental samples (i.e., Metagenome Assembled Genomes (MAGs)). Previous works have already used this approach to expand the tree of life by generating several hundreds of MAGs from environmental metagenomic samples (Parks et al. 2017; Tully, Graham, and Heidelberg 2018).
Box 1. UProC protein annotation workflow

UProC protein annotation workflow consists of the following tasks: First, oligopeptides of the query sequence are placed into an oligopeptide dictionary derived from the sequences of reference protein families (e.g., Pfam (Finn et al. 2014)). This task is performed to determine the query oligopeptide neighboring words in the dictionary and their associated protein families. Subsequently, a similarity assessment between the query and the neighboring reference oligopeptides is performed based on Position Specific Scoring Matrices (PSSMs). UProC uses PSSMs previously computed for all reference oligopeptides in the dictionary using a supervised machine learning approach, with the objective to optimize the scoring of residues for a correct classification of query oligopeptides. Next, the Mosaic Matching score is computed as the maximum possible position-specific score of each residue in the query sequence, summing up over all query oligopeptides matching the same reference protein family. Lastly, the query sequence is classified into a particular protein family if its total Mosaic Matching score is higher than a previously determined length-dependent noise threshold. Creating the oligopeptide dictionary and computing the PSSMs can be a time-consuming process. However, these only have to be computed once, and allow a rapid search and similarity assessment of oligopeptides, which together with the computation of the Mosaic Matching score, provide a very fast and accurate protein annotation.

Mosaic Matching sketch. Oligopeptides of the query sequence are compared with oligopeptides of the reference database. Matching reference oligopeptides of the same protein family are used to compute the Mosaic Matching score as the sum of the maximum PSSM score of each residue (matching words tend to overlap, so there can be different position-specific scores for the same residue). To classify the query sequence, the Mosaic Matching score is compared with a previously determined length-dependent noise threshold. This example shows query oligopeptides matching the reference oligopeptides of the Pfam family PF01370. The color gradient from red to blue represents high to low PSSM substitution scores. Adapted from Meinicke (2015).
Chapter 1. Introduction

1.4 Functional trait-based approaches in metagenomics

By allowing a holistic analysis of the genomic material of microorganisms, metagenomics offers excellent opportunities to perform microbial ecology studies (i.e., studying the interactions between organisms and with the environment). In particular, microbial ecology studies can benefit from comparative metagenomic analyses based on metagenomic samples and environmental data obtained from different environmental conditions. The taxonomic and functional community composition can be integrated with the geographic and environmental data, to study different ecological aspects of microorganisms, for example, their ecosystem functions, environmental adaptation, and biogeography. One of the earliest efforts in this direction is the work of DeLong and co-workers in 2006 (Delong et al. 2006). These authors analyzed the planktonic microbial communities from different water depths and revealed a vertical zonation of taxonomic groups and functional genes (e.g., involved in carbon and energy metabolism, attachment and motility), which reflected the community dynamics and ecological strategies associated with different environmental conditions. A more recent study in this same line of research is the work of Guidi et al. (2016), who linked taxonomic and functional data to an ecosystem process. Based on the analysis of metagenomic and ampli-con data obtained during the TARA Oceans expedition, these authors delineated the planktonic microbial communities and identified bacterial and viral genes associated with the export of carbon to the deep Ocean. Similar to these examples, several other works have been performed to characterize the microbial communities from diverse environments and gain insight into different aspects of their ecology (e.g., Tringe et al. 2005; Dinsdale et al. 2008; The Human Microbiome Project Consortium 2012; Sunagawa et al. 2015).

The studies mentioned above show the potential of applying a functional trait framework to study microorganisms’ ecology based on metagenomic data. Functional traits are defined as characteristics of an organism that are linked to its performance (Violle et al. 2007). Functional trait-based approaches have been extensively applied
in plant ecology. For example, leaf and seed mass, and stem-specific density have been used to explain vegetation distributional patterns (Van Bodegom, Douma, and Verheijen 2014). Similarly, seed mass, wood density and leaf economic traits (i.e., lifespan, specific leaf area, and nitrogen concentration) have been shown to explain variation in life history strategies (Adler et al. 2014). Microbial ecology studies can benefit from borrowing functional trait concepts and approaches from plant ecology. In particular, considering that microorganisms are subjected to high rates of horizontal gene transfer, which result in high functional plasticity and limit the applicability of studies based on taxonomic information (Cordero and Polz 2014; Martiny et al. 2015). Although the functional trait quantification in microbial individuals from a naturally occurring community might not be technically feasible, functional trait-based approaches have a synergetic effect with metagenomics: several functional traits can be measured from metagenomic data, constituting community aggregated functional traits (Fierer, Barberán, and Laughlin 2014). For example, traits like the average genome size and 16S rRNA operon copy number, functional and taxonomic diversity, sporulation and antibiotic resistance, can be readily measured from metagenomic data. Linking the environmental variation to such metagenomic traits has been proposed as a powerful approach to study the ecology of microorganisms (McGill et al. 2006; Barberán et al. 2012; Fierer, Barberán, and Laughlin 2014). Examples of previous works within the framework of functional trait-based approaches in metagenomics include the study of community assembly processes (Burke et al. 2011), ecosystem functioning (Babilonia et al. 2018), and community responses to environmental change (Leff et al. 2015). Together, these show that coupling metagenomics and functional trait-based approaches can produce more quantitative and predictive microbial ecology studies.
1.5 Natural product biosynthetic gene clusters

Metagenomics constitutes a highly promising venue for accessing the reservoir of natural product Biosynthetic Gene Clusters (BGCs) encoded by microorganisms (Medema and Fischbach 2015). BGCs are defined as physically clustered genes that encode the biosynthetic enzymes of a pathway and are responsible for the production of diverse chemical structural classes including polyketides, peptides, oligosaccharides, terpenoids, and alkaloids. Many of these compounds are highly valuable for industrial and medical applications. For example, some BGC products are used as vitamins, antibiotics, antifungals, chemotherapeutics, biofuel, bioplastics and detergents. Additionally, BGCs play diverse roles in the ecology and physiology of microbes like defense, nutrient scavenging, energy production, and quorum sensing (Fischbach and Voigt 2010).

BGCs are composed of several genes and can exceed 100 kbp (Wakimoto et al. 2014). Commonly, BGCs’ constituent genes are composed of different modules, each responsible for producing a specific task in the synthesis of a metabolite. Such organization facilitates gene rearrangement mutations, which favor the rapid diversification of these genomic elements. Common examples of BGCs are the Polyketide synthases (PKSs) and non-ribosomal peptide synthases (NRPSs). These enzyme families account for many natural products with clinical applications. PKS enzymes produce a huge diversity of polyketides based on the combinatorial incorporation of a few building blocks, mainly Malonyl-CoA and methylmalonyl-CoA. On the other hand, NRPS utilize the 20 proteinogenic amino acids, plus a variety of non-proteinogenic amino acids and aryl acids to produce oligopeptides. These enzymes are organized in an assembly line fashion, where the identity and order of each domain specify the sequence and chemistry of the monomers incorporated. Typically, a minimal set of domains that compose a ‘module’ in a PKS are the ketosynthase (KS), acyltransferase (AT) and acyl carrier protein (T) domains (see Fig. 1.5-1). Likewise, in the case of NRPSs, a minimal module is composed of the condensation (C), adenylation (A) and peptidyl carrier protein modules.
(PCP) domains. The primary products produced by PKSs and NRPSs are commonly subjected to tailoring modifications, which increase even further the chemical diversity produced by these enzymes (Walsh and Fischbach 2010).

BGCs impose a metabolic cost for the producer organism, which implies that there must be a selective pressure to maintain these genes. BGCs that provide a strong selective advantage are expected to be consistently distributed within and between different taxa, as is the case for antibiotics and cytotoxins (Jensen 2016). Also, BGCs can act as public goods, that is, specific BGCs carried by a few organisms benefit other members of the community (commonly within an ecological defined population). This phenomenon has been observed for quorum sensing compounds (Sandoz, Mitzmberg, and Schuster 2007), iron-chelating siderophores (Harrison et al. 2008), and antibiotics (Cordero et al. 2012).

The fact that a BGC encodes a complete metabolic pathway, which provides an ecological advantage for the producer organism, promotes the horizontal transferring of these genes (Fischbach, Walsh, and Clardy 2008). Gene exchange can act as a creative force, and it has been shown to be critical for the adaptation to different environments (Cordero et al. 2012; Niehus et al. 2015).

In sum, BGCs are a measurable property of an organism and are responsible for producing specific phenotypes with implications for the organism’s performance. That is, these are functional traits that characterize the ecology of microorganisms.
Figure 1.5-1: Polyketide BGC sequence and synthase. The BGC encoded proteins form a polyketide synthase (PKS) which synthesizes erythromycin A. This PKS is composed of three proteins (harboring 28 domains in total) organized into seven modules. The ketosynthase (KS), acyltransferase (AT) and acyl carrier protein (T) domains constitute a minimal set to form a module and are responsible for adding a building block into the growing chain. The AT domain selects the acyl-CoA monomer to be incorporated and catalyzes its transthiolation to the downstream T domain. The KS domain accepts the growing chain from the previous module and catalyzes the C-C bond formation (decarboxylative condensation) between this and the monomer (elongation group) loaded in the current module. The ketoreductase (KR), dehydratase (DH), and reductase (ER) domains control the oxidation state of the chain, while the thioesterase (TE) domain, terminates the synthesis. Finally, the molecule released from the protein complex (6-deoxyerythronolide B (6-DEB)) is tailored (hydroxylated and glycosylated) to yield erythromycin A (Fischbach and Walsh 2006; Walsh and Fischbach 2010). Adapted from Walsh and Fischbach (2010).
1.6 Research objectives

Overall, this thesis has the objective to contribute to natural product Biosynthetic Gene Cluster (BGC) research and the application of functional trait-based approaches, to improve the exploitation of metagenomic data. The first objective is to develop a tool to mine BGCs in metagenomic data (section 2.2). Metagenomics offers excellent opportunities to explore the broad diversity of BGCs encoded by microorganisms to improve our understanding of the ecological implications of these genes and identify new bioactive compounds. However, due to the inherent complexity of metagenomic data, BGC analyses are commonly limited to the few most abundant BGC sequences. The objective of developing a tool dedicated to BGC research, aims specifically, to improve the use of metagenomic data in BGC exploration analyses. The second objective is to advance the applications of functional trait-based approaches in metagenomics. Metagenomics is particularly well suited to compute functional traits on genomic data at the community level (i.e., metagenomic traits). Two highly informative metagenomic traits are the Average Genome Size (AGS) and 16S rRNA gene Average Copy Number (ACN). These allow an assessment of the dominant ecological strategies in a microbial community. In addition, various other relatively simpler metagenomic traits (e.g., GC content, amino acid usage, and functional diversity) can be easily computed and used to identify environmental drivers to gain insight into microorganisms’ ecology. Precisely, this second objective consists of developing the tools to compute the AGS and ACN metagenomic traits (section 2.3) and extending the work of Barberan et al. 2012, to develop a pipeline to integrate the computation of several metagenomic traits into a single workflow (section 2.4). These tools provide a comprehensive set of consistently computed metagenomic traits, which if properly supplemented with contextual environmental data, have the potential to improve the interrogation of metagenomic data in microbial ecology analyses.
Chapter 1. Introduction

Thesis research lines and objectives. Flowchart depicting the integration of the main research lines of this thesis, which contribute to Natural product Biosynthetic Gene Cluster (BGC) research and functional trait-based approaches in metagenomics.
Chapter 2

Results and Discussion

2.1 Overview

This chapter presents three independent research works that summarize the main achievements of this thesis. An overview of these works is presented below.

**Mining metagenomes for natural product biosynthetic gene clusters: unlocking new potential with ultrafast techniques**

**Authors:** Emiliano Pereira-Flores, Pier Luigi Buttigieg, Marnix Medema, Peter Meinicke, Frank Oliver Glöckner and Antonio Fernández-Guerra

**To be submitted in February 2020 to PLOS Computational Biology.**

**Contribution:** Devised (together with AFG and MM) and programmed the tools, performed the analyses and wrote the manuscript.

**Relevance:** Microorganisms are known to encode a high diversity of biosynthetic gene clusters responsible for the production of compounds with valuable industrial applications. The manuscript presents a toolbox named BiG-MEx dedicated to the exploration of biosynthetic gene clusters in shotgun metagenomic data. BiG-MEx is able to rapidly process large volumes of data to estimate biosynthetic gene cluster domain and chemical class profiles, and perform domain diversity analyses.
Chapter 2. Results and Discussion

Fast and accurate average genome size and 16S rRNA gene average copy number computation in metagenomic data

**Authors:** Emiliano Pereira-Flores, Frank Oliver Glöckner and Antonio Fernández-Guerra

**Published in BMC Bioinformatics; Volume 20; Article Number 453; Year 2019.**

**Contribution:** Devised and programmed the tools, performed the analyses and wrote the manuscript.

**Relevance:** The average genomes size and 16S rRNA gene average copy number of a microbial community, are highly informative traits that allow the characterization of microorganisms’ ecological strategies. The manuscript presents the ags.sh and acn.sh tools, which are able to rapidly and accurately compute these traits on unassembled shotgun metagenomic data.

Mg-Traits pipeline and database: advancing functional trait-based approaches in metagenomics

**Authors:** Emiliano Pereira-Flores, Renzo Kottmann, Albert Barberan, Frank Oliver Glöckner and Antonio Fernández-Guerra

**To be submitted in April 2020 to Nucleic Acid Research.**

**Contribution:** Optimized and reprogrammed the scripts, performed the analyses and wrote the manuscript.

**Relevance:** The quantification of metagenomic traits allows better use of shotgun metagenomic data to study microorganisms’ ecology. In this work, we present a pipeline and database dedicated to the computation and storage of 15 metagenomic traits, ranging from GC content to functional diversity. These traits serve as ecological markers to characterize microbial communities associated with different environmental conditions. Mg-Traits pipeline and database allow a consistent computation and comparative analysis of metagenomic traits.
2.2 Mining metagenomes for natural product biosynthetic gene clusters: unlocking new potential with ultrafast techniques
Chapter 2. Results and Discussion

Mining metagenomes for natural product biosynthetic gene clusters: unlocking new potential with ultrafast techniques

Emiliano Pereira-Flores\textsuperscript{1,2}, Pier Luigi Buttigieg\textsuperscript{3}, Marnix Medema\textsuperscript{4}, Peter Meinicke\textsuperscript{5}, Frank Oliver Glöckner\textsuperscript{1,2} and Antonio Fernández-Guerra\textsuperscript{1,2,6}

1. Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany
2. Department of Life Sciences & Chemistry, Jacobs University Bremen gGmbH, Bremen, Germany
3. HGF-MPG Group for Deep Sea Ecology and Technology, Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany
4. Bioinformatics Group, Wageningen University, Wageningen, The Netherlands
5. Department of Bioinformatics, Göttingen University, Göttingen, Germany
6. Oxford e-Research Centre (OeRC), University of Oxford, Oxford, UK

Contact: epereira@mpi-bremen.de
Abstract

Microorganisms produce an immense variety of natural products through the expression of Biosynthetic Gene Clusters (BGCs): physically clustered genes that encode the enzymes of a specialized metabolic pathway. These natural products cover a wide range of chemical classes (e.g., aminoglycosides, lantibiotics, nonribosomal peptides, oligosaccharides, polyketides, terpenes) that are highly valuable for industrial and medical applications. Metagenomics, as a culture-independent approach, has greatly enhanced our ability to survey the functional potential of microorganisms and is growing in popularity for the mining of BGCs. However, to effectively exploit metagenomic data to this end, it will be crucial to more efficiently identify these genomic elements in highly complex and ever-increasing volumes of data. Here, we address this challenge by developing the ultrafast Biosynthetic Gene cluster Metagenomic Exploration toolbox (BiG-MEx). BiG-MEx rapidly identifies a broad range of BGC protein domains, assesses their diversity and novelty, and predicts the abundance profile of natural product BGC classes in metagenomic data. We show the advantages of BiG-MEx compared to standard BGC-mining approaches, and use it to explore the BGC domain and class composition of samples in the TARA Oceans and Human Microbiome Project datasets. In these analyses, we demonstrate BiG-MEx’s applicability to study the distribution, diversity, and ecological roles of BGCs in metagenomic data, and guide the exploration of natural products with clinical applications.
Introduction

Metagenomics offers unique opportunities to mine natural product BGCs in diverse microbial assemblages from a wide range of environments (Reddy et al. 2012; Charlop-Powers et al. 2015; Lemetre et al. 2017). However, given the complexity of microbial communities found in nature, and the limitations of current sequencing technologies, often only a very small fraction of the short-read sequence data can be assembled in contigs long enough to allow the identification of BGC classes. However, the annotation of individual protein domains of BGCs, is much more straightforward, given that these have comparable length to merged paired-end reads. There are several protein domains known to play important functions in the BGC-encoded enzymes. Specific domains or combinations thereof are commonly found in certain types of BGC classes. Accordingly, these are used for the automatic identification of BGC classes in genome sequences (Van Heel et al. 2013; Cimermancic et al. 2014; Weber et al. 2015) and to study the distribution and diversity of particular BGC classes in the environment (Reddy et al. 2012; Ziemert et al. 2012; Charlop-Powers et al. 2015; Borchert et al. 2016; Lemetre et al. 2017). Although there are various BGC mining tools with practical applications (Weber and Kim 2016), only the Natural Product Domain Seeker (NaP-DoS) (Ziemert et al. 2012) and the environmental Surveyor of Natural Product Diversity (eSNaPD) (Reddy et al. 2014) are dedicated to the study of BGC domains. Both of these tools focus on nonribosomal peptides and polyketide synthases (NRPSs and PKSs, respectively), and take assembled or amplicon data as input. Currently, there is no technology available capable of efficiently exploiting raw metagenomic data to study the composition and diversity of natural product BGC classes and domains in the environment.

Capitalizing on the fact that BGC domains can be readily annotated in unassembled metagenomic data, and used to identify the different natural product BGC classes, we developed BiG-MEx. This tool generates ultrafast BGC domain annotations in short-read sequence data and applies a machine-learning approach to predict the BGC class coverage-based abundances (for simplicity, we will refer to these as BGC class
Chapter 2. Results and Discussion

abundance profiles). Additionally, the identified domain sequences are used to carry out a domain-based diversity analysis. This allows BiG-MEx both to deeply exploit metagenomic data, and to adapt to their ever-increasing volume. BiG-MEx consists of three interacting modules that are described below and illustrated in Fig. 2.2-1:

1. **BGC domain identification module.** We use the Ultrafast Protein domain Classification UProC (Meinicke 2015) tool to identify BGC protein domains in short-read sequence data. For this purpose, we created an UProC database, which includes 150 BGC domains covering 44 BGC classes.

2. **BGC domain-based diversity analysis.** This module performs a domain-targeted assembly, clusters the assembled domain sequences to create Operational Domain Units (ODUs) (Ziemert et al. 2014) and computes the ODU alpha diversity. Further, assembled domain sequences are placed onto reference phylogenetic trees. The module includes pre-computed phylogenies for 48 BGC domains. These were selected based on domain sequences from experimentally characterized biosynthetic gene clusters with enough sequence information for phylogenetic analysis.

3. **BGC class abundance prediction module.** We created machine-learning models that predict the abundance of BGC classes based on the domain annotation. The models are class-specific and consist of a random forest (RF) classifier to predict the presence/absence of a BGC class, and a multiple linear regression (MLR) to predict its abundance. These models can be customised to target metagenomic and genomic data from different environments and taxa, respectively.

**Results and discussion**

To evaluate the performance of BiG-MEx, we first assessed how the UProC-based domain identification used in BiG-MEx improves the data processing speed compared to
Chapter 2. Results and Discussion

HMMER (Eddy 2011) (i.e., the traditional approach for domain annotation) for the annotation of the 150 BGC domains. This comparison showed that UProC was on average 18 times faster than HMMER (Supplementary Fig. 6.1-1a). We then evaluated the accuracy of BiG-MEx Operational Domain Unit (ODU) diversity estimation approach. We used BiG-MEx to compute the ODU diversity of the NRPS adenylation (AMP-binding) and condensation domains, as well as the PKS ketosynthase (KS) and acyltransferase (AT) domains in a simulated metagenomic dataset (Marine-TM dataset; see Materials and Methods section 3). Additionally, we computed the ODU diversity of these domains based on the domain sequences obtained from the genome sequences used to simulate the Marine-TM metagenomes. The latter estimates (henceforth, the reference estimates) were assumed to accurately reflect the ODU diversity, as they were computed using the complete domain sequences. We compared BiG-MEx ODU diversity estimates against the reference ODU diversity and observed that these were highly correlated: KS domains had a Pearson’s r of 0.77, while for the other domains the Pearson’s r was greater than 0.9 (Supplementary Fig. 6.1-1b). Lastly, we evaluated BiG-MEx’s BGC class abundance prediction module. We point out that although we modeled the abundance of a few BGC subclasses, we refer to all as BGC classes. For this analysis, we used two different simulated metagenomic datasets, one for training and the other for testing the BGC class abundance models (Marine-RM and Marine-TM, respectively) (see Supplementary Table 6.1-1). We predicted the BGC class abundances in the Marine-TM metagenomes, using BiG-MEx BGC class abundance prediction module, and additionally, computed the BGC class abundances based on the complete genome sequences used to simulate the Marine-TM metagenomes. Similarly as indicated previously, the latter abundances were taken as a reference to evaluate the accuracy of the predictions. We observed that the predicted vs. reference abundance comparison for 20 of the 23 BGC classes we modeled (i.e., the total number of classes detected in the Marine-RM training dataset) had a Pearson’s r correlation coefficient greater than 0.5 and a Median Absolute Percentage Error (MdAPE) lower than 25% (Supplementary Fig. 6.1-2). Figure 2.2-2a displays the
scatter plots of this comparison for the NRPS, terpene, and type I and II PKS BGC classes. To benchmark BiG-MEx BGC class abundance prediction module, we compared its abundance predictions against the abundance estimates derived from running antiSMASH on assemblies of the Marine-TM metagenomes (hereafter referred to as the “assembly approach”). The plots in Figure 2.2-2b display the Pearson correlation coefficients and the Absolute Percentage Error (APE) distributions with respect to the reference abundances comparing both approaches for the same four BGC classes mentioned above. All BGC class abundance models included in this analysis were considerably more accurate than the assembly approach (Supplementary Fig. 6.1-3).

To illustrate the application of BiG-MEx, we performed a Principal Coordinates Analysis (PCoA) based on BiG-MEx-derived BGC class abundance profiles of the 139 prokaryotic metagenomes of TARA Oceans (Karsenti et al. 2011). In Figure 2.2-3a, we ordinate the first two axes of the PCoA. The first axis (PCo1; 73.5% of the total variance) differentiated the mesopelagic (MES) from the surface (SRF) and deep chlorophyll maximum (DCM) water layers (Wilcoxon rank sum test; all p-values < 0.0001; see Supplementary Table 6.1-2). Further, the ordination values of the metagenomes along the PCo1 axis correlated with temperature (Pearson’s r = -0.73; p-value < 0.0001). The differences in the BGC class composition between water layers were additionally confirmed with a Permutational Multivariate Analysis of Variance (PERMANOVA) (see Supplementary Table 6.1-3). We also performed a PCoA to explore the BGC domain composition and obtained a similar ordination of the metagenomes (Supplementary Fig. 6.1-4). These results are in agreement with previous work showing the stratification of microbial communities along depth and temperature gradients (Sunagawa et al. 2015; Walsh et al. 2015). In particular, a very similar differentiation of the MES water layer along the first axis was also observed in the PCoA performed by Sunagawa et al. (2015) based on the 16S rDNA (i.e., mtags (Logares et al. 2014)) taxonomic composition of these same TARA Oceans metagenomes.

As shown in Supplementary Figures 6.1-5 and 6.1-6, the SRF and DCM water layers appear dominated by terpenes. Such pattern is likely to derive from the fact that
these water layers contain a relatively high abundance of carotenoid pigments involved in light harvesting and oxidative stress protection processes (Latifi, Ruiz, and Zhang 2009). Conversely, the MES layer has a more diverse BGC class composition: besides terpene, other BGC classes like bacteriocin, phosphonate, PKS, and NRPS also have a high abundance. The chemical warfare that occurs in densely populated and highly competitive microenvironments, which are more frequently found in deeper water layers (e.g., particle attached, or insensibly exploited (transient) nutrient-rich microenvironments), is likely to explain this latter observation (Lauro et al. 2009).

Next, we used BiG-MEx domain-based diversity module to compare the Operational Domain Unit (ODU) diversity of the NRPS adenylation (AMP-binding) and condensation domains between the SRF, DCM and MES water layers. These domains provide information about the chemical characteristics of the peptides synthesized by NRPS enzymes. AMP-binding domains recruit the amino acid monomers to be incorporated, while condensation domains catalyze the peptide bond formation (Rausch et al. 2005; Rausch et al. 2007). In this analysis, we aimed to assess the potential chemical diversity of the NRPS products. NRPSs are one of the most studied BGC classes and are responsible for the production of many compounds with clinical applications. The results show that the ODU diversity of both domains increased from the surface to the mesopelagic water layers and differentiated significantly between water layers (pairwise Wilcoxon rank sum test; all p-values < 0.005; see Supplementary Table 6.1-2) (Fig. 2.2-3b). These results indicate that the microbial communities inhabiting deeper water layers contain a significantly higher diversity of NRPS products. The ODU diversity gradients resemble the Operational Taxonomic Unit (OTU) richness and functional diversity distributions shown in Sunagawa et al. (2015). We found highly significant correlations between the ODU diversity estimates and the taxonomic and functional richness and diversity obtained by Sunagawa and colleagues (see Supplementary Table 6.1-4).

To exemplify a more fine-grained analysis with BiG-MEx’s domain-based diversity module, we explored the ODU diversity of condensation domains in the three TARA Oceans
metagenomes obtained from the SRF, DCM, and MES water layers at the sampling station TARA_085 (Antarctic Ocean). As observed previously, the metagenome from the MES water layer had a higher ODU diversity (Fig. 2.2-4a). It contains many low abundance ODUs scattered throughout the reference phylogeny (Fig. 2.2-4b). The phylogenetic diversity (Faith 1992) (PD) of ODU representative sequences of the MES metagenome, was 5.24 and 2.65 times greater than the PD estimates of the SRF and DCM metagenomes, respectively. Besides indicating a higher chemical diversity, this result indicates that there is greater potential chemical novelty of nonribosomal peptides. Additionally, the phylogenetic placement analysis revealed that the most abundant condensation ODU is placed close to the reference condensation domain sequences of NRPSs that produce albicidin and cystobactamide antibiotics (both topoisomerase inhibitors) (Fig. 2.2-4c). As albicidin is also a phytotoxin, the dominance of such ODU, which originates from the DCM layer, could be explained by the presence of a large number of NRPSs that act on the photosynthetic organisms that concentrate therein. The DCM layer had a notably higher chlorophyll concentration than the other two layers (0.01, 0.28, and 0 mg/m3 for the SRF, DCM, and MES respectively). The NRPS producing albicidin belongs to the class **Gammaproteobacteria** and order **Xanthomonadales**. This is in agreement with the ODU taxonomic affiliation, which was annotated as a **Gammaproteobacteria** (lowest common ancestor). This finding is also supported by the fact that the BLASTP search against the reference MIBiG database, showed that condensation domains significantly similar to NRPS domains producing albicidin (e-value < 1e-5), where only found in the DCM layer. We cannot exclude other possible explanations of these results; however, this line of exploration might be worth considering for further research. Rising ocean temperatures, as a consequence of global warming, are predicted to increase the frequency of events of bacteria affecting the algae populations, which in turn can impact marine ecosystems on a global scale (Mayers et al. 2016). Regarding potential biotechnological applications, these results are relevant for bioprospecting, given that albicidin and cystobactamide are antibiotics of interest for clinical treatments (Baumann et al. 2014; Cociancich et al. 2015).
Chapter 2. Results and Discussion

We note that neither the TARA Oceans Metagenomes Assembled Genomes (MAGs) (Delmont et al. 2018), nor the DCM assembled metagenome from TARA_085 sampling site, contained albicidin or cystobactamide NRPS-like sequences. The difference between our findings in comparison to standard approaches based on assembled data was expected to occur, given the limitations of the latter to identify BGC classes (as shown in Fig. 2.2-2b). In Supplementary Figure 6.1-7, we illustrate this problem by comparing the sequence length between MIBiG BGCs, and the TARA Oceans MAGs, and assembled metagenomic contigs.

Considering the relevance of human microbiome-derived natural product BGCs in medical research, we demonstrate the applicability of BiG-MEx to explore the BGC composition in the Human Microbiome Project (HMP) dataset (The Human Microbiome Project Consortium 2012b). Our analyses traversed metagenomes from the buccal mucosa, tongue dorsum, and supragingival plaque body sites as well as stool samples (491 metagenomes in total). We used BiG-MEx to compute the BGC domain and class abundance profiles, and applied the same methodology as described for TARA Oceans, to compute the domain and class-based PCoAs. In agreement with previous analyses based on the taxonomic and functional annotation (The Human Microbiome Project Consortium 2012b; Segata et al. 2012), we observed that metagenomes grouped according to the body site they were sampled from in the first two ordination axes (Supplementary Fig. 6.1-8a and b). We conducted a PERMANOVA to test and assess the strength of the differences between body sites according to their BGC class composition, which showed significant differences in all body site comparisons (Supplementary Table 6.1-5). Additionally, we used BiG-MEx to compare the ODU diversity of the AMP-binding and condensation domains between body sites and observed that supragingival plaque metagenomes contain significantly higher diversity than the other body sites (pairwise Wilcoxon rank sum test; p-value < 0.0001) (Supplementary Fig. 6.1-9 and Supplementary Table 6.1-6). This is in line with previous work showing that the supragingival plaque is one of the most functionally and taxonomically diverse body sites in the HMP dataset (The Human Microbiome Project Consortium 2012b).
Besides the mining analyses, BiG-MEx BGC class profiling can be used for the screening and prioritization of (meta)genomic samples. BGC class abundance profiles derived from shallow sequencing depth (meta)genomic data can be used for the identification of strains or environments with high biosynthetic potential, before investing in deep sequencing or long read sequencing technologies. As a proof-of-concept for this application, in Supplementary Figure 6.1-10 we show a comparison of the BGC class abundance predictions computed in metagenomes of 100 and 5 million reads.

**Conclusions**

In our example applications, we processed 630 metagenomes, which sum to more than 85 billion paired-end reads. The analyses showed that BiG-MEx ultrafast domain and class profiling, and ODU diversity estimates provide biologically meaningful information, which can be used to mine BGCs in metagenomic data and as a basis from which to assess the ecological roles of their products in specific environments. BiG-MEx extends BGC-based research and exploitation into large environmental datasets. It can be used to study the biogeography, distribution, and diversity of natural product BGCs either at the class, domain or ODU levels. Such analyses have the potential to accelerate the discovery of new bioactive products (Medema and Fischbach 2015).
Materials and Methods

1. Data acquisition, pre-processing and annotation

We retrieved the 139 prokaryotic metagenomes of the TARA Oceans dataset from the European Nucleotide Archive (Harrison et al. 2019) (ENA:PRJEB1787, filter size: 0.22-1.6 and 0.22-3). To pre-process the metagenomic short-read data, we clipped the adapter sequences (obtained from Shinichi Sunagawa personal communication, July 21, 2015) using the BBduk tool from the BBMap 35.00 (Bushnell 2014) suite with a maximum Hamming distance of one (hdist=1). We then merged the paired-end reads using VSEARCH 2.3.4 (Rognes et al. 2016), quality trimmed all reads at Q20 and filtered out sequences shorter than 45 bp using BBduk, and de-replicated the resulting quality-controlled sequences with VSEARCH. We annotated the BGC domains by first predicting the Open Reading Frames (ORFs) in the pre-processed data with FragGeneScan-plus (Kim et al. 2015) and then running BiG-MEx on the predicted ORF’s amino acid sequences.

We downloaded 491 human microbiome metagenomes from the Data Analysis and Coordination Center (DACC) for the Human Microbiome Project (HMP) (https://www.hmpdacc.org/hmp/HMASM). Our dataset included the metagenomes of the supragingival plaque (118), tongue dorsum (128), buccal mucosa (107), and the stool (138) body sites. These metagenomes have been already pre-processed as described in The Human Microbiome Project Consortium (2012a). The additional pre-processing tasks we performed consisted of merging the metagenomic reads with VSEARCH, quality trimming all reads at Q20 and filtering out sequences shorter than 45 bp with BBduk. To annotate the BGC domains, we predicted the ORFs with FragGeneScan-plus and ran BiG-MEx BGC domain identification module on the ORF’s amino acid sequences (Supplementary Table 6.1-7).

2. Exploratory analysis performed on TARA Oceans and HMP datasets

The domain abundance profiles of the TARA Oceans and HMP metagenomes
were used to predict the BGC class abundance profiles with BiG-MEx BGC class abundance prediction module. The models used to generate the predictions for the TARA Oceans, and the oral and stool HMP metagenomes, were trained with the Marine-RM, Human-Oral and Human-Stool simulated metagenomic datasets, respectively. For each dataset, we performed a Principal Coordinate Analysis (PCoA) as follows: 1) We applied a total sum scaling standardization to both the domain and class abundance matrices; 2) We used the standardized matrices to compute the domain and class Bray-Curtis dissimilarity matrices; 3) We performed the PCoAs on the dissimilarity matrices with vegan R package utilizing the function capscale (Oksanen et al. 2017). We applied a Permutational Multivariate Analysis of Variance (PERMANOVA) (Anderson 2001) to quantify the strength and test the differences between water layers and body sites according to their BGC class composition. For these analyses, we selected a balanced subset of metagenomes from the TARA Oceans and HMP datasets (63 and 216 metagenomes, respectively; see below).

We performed a PERMANOVA on the Bray-Curtis dissimilarity matrix, computed for the TARA Oceans and HMP metagenome subsets as described above, to test the differentiation between all groups simultaneously. Subsequently, we tested each pair of groups independently, applying the Bonferroni correction for multiple comparisons. To perform the PERMANOVA, we employed the adonis function of the vegan R package, with the permutation parameter set to 999.

To compare the domain ODU diversity of the NRPS adenylation (AMP-binding) and condensation domains between the surface (SRF), deep chlorophyll maximum (DCM) and mesopelagic (MES) water layers, we used a subset of 63 TARA Oceans metagenomes, representing the three water layers in 21 sampling stations. We computed the ODU Shannon diversity in these metagenomes, using routines implemented in the BiG-MEx domain-based diversity module. Additionally, we used the same BiG-MEx module to examine the diversity of the condensation domains in the metagenomes representing the three water layers at
sampling station TARA_085.

To perform the ODU taxonomy annotation, we used MMseqs2 taxonomy assignment function (Burke et al. 2011) based on UniRef100 (Suzek et al. 2015) sequences (release-2018_08), with the e-value and sensitivity parameters set to 0.75 and 0.01, respectively.

To compare the AMP-binding and condensation ODU diversity between body sites, we applied a similar approach as described above. We selected a subset of 216 metagenomes, 54 from each of the supragingival plaque, tongue dorsum, buccal mucosa, and stool body sites. This subset includes only the metagenomes obtained from individuals for whom the four body sites were sampled. We applied BiG-MEx domain-based diversity module to compute the ODU Shannon diversity estimates.

The Wilcoxon rank-sum tests (two-sided) to assess the significance of the differentiations between metagenomes from different groups (i.e., water layers or body sites), were performed with the wilcox.test function from the R package stats (R Core Team 2016).

3. Data simulation, pre-processing and annotation

(a) Construction of simulated metagenomic datasets

We created four simulated metagenomic datasets: Two of these approximate the taxonomic composition found in marine environments (Marine-RM and Marine-TM), and the other two, the taxonomic composition found in the human oral cavity and stool body sites (Human-Oral and Human-Stool, respectively). Each dataset is composed of 150 metagenomes, all of which have a size of two million paired-end reads. To simulate a metagenomic dataset, we first created a dataset of reference genome sequences and the genome abundance profiles to specify the metagenomes’ taxonomic composition. That is, we defined a hypothetical microbial community from which a metagenome is simulated by specifying which reference genomes and the
number of times each genome occurs in the community.

To create the Marine-RM (Marine Reference Microbiome) genome dataset, we downloaded all genomes belonging to the Ocean Microbial Reference Gene Catalogue (OM-RGC) (Sunagawa et al. 2015) having an assembly status of “Complete genome” from RefSeq (O’Leary et al. 2016) (on December 7, 2017). If a given species did not have a complete genome sequence available, we randomly selected another species of the same genus. In total, we obtained 378 genomes corresponding to 363 species.

We applied a similar methodology to create the Marine-TM (Marine TARA Microbiome) genome dataset. To determine the taxonomic composition, we used the genus affiliation of the 16S rDNA (i.e., miTags (Logares et al. 2014)) Operational Taxonomic Units (OTUs) of TARA Oceans (Sunagawa et al. 2015). We only included 30 shared genera (randomly selected) between TARA OTUs and the Marine-TM genome dataset. This latter filtering was necessary to reduce the taxonomic overlap, given that we used the Marine-TM dataset to evaluate the performance of the BGC class abundance models trained with the Marine-RM dataset (see section 4(c)). For the remaining genera for which there was at least one representative completely sequenced genome, we downloaded a maximum of three genomes per genus from RefSeq, irrespective of their species affiliation. This resulted in a database composed of 344 genomes from 308 species.

To create the genome datasets for the Human-Oral and Human-Stool metagenomic datasets, we used the genomes sequenced by the HMP derived from samples of the oral cavity and stool body sites. Given that few of these genomes were completely sequenced, we also included partially complete sequenced genomes. We downloaded all genomes with an assembly status of “Complete genome” or “Chromosome” or “Scaffold” generated by the HMP from the GenBank database (Benson et al. 2017) (on March 15, 2018). In the cases where a genome (sequenced by the HMP) had an as-
assembly status lower than “Scaffold”, we downloaded another genome with the same species affiliation and an assembly status of “Complete genome” or “Chromosome”. The Human-Oral and Human-Stool reference genome datasets contain 209, and 479 genomes representing 140 and 338 species, respectively.

To create the community abundance profile of a metagenomic dataset, we randomly selected between 20 and 80 genomes from its genome reference dataset and defined the number of times each genome occurs by sampling from a lognormal distribution with mean 1 and standard deviation of 0.5. Lastly, we simulated the metagenomes with MetaSim v0.9.5 (Richter et al. 2011). MetaSim was set to generate paired-end reads with a length of 101 bp, and a substitution rate increasing constantly along each read from $1 \times 10^{-4}$ to $9.9 \times 10^{-2}$. With this data, we aimed to simulate the short-read sequences generated by an Illumina HiSeq 2000 platform. Dataset statistics are shown in Supplementary Table 6.1-1. The assembly accessions, organism names, and taxids of the genome sequences used to create the genome databases are found at https://github.com/pereiramemo/BiG-MEx/raw/master/data/supplementary_file1.xlsx. The workflow used to create the simulated metagenomic datasets can be found at https://github.com/pereiramemo/BiG-MEx/wiki/Data-simulation.

(b) Annotation of the simulated metagenomes

To estimate the reference BGC class abundances in a simulated metagenome, we annotated the BGC classes in its reference genome sequences with antiSMASH 3.0, mapped the paired-end reads to the identified BGC sequences with BWA-MEM 0.7.12 (Li 2013), and filtered out read alignments with a quality score lower than 10. Next, we removed read duplicates with Picard tools v1.133 (http://broadinstitute.github.io/picard), and computed the mean coverage with BEDTools v2.23 (Quinlan and Hall 2010). The coverage estimates were assumed
to accurately reflect the BGC class coverage-based abundances, as they were computed using complete BGC sequences, obtained from the genome sequences used to simulate the metagenomes. Additionally, we merged the paired-end reads of the simulated metagenomes with VSEARCH 2.3.4, predicted the ORFs with FragGeneScan-plus, and used BiG-MEx domain identification module to annotate the BGC domains in the ORF's amino acid sequences. The workflow to annotate the synthetic metagenomes can be found at https://github.com/pereiramemo/BiG-MEx/wiki/Data-simulation#7-bgc-domain-annotation.

4. Performance evaluation

(a) BGC domain identification module
We compared the running time (wall-clock) of UProC (i.e., uproc-prot) against a typical search using hmmsearch from the HMMER3 package (Eddy 2011), for the identification of the 150 BGC domains included in BiG-MEx, in nine prokaryotic metagenomes of the TARA Oceans dataset (Supplementary Table 6.1-8). To run hmmsearch, we used the domain HMM profiles of antiSMASH. We annotated the nine metagenomes with both these tools in four independent rounds, each round using a different thread number (i.e., 4, 8, 16 and 32 threads). All parameters of uproc-prot and hmmsearch were set to default. The annotations were carried out on a workstation with Intel(R) Xeon(R) CPU E7-4820 v4 2.00GHz processors.

(b) BGC domain-based diversity analysis module
We evaluated BiG-MEx Operation Domain Unit (ODU) diversity estimation approach using NRPS adenylation (AMP-binding) and condensation, and PKS ketosynthase and acyltransferase domains (KS and AT, respectively). In this analysis, we used the BGC domain-based diversity analysis module to compute the ODU diversity in the Marine-TM dataset, and compared these estimates with the ODU diversity computed using the complete do-
main sequences. To obtain the latter ODU diversity, we applied the workflow implemented in BiG-MEx, with the exception that instead of assembling the domain sequences, we extracted these from the complete genome sequences used to simulate the Marine-TM metagenomes. We annotated the four domains in the complete genome sequences with hmmsearch using the antiSMASH HMM profiles.

(c) BGC class abundance predictions
We used the BGC class models trained with the Marine-RM metagenomic dataset to predict the BGC class abundances in the Marine-TM metagenomic dataset. We applied the methodology described in section 3(b) to compute the BGC class abundances in the Marine-TM metagenomes based on the complete genome sequences (i.e., reference abundance). To predict the BGC class abundances using machine-learning models, we annotated the Marine-TM metagenomes with the BiG-MEx domain identification module and used the domain abundance profiles as an input for the BiG-MEx BGC class abundance prediction module. The evaluation consisted of computing the Pearson correlation and Median Absolute Percentage Error (MdAPE) between the predicted and reference BGC class abundances. The Absolute Percentage Error (APE) was computed as \[ \frac{100 \times |\hat{A} - A|}{\hat{A}} \], where \( \hat{A} \) and \( A \) are the predicted and reference abundance, respectively. To benchmark the machine-learning models, we compared the BGC class abundance predictions against the abundance estimates based on the assembly of 50 metagenomes of the Marine-TM dataset (assembly approach). The assembly approach consisted of assembling the metagenomes with MEGAHIT v1.0 (Li et al. 2016) (default parameters), running BiG-MEx domain identification module to select the contigs with potential BGC sequences, annotating the selected contigs with antiSMASH 3.0, and estimating the BGC class abundance following the same approach as described in section 3(b) (Supplementary Table 6.1-9). We computed the Pearson correlation coefficient.
and APE of BGC class abundance estimates obtained by the assembly approach and predicted by BiG-MEx, with respect to the reference BGC class abundances. The analysis performed to evaluate the accuracy of the models can be reproduced here: https://rawgit.com/pereiramemo/BiG-MEx/master/machine_leaRning/bgcpred_workflow.html.

(d) Evaluation of the BGC class abundance predictions in shallow metagenomes

We selected 30 merged pre-processed TARA Oceans metagenomes and randomly subsampled these to generate two sets of metagenomes, one with 100 million and the other with 5 million reads, using the seqtk v1.0 tool (https://github.com/lh3/seqtk). We then annotated the BGC domains and predicted the BGC class abundances in this data using BiG-MEx (as described in sections 1 and 2), and compared the BGC class abundance predictions between the two sets of metagenomes.

5. BiG-MEx implementation

(a) BGC domain identification module

BiG-MEx BGC domain identification module uses the UProC 1.2.0 (Meinicke 2015) software to classify short-read sequences using BGC domain references. To train UProC for this purpose, we manually curated all amino acid sequences matching 150 antiSMASH hidden Markov model profiles (HMMs) (Weber et al. 2015). In this task, we removed sequences shorter than 25 amino acids and checked for the presence of overlaps between sequences of different HMM profiles. In addition, we categorized multi-domain proteins into multiple families. For the training process, we included a set of negative control profiles to assess the ratio of false positive hits. Namely, we used the t2fas, fabH, bt1fas, ft1fas profiles as negative controls for the PKS_KS, t2ks, t2ks2, t2clf, Chal_sti_synt_N, Chal_sti_synt_C, hglD and hglE profiles. Once we curated the amino
acid sequence data, we applied the SEG(mentation) low complexity filter from the NCBI Blast+ 2.2 Suite (Camacho et al. 2009) and created the UProC database. This UProC database can be downloaded from https://github.com/pereiramemo/BiG-MEx. Based on the identified reads containing a BGC domain sequence, the module computes a count-based abundance profile of BGC domains.

(b) BGC domain-based diversity analysis module

This module performs two different analyses: Operational Domain Unit (ODU) diversity estimation and phylogenetic placement of domain sequences. The pipeline to estimate the ODU diversity, analyses each domain independently, and consists of the following steps: 1) Short-read sequences, where the domain being studied was identified, are recruited to perform a targeted assembly metaSPAdes 3.11 (Nurk et al. 2017) with default parameters; 2) The Open Reading Frames (ORFs) in the resulting contigs are predicted with FragGeneScan-Plus; 3) Domain sequences are identified within the ORF amino acid sequences with hmmsearch from HMMER v3 and extracted; 4) Domain amino acid sequences are clustered into ODUs using MMseqs2 (Hauser, Steinegger, and Söding 2016) with the cascaded clustering option and the sensitivity parameter set to 7.5; 5) Annotated unassembled reads are mapped to the domain nucleotide sequences with BWA-MEM 0.7.12, and the mean depth coverage is calculated using BEDTools v2.23; 6) Based on this information, the coverage-based abundance of the ODUs is computed and used to estimate an ODU alpha Shannon diversity. To allow a comparison of the ODU diversity estimates between samples with different sequencing depth, we include an option to estimate the diversity for rarefied subsamples.

To perform the phylogenetic placement of domain sequences, we applied an approach similar to NaPDoS (Ziemert et al. 2012). However, we extended the phylogenetic placement analysis to 48 domains and included
more comprehensive reference trees, which are critical for the analysis of large metagenomic samples. In detail, the phylogenetic placement consists of aligning the target domain sequences to their corresponding reference multiple sequence alignment (MSA) with MAFFT (Yamada, Tomii, and Katoh 2016) (using –add option). Subsequently, the extended MSA together with its reference tree are used as the input to run pplacer (Matsen, Kodner, and Armbrust 2010) (with parameters: –keep-at-most 10 and –discard-nonoverlapped; all other parameters set to default). pplacer performs the phylogenetic placement using the maximum-likelihood criteria and outputs the extended tree in Newick and jplace formats (Matsen et al. 2012), and a table with statistics and information about the placement of each sequence (i.e., likelihood, posterior probability, expected distance between placement locations (EDPL), pendant length, and edge number). To visualize the phylogenetic placement, a tree figure is generated using the ggtree R package (Yu et al. 2017), where the coverage of the placed sequences is mapped on their tree tips and used to scale a bubble representation. Besides the phylogenetic placement, we included in this module an option to perform a BLASTP search of the assembled domain sequences against the reference domain sequences.

To construct the reference phylogenies, we first downloaded all the BGC amino acid sequences from the MIBiG database (Medema et al. 2015). We identified the domain sequences with hmmsearch using the BGC domain HMM profiles from antiSMASH. Subsequently, we extracted and clustered these sequences with MMseqs2 to create a non-redundant dataset of amino acid sequences for each domain. If the number of reference sequences identified in the MIBiG database was greater than 500, we used a clustering threshold of 0.7 identity at the amino acid level; otherwise, the threshold was set to 0.9; all other parameters of MMseqs2 were set as specified previously. All domains with less than 20 representative sequences were
discarded. This resulted in a subset of 48 domains that were considered for the phylogenetic reconstructions. For each set of domain representative sequences, we generated an MSA with MAFFT using the E-INS-I algorithm, removed sequence outliers with OD-seq (Jehl, Sievers, and Higgins 2015) and constructed a phylogenetic tree with RAxML (Stamatakis 2014). To select the protein evolutionary model for the phylogenetic reconstruction, we used the automatic model selection implemented in RAxML with the maximum likelihood criterion. We used the GAMMA model of rate heterogeneity and searched the tree space using the rapid hill-climbing algorithm (Stamatakis et al. 2007), starting from a maximum parsimony tree. For the sake of reproducibility, we specified a random seed number (i.e., -p 12345). Finally, we used RAxML to root the trees and compute the SH-like support scores (Guindon et al. 2010). In Supplementary Table 6.1-10, we provide for each domain phylogeny the number of sequences and amino acid substitution model used, the mean, standard deviation, maximum and minimum cophenetic distances between sequences, Faith’s phylogenetic diversity (Faith 1992) and the name of its corresponding BGC class.

(c) BGC class abundance prediction module

BiG-MEx uses machine-learning models to predict the abundance of the BGC classes, based on the counts of annotated domains in unassembled metagenomes. Each model is class-specific and was trained using the abundance of the BGC class and its corresponding protein domains, as the response and predictor variables, respectively. We used the classification rules defined in antiSMASH for the annotation of BGC classes, to determine the protein domains used as predictor variables in each model. To model the abundance of a given BGC class, we implemented a two-step zero-inflated process. First, the presence or absence of the target BGC class is predicted using a random forest (RF) binary classifier (Breiman 2001). Second, a multiple linear regression (MLR) is applied to predict the class
abundance, but only if the class was previously predicted as present. In the cases where the number of zero values was lower than 10 or non-existent, we directly applied an MLR. We trained the models using simulated metagenomic data (i.e., Marine-RM, Human-Oral and Human-Stool datasets). The models predict a coverage-based abundance since this was the response variable used in the training process. The RF binary classification models were created with the randomForest function of the randomForest R package (Liaw and Wiener 2002), with the parameters ntree set to 1000 (number of trees grown), nodesize set to 10 (minimum size of terminal nodes), and mtry set to 1 (number of variables randomly sampled as candidates at each split). For the MLR, we used the lm function of the stats R package with default parameters.
Chapter 2. Results and Discussion

Code availability

BiG-MEx is freely distributed using Docker container technology (www.docker.com), under the GNU General Public License v3.0. It can be downloaded from https://github.com/pereiramemo/BiG-MEx, where we also provide thorough documentation. Currently, we provide BGC class abundance models targeting the marine environment, four different human body sites, and the genus Streptomyces. To help users create their own BGC class abundance models and compute the predictions, we developed the R package bgcpred: https://github.com/pereiramemo/bgcpred. bgcpred is integrated in BiG-MEx, and is used to generate the BGC class abundance predictions.

In addition, we provide a tutorial on how to run BiG-MEx at https://github.com/pereiramemo/BiG-MEx/wiki/Getting-started and environment specific BGC class abundance models at https://github.com/pereiramemo/BiG-MEx/wiki/Models-download-and-evaluation. Also, at https://github.com/pereiramemo/BiG-MEx/wiki/Data-simulation we show the code used to generate the simulated metagenomic datasets. Lastly, we provide an example code on how to train and test the BGC class abundance models at https://rawgit.com/pereiramemo/BiG-MEx/master/machine_learning/bgcpred_workflow.html.

Data availability

At https://github.com/pereiramemo/BiG-MEx we provide the GenBank and RefSeq assembly accessions of the genomes used to generate the simulated metagenomic datasets, and the BGC domain and class abundance tables, obtained from the simulated data.

Acknowledgements

EP-F was supported by a Ph.D. fellowship of the German Academic Exchange Service (DAAD) and the Uruguayan National Research and Innovation Agency (ANII). PLB
was supported by the HGF Infrastructure Programme FRAM of the Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung. MM was supported by Rubicon (825.13.001) and Veni (863.15.002) grants from the Netherlands Organization for Scientific Research (NWO). P-M was partially supported by German Research Foundation (DFG). AF-G received funding from the European Union’s Horizon 2020 research and innovation program [Blue Growth: Unlocking the potential of Seas and Oceans] under grant agreement no. [634486] 436 (project acronym INMARE). This work was accomplished using computational facilities provided by the Max Planck Society.
References


Kim, Dongjae et al. (2015). “FragGeneScan-plus for Scalable High-Throughput Short-Read Open Reading Frame Prediction”. In: Computational Intelligence in Bioinfor-
Chapter 2. Results and Discussion


Logares, Ramiro et al. (2014). “Metagenomic 16S rDNA Illumina Tags Are a Powerful Alternative to Amplicon Sequencing to Explore Diversity and Structure of Microbial Communities”. In: *Environmental Microbiology* 16 (9), pp. 2659–71. DOI: 10.1111/1462-2920.12250.


Chapter 2. Results and Discussion


Segata, Nicola et al. (2012). “Composition of the Adult Digestive Tract Bacterial Microbiome Based on Seven Mouth Surfaces, Tonsils, Throat and Stool Samples”. In: *Genome Biology* 13 (6), R42. DOI: 10.1186/gb-2012-13-6-r42.


Walsh, Emily A. et al. (2015). “Bacterial Diversity and Community Composition from Seasurface to Subseafloor”. In: *The ISME Journal* 10 (4), pp. 979–89. DOI: 10.1038/ismej.2015.175.


Ziemert, Nadine et al. (2014). “Diversity and Evolution of Secondary Metabolism in the Marine Actinomycete Genus Salinispora”. In: *Proceedings of the National Academy
of Sciences of the United States of America 111 (12), E1130–9. DOI: 10.1073/pnas.1324161111.
Chapter 2. Results and Discussion

Figures

Figure 2.2-1: BiG-MEx analysis workflow. 1) BGC domain identification module. To annotate the BGC domains with UProC, we created an UProC database including 150 domains, which originate from 44 different BGC classes. This database was generated based on the amino acid sequences of antiSMASH hidden Markov model (HMM) profiles (Weber et al. 2015). Using UProC output, this module generates a count-based abundance profile of BGC domains; 2) BGC domain-based diversity analysis module. Using the previously identified domains, this module performs a targeted assembly with metaSPAdes (Nurk et al. 2017) to reconstruct the domain sequences. Assembled domain sequences are clustered into Operational Domain Units, and the number of ODU and the coverage of the domain sequences within each ODU (used to approximate the abundance of the ODU) are used to compute the ODU alpha diversity. The environmental reconstructed domain sequences are placed onto reference phylogenetic trees with pplacer (Matsen, Kodner, and Armbrust 2010) (maximum likelihood criteria). In this module, we include pre-computed phylogenies for 48 domains, which are based on sequence data contained in the Minimum Information about a Biosynthetic Gene cluster (MIBiG) (Medema et al. 2015) database, allowing us to identify the relationships of query sequences with domains from pathways of known function; 3) BGC class abundance prediction module. The domain abundance profiles are used to predict the BGC class coverage-based abundance profiles using class-specific machine-learning models. These models consist of a two-step process: First, the presence/absence of the BGC class is predicted using a random forest (RF) classifier; Secondly, the abundance is predicted with a multiple linear regression (MLR) only if the class was previously predicted as present.
Figure 2.2-2: Evaluating and benchmarking the BGC abundance prediction models. (a) Scatter plots comparing the reference and predicted abundances of the NRPS, terpene, T1PKS and T2PKS BGC classes. MdAPE: Median Absolute Percentage Error. The black, solid line represents the one-to-one relationship between the reference and predicted BGC class abundances. The BGC class abundance models were trained with the Marine-RM metagenomes and used to predict the abundances in the Marine-TM metagenomes. (b) Plots of the Pearson correlation coefficients (upper panel) and the Absolute Percentage Error (APE) distributions (lower panel) of the BGC class abundances predicted by the models and estimated by the assembly approach, with respect to the reference abundances. In this comparison, we used 50 Marine-TM metagenomes. For the sake of clarity, 12 outlying APE values (3% of the total comparisons) were excluded from the plot. The assembly approach consisted of the following tasks: 1) Assembling the metagenomes of the Marine-TM dataset; 2) Selecting the contigs with potential BGC sequences using BiG-MEx domain identification module; 3) Annotating the contigs with antiSMASH; 4) Mapping the short-read sequences to the identified BGC sequences; 5) Estimating the BGC class abundances.
Figure 2.2-3: BiG-MEx BGC class composition and domain-based diversity analysis in the TARA Oceans dataset. (a) Principal Coordinates Analysis (PCoA) performed on a Bray-Curtis dissimilarity matrix of BGC class relative abundance profiles of the 139 prokaryotic metagenomes of TARA Oceans. BGC class abundance profiles were generated with BiG-MEx BGC class abundance module, using machine-learning models trained with the simulated Marine-RM metagenomic dataset. The abbreviations SRF, DCM, MES, and MIX correspond to surface, deep chlorophyll maximum, mesopelagic, and subsurface epipelagic mixed water layers, respectively. The boxplot in the bottom section of the panel shows the PCo1 value distributions for the metagenomes from the SRF, DCM and MES water layers. The PCo1 axis differentiated the MES water layer from the other two layers (Wilcoxon rank sum test; all p-values < 0.0001). (b) Bar plots showing the distribution of the ODU Shannon alpha diversity indices for the AMP-binding and condensation domains (NRPSs). The ODU diversity was computed for a match subset of 63 TARA Oceans metagenomes representing SRF, DCM, and MES water layers in 21 sampling stations. The AMP-binding and condensation ODU diversity estimates were significantly different between the three water layers (pairwise Wilcoxon rank sum test; all p-values < 0.005).
Figure 2.2-4: BiG-MEx diversity analysis of condensation domains in three metagenomes from TARA Oceans sampling station TARA_085. (a) Rarefaction curves of the Shannon alpha diversity indices generated by BiG-MEx domain-based diversity analysis module, comparing the diversity of condensation ODUs in the metagenomes of the SRF, DCM, and MES water layers. Condensation domain sequences were clustered into ODUs using a 0.75% amino acid identity threshold. The diversity was computed using the number and abundance of distinct condensation ODUs. (b) Phylogenetic placement of the condensation ODU representative sequences, as performed by the BiG-MEx domain-based diversity analysis module. The SRF, DCM and MES had a phylogenetic diversity (Faith’s PD) (Faith 1992) of 58.15, 114.98 and 304.88, respectively. The size and color of the bubbles on the leaves represent the ODU abundance and sample origin, respectively. (c) Detail of the clade contained in the orange, hollow square in (c), including the most abundant ODU (obtained in the TARA_085_DCM_0_22-3 sample; indicated with an orange arrow).
2.3 Fast and accurate average genome size and 16S rRNA gene average copy number computation in metagenomic data
Fast and accurate average genome size and 16S rRNA gene average copy number computation in metagenomic data

Emiliano Pereira-Flores¹,²*, Frank Oliver Glöckner¹,³ and Antonio Fernandez-Guerra¹,²,⁴*

Abstract

Background: Metagenomics caused a quantum leap in microbial ecology. However, the inherent size and complexity of metagenomic data limit its interpretation. The quantification of metagenomic traits in metagenomic analysis workflows has the potential to improve the exploitation of metagenomic data. Metagenomic traits are organisms’ characteristics linked to their performance. They are measured at the genomic level taking a random sample of individuals in a community. As such, these traits provide valuable information to uncover microorganisms’ ecological patterns. The Average Genome Size (AGS) and the 16S rRNA gene Average Copy Number (ACN) are two highly informative metagenomic traits that reflect microorganisms’ ecological strategies as well as the environmental conditions they inhabit.

Results: Here, we present the ags.sh and acn.sh tools, which analytically derive the AGS and ACN metagenomic traits. These tools represent an advance on previous approaches to compute the AGS and ACN traits. Benchmarking shows that ags.sh is up to 11 times faster than state-of-the-art tools dedicated to the estimation AGS. Both ags.sh and acn.sh show comparable or higher accuracy than existing tools used to estimate these traits. To exemplify the applicability of both tools, we analyzed the 139 prokaryotic metagenomes of TARA Oceans and revealed the ecological strategies associated with different water layers.

Conclusion: We took advantage of recent advances in gene annotation to develop the ags.sh and acn.sh tools to combine easy tool usage with fast and accurate performance. Our tools compute the AGS and ACN metagenomic traits on unassembled metagenomes and allow researchers to improve their metagenomic data analysis to gain deeper insights into microorganisms’ ecology. The ags.sh and acn.sh tools are publicly available using Docker container technology at https://github.com/pereiramemo/AGS-and-ACN-tools.

Keywords: Microbial ecology, Metagenomics, Functional traits, Average genome size, 16S rRNA gene average copy number

Background

Advances in high-throughput sequencing technologies have pushed forward metagenomic studies, allowing the generation of massive amounts of data. As a consequence, metagenomics has become crucial to study microorganisms’ ecology [1]. Nonetheless, making sense of the metagenomic data is a complex and computationally intensive task. Commonly, metagenomes consist of many short-read sequences obtained from numerous different species, many of which are unknown.

Functional trait based-analyses offer an opportunity to improve our understanding of microorganisms’ ecology [2–4]. In particular, community functional traits measured at the genome level in a random sample of individuals (i.e., metagenomic traits), can help to uncover ecological patterns in short-read metagenomic data [5]. Functional traits are defined as characteristics of an organism that are linked to its performance, and consequently, influence its ecology and evolution [6]. Previous studies have used metagenomic traits to explain different aspects of microbial ecology, including why microorganisms live
in a particular environment or how they respond to environmental changes [7–9].

The Average Genome Size (AGS) and the 16S rRNA gene Average Copy Number (ACN) are two metagenomic traits that can be computed from unassembled metagenomic data and provide valuable information to study the ecology of microbes. The genome size is known to be associated with environmental complexity and the organisms’ lifestyle [9–11]. Larger genomes tend to contain a more diverse metabolic repertoire, which in turn allows organisms to metabolize a greater diversity of substrates and inhabit heterogeneous environments [12]. Further, the AGS of a metagenome is important from a statistical perspective: the larger the AGS, the lower is the probability of sampling a specific gene. Hence, in order to avoid potential biases, this trait should be taken into account in gene-centric comparative metagenomics [13]. Lastly, the AGS can be used to estimate the proportion of an average-sized genome that has been sequenced to exhaustion, which can help to determine an appropriate sequencing depth, in particular when the aim is to generate metagenome-assembled genomes (MAGs) [14]. On the other hand, the 16S rRNA gene average copy number provides additional insights into the ecology of microorganisms. The 16S rRNA gene copy number in prokaryote genomes is known to vary from 1 to 15 [15]. This trait is associated with different growth strategies: organisms with low copy numbers tend to utilize resources more efficiently and inhabit oligotrophic environments, while those with high copy numbers can grow more rapidly under favorable conditions [16–18].

Currently, there are two publicly available tools dedicated to the computation of the AGS in metagenomes: the Genome relative Abundance and Average Size (GAAS) [19] and MicrobeCensus [20]. GAAS computes the AGS based on a BLAST search [21] against a reference database of microbial genomes. It was the first tool developed for the computation of the AGS, and although useful at its time, the runtime renders it highly impractical due to the now available large volume of metagenomes. Also, the fact that GAAS relies on genome databases to estimate the AGS, limits its accuracy when analyzing metagenomic samples containing novel taxa [20]. Alternatively, MicrobeCensus computes the AGS based on the abundance information of 30 universally distributed single-copy genes, following an approach initially proposed by Raes [22]; the AGS is estimated based on the abundance-weighted average of these marker genes, using optimized gene weights and empirically determined proportionality constants. Although MicrobeCensus has been shown to be considerably more accurate and faster than GAAS, the rapid increase of data generated by high-throughput sequencing technologies can still challenge its applicability.

For the ACN estimation, there are three publicly available tools able to predict this trait in metagenomic and amplicon data (i.e., PICRUSt [23], CopyRighter [24] and PAPRICA [25]). The approaches implemented in these tools are based on the work of Kembel et al. (2012) [26], which showed that the 16S rRNA gene copy number can be predicted based on the phylogenetic relationships of environmental sequences to reference organisms with known gene copy numbers. Although these tools can be used to estimate the ACN, their objective is to correct for copy number counts when estimating organisms’ abundances. They comprise a series of computationally intensive tasks, and their accuracy has been shown to be limited when analyzing taxa for which there are no close representatives in the reference phylogenies [27].

In this work, we developed two tools, which analytically derive the Average Genome Size (AGS) and 16S rRNA gene Average Copy Number (ACN) in prokaryotic metagenomes (ags.sh and acn.sh, respectively). Our implementations exploit recent advances in gene annotation algorithms to make methodological improvements for the estimation of these traits. We show that the ags.sh and acn.sh tools can rapidly and accurately predict the AGS and ACN, respectively. Compared to other tools used to estimate these traits, ags.sh and acn.sh represent an improvement in terms of accuracy and computational speed. Lastly, we analyzed the AGS and ACN in the TARA Oceans dataset [28], where we demonstrate the applicability of our tools and the value of these traits to reveal the ecological strategies adopted by microbial communities to cope with different environmental conditions.

**Materials and methods**

**Implementation**

The ags.sh and acn.sh tools were written in AWK, Bash, and R, and are provided as command line applications.

**Average genome size computation tool (ags.sh)**

ags.sh computes the Average Genome Size (AGS) and Number of Genomes (NGs) in metagenomic samples, based on the annotation of 35 single-copy genes that are universally present in prokaryotes [22] (Additional file 1). The workflow of ags.sh consists of the following steps: 1) Short-read sequences are filtered by length and trimmed with BBDuk [29] (optional step); 2) Open Reading Frames (ORFs) are predicted in the short-read sequences with FragGeneScan-Plus [30, 31] (optional step); 3) Single-copy genes are annotated with UProC [32]; 4) The gene coverage is estimated as the total number of annotated base pairs divided by the gene length; 5) The NGs is computed as the mean coverage of the 35 single-copy genes (see Eq. 1); 6) The AGS is computed as the ratio of the total number of base pairs to the NGs (see Eq. 2).
\[ \text{NGs} = \frac{1}{35} \sum_{i=1}^{35} \frac{\text{gene_bp}_i}{\text{gene_length}_i} \]  

(1)  

\[ \text{AGS} = \frac{\text{total_bp}}{\text{NGs}} \]  

(2)  

Where "gene_bp," and "gene_length," are the number of annotated base pairs and length of marker gene "i," and "total_bp" is the total number of base pairs in the target metagenome.

To annotate the single-copy genes, we created a UProC database. We downloaded the eggnog database version 4.5 [33], selected the amino acid sequences (full-alignment files) used to create the Hidden Markov Model profiles of the 35 single-copy genes, applied the SEG low complexity filtering tool of the NCBI Blast+ 2.2 Suite [21] on these sequences, and created the UProC database with the uproc-makedb command.

**16S rRNA gene average copy number computation tool (acn.sh)**

The workflow implemented in the acn.sh tool, consists of annotating the 16S rRNA genes using SortMeRNA 2.0 [34], estimating the 16S rRNA gene coverage as the ratio of the 16S rRNA gene length, and computing the ACN as the ratio of the 16S rRNA gene length to the NGs (see Eq. 3). The 16S rRNA gene length in this tool is set to 1542 bp, which corresponds to the full-length 16S rRNA gene of *Escherichia coli*. To run SortMeRNA, we use its pre-packaged silva-bac-16s-id90 and silva-arc-16s-id95 16S rRNA gene sequence databases.

\[ \text{ACN} = \frac{16S\text{_gene_bp}/16S\text{_gene_length}}{\text{NGs}} \]  

(3)  

Where "16S_gene_bp" and "16S_gene_length" are the number of annotated base pairs and the 16S rRNA gene length, respectively, and "NGs" is the number of genomes in the target metagenome.

**Data acquisition, pre-processing, and analysis**

The 139 prokaryotic metagenomes of the TARA Oceans dataset were downloaded from the European Nucleotide Archive [35] (ENA:PRJEB1787). To pre-process the metagenomic short-read data, we applied the following procedure. We clipped the adapter sequences (obtained from Shinichi Sunagawa personal communication, July 21, 2015) with the BBduk tool of the BBMap 35.00 suite [29]; We then merged the paired-end reads with VSEARCH 2.3.4 [36], quality trimmed all reads at Q20 and filtered out sequences shorter than 45 bp using BBduk; Lastly, we de-replicated the quality-controlled sequences with VSEARCH.

We estimated the Average Genome Size (AGS) and the 16S rRNA gene Average Copy Number (ACN) in the 139 metagenomes with the ags.sh and acn.sh tools, respectively. To run the ags.sh tool we used the minimum length parameter set to 100 bp. To filter significant 16S rRNA gene sequence alignments when running the acn.sh tool, we used an e-value of 1e-5. We selected a matching subset of 63 TARA Oceans metagenomes representing the Surface (SRF), Deep Chlorophyll Maximum (DCM) and Mesopelagic (MES) water layers in 21 sampling stations [28], to analyze the changes of the AGS and ACN between water layers.

To test whether the AGS and ACN values differ between water layers, we applied a paired Wilcoxon rank-sum test between each pair of water layers, using the wilcox.test function of the vegan R package [37].

We used TARA Oceans’ taxonomic abundance profile computed by Sunagawa et al. 2015 based on the annotation of 16S rDNA Operational Taxonomic Units (OTUs), to search for genera that correlated with the AGS and ACN. First, we removed singletons and genera with a total relative abundance lower than 0.001%, and computed the genera relative abundance in each metagenome (total sum scaling standardization). Subsequently, we selected the genera for which their correlation with either of these traits had a p-value lower than 0.001 after applying the Bonferroni correction for multiple comparisons. Additionally, we compared the AGS with the functional richness computed by Sunagawa et al. in the 139 TARA Oceans metagenomes.

**Simulation of metagenomic datasets**

To assess the performance of our tools, we created three simulated metagenomic datasets (i.e., General, Infant Gut, and Marine). Each dataset is composed of 50 metagenomes, and all metagenomes have a size of two million reads. The metagenomes of the General dataset were simulated based on a random selection of prokaryotic species. The Infant Gut and Marine datasets approximate the taxonomic composition of the microbial communities found in the infant’s gut and marine environments, respectively.

To create each of the three simulated metagenomic datasets, we first created a reference dataset of complete genome sequences and the abundance profiles to define the community composition of each metagenome. In the case of the General genome reference dataset, we randomly selected 500 genera from all the prokaryotic genera in the NCBI RefSeq database [38] and downloaded a maximum of three genome sequences per genus with an assembly status of “Complete genome” (on November 8, 2017). The resulting dataset comprised 751 different species. For the Infant Gut genome reference dataset, we used the genus taxonomic annotation of the metagenome-assembled genomes
(MAGs) generated by Sharon et al. (2013) [39] obtained from fecal samples collected from a premature infant, to guide the selection of species. We downloaded for each species one genome with an assembly status of “Complete genome” from the NCBI RefSeq database. If a species did not have a complete representative genome, we randomly selected another species with the same genus affiliation. The Infant Gut reference genome dataset contains 95 different species. Finally, the Marine reference genome dataset was created based on the taxonomic composition of TARA Oceans 16S rDNA Operational Taxonomic Units (OTUs) described by [28]. We selected 172 genera for which there was at least one representative completely sequenced genome and downloaded a maximum of three genomes per genus from RefSeq, irrespective of their species affiliation. This database comprises 308 species.

To define the community profile of each metagenome, we randomly selected between 20 and 80 genomes from a reference dataset and assigned the genome abundances by sampling from a lognormal distribution with mean 1 and standard deviation of 0.5. We used these profiles together with the corresponding reference genome sequence data as an input to run MetaSim v0.9.5 [40], where we set the read length to 300 bp and the substitution rate to $1 \times 10^{-3}$.

Lastly, we simulated a second Marine metagenomic dataset of 50 metagenomes (i.e., Marine dataset-2). This dataset was generated using the marine abundance profiles and reference genome dataset mentioned above; however, in this case, we simulated merged paired-end reads and varied their length according to the following distribution: $p(50 \text{ bp}) = 0.05; p(110 \text{ bp}) = 0.15; p(150 \text{ bp}) = 0.15; p(165 \text{ bp}) = 0.5; p(180 \text{ bp}) = 0.15$. In addition, the substitution rate along each read was set to increase from $1 \times 10^{-4}$ to $9.9 \times 10^{-2}$. The simulated short-read sequences were merged using VSEARCH [36] with default parameters. With this read length distribution and error rates, we aimed to generate a more realistic dataset. It has similar characteristics as the metagenomic data obtained using Illumina HiSeq 2000 sequencing technology (as the prokaryotic metagenomes of the TARA Oceans dataset), and the majority of Illumina sequencing technologies in general [41].

In Additional file 2, we show the taxonomic composition of the reference datasets of complete genome sequences, and in Additional file 3, we show details of the simulated metagenomic datasets.

**Benchmarking and validation**

To benchmark the wall-clock runtime of ags.sh against MicrobeCensus, we randomly selected five (pre-processed) metagenomes of the TARA Oceans dataset and subsampled these to two million paired-end reads with the seqtk v1 tool [42]. Next, we ran the AGS computation on each metagenome three times with both tools, using a different number of threads in each run (i.e., 4, 8, and 16). All the computations were performed in a workstation with Intel(R) Xeon(R) CPU E7–4820 v4 @ 2.00GHz.

To benchmark the accuracy of ags.sh against MicrobeCensus, we computed the AGS in the metagenomes of the General, Infant Gut, and Marine simulated datasets with both tools and compared it with the real AGS. To assess the accuracy of these tools in relation to the read length, we trimmed the 3’ end of the reads to simulate metagenomic datasets of different read lengths. Namely, for each of the General, Marine, and Infant Gut datasets, we trimmed the simulated 300 bp reads to 100 bp, 120 bp, 130 bp, 140 bp, 150 bp, 160 bp, 170 bp, 180 bp, and 200 bp. We processed a total of 500 metagenomes per dataset.

In these analyses, ags.sh was run with default parameters and MicrobeCensus was set to process the total number of reads in a metagenome. To derive the real AGS of a simulated metagenome, we computed the sum of the lengths of its component genomes weighted by their respective abundance, divided by the total abundance of genomes (see Eq. 4). The genome abundances were obtained from the abundance profiles used to simulate the metagenomes. To quantify the accuracy, we computed the Pearson’s correlation and Absolute Percentage Error (APE) (see Eq. 5) of the AGS computed by ags.sh and MicrobeCensus, with respect to the real AGS.

$$\text{AGS}_{\text{real}} = \frac{1}{\text{total_abund}} \sum_{i=1}^{k} \frac{\text{genome_length}_i}{\text{genome_abund}_i}$$

$$\text{APE} = 100 \times \frac{\text{Ref} - \text{Est}}{\text{Ref}}$$

Where “total_abund” is the total abundance in the target metagenome, and “genome_length,” and “genome_abund,” are the length and abundance of genome “i,” respectively.

To compare ags.sh vs. MicrobeCensus on real data, we computed the AGS with these tools on a randomly selected subset of 50 (pre-processed) metagenomes of TARA Oceans. To accelerate the computation of the AGS, we randomly subsampled the metagenomes to two million reads using the seqtk v1 tool [42].

To measure the wall-clock running time of acn.sh, we used the same five TARA Oceans metagenomes subsampled to two million paired-end reads and previously used to benchmark the ags.sh running time. Next, we ran acn.sh using different thread numbers (i.e., 4, 8, and 16).
and measured the wall-clock running time. These computations were performed in a workstation with Intel(R) Xeon(R) CPU E7–4820 v4 @ 2.00 GHz.

To benchmark the accuracy of acn.sh, we compared it with PICRUSt [23], CopyRighter [24], and PAPRICA [25]. We computed the ACN with these four tools in the three simulated metagenomic datasets, where we also trimmed the 3’ end of the metagenomic reads to produce different read lengths (i.e., 100 bp, 150 bp, 200 bp, and 300 bp). We then computed the Pearson’s correlation and APE between the predicted and real ACNs.

To compute the ACN with acn.sh, we ran the tool with default parameters. To compute the ACN with CopyRighter and PICRUSt, we applied the following procedure: 1) Reads containing 16S rRNA genes were identified with SortMeRNA 2.0 [34]; 2) 16S rRNA gene sequences were extracted and clustered at 99% identity with VSEARCH 2.34 [36]; 3) Cluster centroid sequences were blasted against the GreenGenes databases GG2012 (release October 2012) and GG2013 (release May 2013) [43], using BLASTN [21] with an e-value of 0.001 and an identity threshold of 99%; 4) The 16S rRNA gene copy numbers of the best hits were parsed out of the respective lookup tables (ssu_img40_gg201210.txt and 16S_13_5_precalculated.tab for CopyRighter and PICRUSt, downloaded from [44, 45], respectively); 5) The ACN was computed as the average of the predicted 16S rRNA gene copy number (i.e., 16S rRNA gene copy number of the best hits), weighted by the abundance of the cluster represented by the respective query sequence (see Eq. 6).

To compute the ACN with PAPRICA, we used the cluster centroid sequences to run the paprica-run.sh script using the --large parameter for the paprica-place_it.py script (i.e., to increase the memory utilization). Then, we computed the average of the 16S rRNA gene copy numbers predicted for these sequences, weighted by the respective cluster abundances (see Eq. 6).

\[
\text{ACN}_{\text{est}} = \frac{1}{\text{total_abund}} \sum_{i=1}^{k} \text{pred_copy_num}_i \times \text{cluster_abund}_i
\]  

Where “pred_copy_num,” and “cluster_abund,” are the predicted 16S rRNA gene copy number and cluster size of query sequence “i,” and “total_abund,” is the total number of identified 16S rRNA gene sequences in a metagenome.

Similarly, as described above, we computed the real ACN of a metagenome, as the sum of the 16S rRNA gene copy numbers of its component genomes weighted by their respective abundance and divided by the total abundance of genomes (see Eq. 7). The 16S rRNA gene copy numbers were obtained from the NCBI features annotation.

\[
\text{ACN}_{\text{real}} = \frac{1}{\text{total_abund}} \sum_{i=1}^{k} \text{genome_copy_num}_i \times \text{genome_abund}_i
\] 

Where the “total_abund” is the total abundance in the target metagenome, “genome_copy_num,” and “genome_abund,” are the 16S rRNA gene copy number and abundance of genome “i,” respectively.

### Results and Discussion

Our implementation to compute the AGS is based on the annotation of 35 prokaryotic universally distributed single-copy genes identified by Raes et al., (2007) [22]. Most of these genes are part of the translation machinery and essential for cellular life. The main finding that allowed us to develop ags.sh (and in turn acn.sh), is that the annotation of the 35 marker genes in unassembled metagenomes, using new, fast and accurate tools, can be used to rapidly estimate the genes’ coverage, which accurately approximates the total number of genomes (NGs). Thus, we can derive the AGS analytically, as the ratio of NGs and the total number of base pairs in a metagenome. However, to estimate the NGs it is crucial to perform a precise annotation of the single-copy genes. To this end, we include in ags.sh an option to filter and trim sequence reads to obtain the optimal read lengths for the annotation of single-copy genes (see Fig. 1a). The computation of the 16S rRNA gene average copy number follows a similar methodology: we estimate the coverage of the 16S rRNA genes and divided it by the NGs (see Fig. 1b).

### Benchmarking the average genome size computation tool (ags.sh)

Firstly, we benchmarked the wall-clock runtime of ags.sh against MicrobeCensus (see Fig. 2a and Additional file 4). We used a subset of five TARA Oceans metagenomes subsampled to two million paired-end reads (see Materials and Methods), to measure the runtime of these tools with increasing number of threads (i.e., 4, 8 and 16). In this analysis, we also benchmarked the runtime of ags.sh using previously predicted Open Reading Frame (ORF) amino acid sequences. We consider this is a likely scenario when using our tool, given that the prediction of ORFs is a standard procedure in most metagenomic analysis workflows. With and without the ORF prediction step, ags.sh was on average 6.6 and 12.6 times faster than MicrobeCensus, respectively. Ags.sh also showed a greater runtime improvement when we increased the number of threads from 4 to 16. The acceleration achieved by our implementation is the
result of using a fork of FragGeneScan-plus [30, 31] and the UProC program [32] for the ORF prediction and gene annotation, respectively. FragGeneScan-plus is several times faster than FragGeneScan [46], and has the same prediction accuracy, while UProC is up to three orders of magnitude faster and more sensitive than profile-based methods on unassembled short-read sequences. Reducing the runtime also has consequences for the AGS estimation accuracy, given that all the metagenomic data available can be readily used to compute this trait.

Secondly, we compared the accuracy of ags.sh against MicrobeCensus. We used both tools to estimate the AGS in the simulated metagenomes of three different datasets we generated (i.e., General, Infant Gut, and Marine datasets; see Material and Methods). Simulated data allowed us to calculate the real AGS in the metagenomes, which we used as a reference to evaluate the accuracy of these tools. Each simulated dataset is composed of 50 metagenomes of two million reads of 300 bp, which were trimmed to different read lengths, to evaluate the accuracy of the tools as a function of the read length.

The analysis showed that the performance of both tools changed very little between datasets, and ags.sh had comparable or higher accuracy than MicrobeCensus in metagenomes with a read length between 120 and 200 bp (see Additional file 5 and Additional file 6). The Pearson’s correlation values showed marginal differences between both tools and remained practically constant within each dataset when the read length was changed. However, the error values revealed that the accuracy of ags.sh varies with the read length, with a trend of having higher accuracy than MicrobeCensus in metagenomes with a read length between 120 and 200 bp. MicrobeCensus was less affected by the read length and outperformed ags.sh in metagenomes with read lengths of 100 and 300 bp.

The optimal read length range observed for our tool reflects the read length in which UProC has an optimal sensitivity and specificity. In essence, if the reads are too short there is a lower sensitivity, and consequently, the AGS is overestimated (NGs underestimated). Conversely, when the reads are too long, there is a lower specificity and the AGS is underestimated (NGs overestimated).

It is important to point out that ags.sh has an option to remove and trim metagenomic reads to obtain appropriate read lengths. This way, it can always process metagenomic reads where it has its highest accuracy.

In addition, to test these tools in a more realistic scenario, we generated a second simulated Marine
metagenomic dataset (i.e., Marine dataset-2). Marine dataset-2 is composed of 50 metagenomes of two million merged paired-end reads. We also compared the ags.sh runtime using previously predicted Open Reading Frames (ORFs). When the ORF prediction procedure was included, ags.sh was 11 times faster than MicrobeCensus using 16 threads. b Scatter plots comparing the accuracy of the AGS computed by ags.sh (upper panel) and MicrobeCensus (lower panel) with the reference AGS in the metagenomes of the Marine dataset-2. c Scatter plot comparing the AGS computed by ags.sh and MicrobeCensus in 50 TARA Oceans metagenomes, randomly subsampled to two million reads. The black line shown in the scatter plots from b) and c) represents the one-to-one relationship. The absolute percentage error was computed as 100 x |(AGS ref - AGS est)/AGSref|, where AGS est and AGS ref are the estimated and reference AGS, respectively.

Benchmarking the 16S rRNA gene average copy number estimation tool (acn.sh)
Analog to the previous benchmark analysis, we measured the wall-clock runtime of acn.sh using 4, 8, and 16 threads on the five subsampled metagenomes of TARA Oceans (see Fig. 3a and Additional file 7). In this evaluation, we observed that acn.sh scales very well with the number of threads and is able to process approximately one million reads per minute using eight threads. The most computationally intensive task performed by acn.sh, is the annotation of the 16S rRNA genes with SortMeRNA [34], which determines the runtime. SortMeRNA has an optimal accuracy-speed trade-off, making it very convenient for the computation of this trait in metagenomic data. In addition, given that acn.sh depends on the NGs estimated with ags.sh, we benchmarked the running time of
ags.sh plus acn.sh. When both tools are taken into account, we observed that a two million paired-end reads metagenome is processed in less than four minutes using eight threads.

We also compared the accuracy of acn.sh with PICRUSt, CopyRighter, and PAPRICA (see Additional file 8 and Additional file 9). As described above, we used acn.sh to compute the ACN in the simulated metagenomes of the General, Infant Gut, and Marine datasets, trimmed to different read lengths. We then computed the real ACN, which we used as a reference to assess the accuracy of these tools. This analysis showed that acn.sh is considerably more accurate than the other three tools. As observed for ags.sh, acn.sh had comparable performance in the three datasets; while the correlation coefficients were not affected by the read length, the APE values increased in metagenomes of 100 and 300 bp. The under- and overestimation of the NGs computed by ags.sh limited the accuracy in these latter two cases. When considering an appropriate read length (i.e., 150 and 200 bp), the correlation and MdAPE values ranged from 0.9 to 0.94, and from 2.9 to 6.4%, respectively.

Conversely, the performance of PICRUSt, CopyRighter, and PAPRICA varied between datasets, especially between the Infant Gut and Marine datasets, where we observed their highest and lowest performance, respectively. For instance, the correlation and MdAPE values of PAPRICA, which showed the highest accuracy of the three in the Infant Gut dataset, went from 0.78 to 0.48, and from 3.2 to 15.8%, respectively, in the metagenomes of 100 bp. Such finding is likely to reflect the low representation of sequenced environmental taxa in reference phylogenies, which is known to limit the prediction power of these tools [27].

In Fig. 3b and c, we show the scatter plots comparing the ACN computed with acn.sh and CopyRighter vs. the reference ACN in the metagenomes of the Marine dataset-2. The black line shown in the plot represents the one-to-one relationship. Similarly as above, we applied the following formula to compute the absolute percentage error: $100 \times |(\text{ACN}_{\text{ref}} - \text{ACN}_{\text{est}})/\text{ACN}_{\text{ref}}|$, where ACN$_{\text{est}}$ and ACN$_{\text{ref}}$ are the estimated and reference ACN, respectively.
more complex approach based on the phylogenetic annotation of the 16S rRNA genes.

**Analysis of the average genome size and 16S rRNA gene copy number in TARA oceans metagenomes**

We computed the AGS and ACN in the 139 prokaryotic metagenomes of TARA Oceans (using complete metagenomic samples) to analyze microorganisms’ ecological strategies associated with different marine environmental conditions. Firstly, we conducted pairwise comparisons of the AGS and ACN between water layers. For this task, we used a matching subset of 63 TARA Oceans metagenomes that represent the surface (SRF), deep chlorophyll maximum (DCM) and mesopelagic (MES) water layers in 21 sampling stations across the globe [28]. The results showed that apart from the ACN of the SRF and DCM water layers, all other water layers have significant differences in their trait values (paired Wilcoxon rank-sum test; all \( p \)-values < 0.05; see Additional file 10). As observed previously [16], we found that the two traits were significantly correlated among themselves (Pearson’s \( r = 0.38; p\text{-value} = 0.0023 \) (see Fig. 4a; Additional file 11).

The distributions of the ACN values resemble the differences in minimum generation times between water layers, computed by Sunagawa et al. 2015 for the same metagenomes. As shown by Vieira-Silva & Rocha (2010) [17], the 16S rRNA gene copy number has a strong correlation with the growth rate of microorganisms.

Additionally, based on the analysis of the 139 TARA Oceans metagenomes, we observed significant correlations between both traits and the sampling depth in the water column (AGS vs. water depth: Pearson’s \( r = 0.46 \); ACN vs. water depth: Pearson’s \( r = 0.28 \); all \( p \)-values < 0.001; Additional file 11). Similar findings relating the genome size and 16S rRNA copy number with water depth have been previously described [47, 48]. We also obtained a significant correlation between the ACN and water temperature after controlling for water depth (Pearson’s \( r = -0.34 \) and \( p\text{-value} < 0.001 \); Additional file 11), which was not observed for the AGS. This result could be

---

**Fig. 4** Exploratory analyses performed on TARA Oceans metagenomes. **a** Scatter plot comparing the AGS and ACN in the matching subset of 63 TARA Oceans metagenomes representing the surface, deep chlorophyll maximum and mesopelagic water layers (SRF, DCM, and MES, respectively) in 21 sampling sites. The box plots in the lower and left-hand side panels show the distributions of the Average Genome Size (AGS) and 16S rRNA gene Average Copy Number (ACN) in the SRF, DCM, and MES water layers. For the sake of clarity, two metagenomes with relatively large AGS or ACN values were not included in the plot. These are the TARA_076_DCM_0.22–3 with an AGS = 5,036,010 bp and TARA_064_DCM_0.22–3 with an ACN = 2.4. **b** Scatter plots comparing the AGS with the log relative abundance of the *Herbiconiux* and *Candidatus Pelagibacter* genera (upper and lower panel, respectively) in TARA Oceans metagenomes. *Herbiconiux* and *Candidatus Pelagibacter* genera had the strongest positive and negative Pearson’s correlations with the AGS, respectively. **c** Scatter plot comparing the ACN with the log relative abundance of the *Glaciecola* genus in TARA Oceans metagenomes. This genus showed the strongest positive Pearson’s correlation with the ACN. The abundance of these genera was computed by Sunagawa et al. based on the annotation of 16S rDNA Operational Taxonomic Units (OTUs). **d** Scatter plot comparing the AGS with the functional richness of TARA Oceans metagenomes. The functional richness was computed by Sunagawa et al. based on the abundance estimation of eggNOG orthologous groups.
explained by the fact that psychrophiles are slower growers than expected, given their growth-associated genomic traits [17]. That is, for the same minimal generation times, organisms inhabiting colder marine environments tend to have a greater 16S rRNA gene copy number to compensate for slower enzymatic activity.

Based on the taxonomic annotation of the rDNA Operational Taxonomic Units (OTUs) obtained by Sunagawa et al. 2015 for TARA Oceans metagenomes, we looked for genera associated with the variability of the AGS and ACN traits. This analysis revealed that the genera *Herbiconius* and *Candidatus Pelagibacter* had the strongest positive and negative correlations with the AGS, respectively, and the genus *Glaciecola* had the strongest positive correlation with the ACN (see Fig. 4b and c; Additional file 11). *Herbiconius* species tend to have a relatively large genome size, often above 6 Mbp, while *Candidatus Pelagibacter* has a streamlined genome, which is around 1.3 Mbp. On the other hand, sequenced species of *Glaciecola* contain from four to five rRNA gene operons [49]. These are marine microorganisms with extreme values in their genomic traits, and represent different ecological strategies associated with different environments (see below). We note that we did not find any genus with a significant negative correlation with the ACN. That is, the ACN, which tends to be low, appears to vary between metagenomes, mainly when these have a higher abundance of organisms with a high 16S rRNA gene copy number.

Lastly, we compared the AGS of TARA Oceans metagenomes, with the functional richness estimated by Sunagawa et al. 2015 based on the annotation of eggNOG orthologous groups [33]. We observed a highly significant correlation between these two metagenomic traits (Pearson’s $r = 0.52$ and $p$-value $< 0.0001$; see Fig. 4d and Additional file 11). This result was expected, given that the number of genes in prokaryotes is known to be linearly related to the genome size [50]. However, the nature of this relationship using the community AGS and functional richness offers new insights into the ecology of marine microbial communities. The AGS explained a moderate amount of the functional richness variation ($R^2 = 0.27$). Several other factors can influence the community functional richness and AGS. For example, the functional richness is also highly correlated with the taxonomic richness and water depth (Pearson’s $r = 0.85$ and 0.65, respectively; all $p$-values $< 0.0001$; Additional file 11). A plausible explanation, in line with previous work characterizing prokaryotic ecological strategies [9, 51], is that more heterogeneous environmental conditions, which appear to be found in deeper water layers [28, 52], result in a higher functional and taxonomic richness and more complex ecological interactions. In turn, a higher complexity of the ecological interactions and a greater environmental heterogeneity, increase the demand for larger gene repertoires, and consequently, larger genomes.

Taken together, these exploratory analyses indicate that surface marine microbial communities are characterized by a small AGS and low ACN. Such trait values denote the presence of efficient, slow growers and specialist organisms (i.e., k-strategist), in agreement with the oligotrophic environmental conditions commonly found in marine surface waters. On the contrary, microbial communities from the DCM and MES water layers exhibited a larger AGS and greater ACN, which indicate that organisms inhabiting deeper layers tend to have a more diverse metabolism and grow faster (i.e., r-strategist). As such, these organisms respond better to environmental changes and can exploit intensively nutrient rich micro-niches [10, 53, 54].

**Conclusions**

In this work, we developed the ags.sh and acn.sh tools that accurately and rapidly compute the average genome size and 16S rRNA gene average copy number in unassembled prokaryotic metagenomes. The quantification of these traits provides a powerful approach to characterize microbial ecological strategies. We benchmarked and evaluated the performance of these tools using simulated metagenomic datasets composed of contrasting microbial communities. In these analyses, we showed that the ags.sh tool is up to 11 times faster with comparable or higher accuracy than MicrobeCensus. Reducing the computation time is a valuable improvement given the large data volumes generated by current sequencing technologies. Ags.sh can be readily used to process a comprehensive metagenomic sample for the estimation of the AGS, as exemplarily applied here on TARA Oceans metagenomes. Given that MicrobeCensus is already a highly accurate tool, there was little room for improvement in this sense, and ags.sh only showed a moderate improvement in accuracy regarding the absolute error rates. However, the fact that ags.sh derives the AGS analytically makes it more reliable in comparison to MicrobeCensus. Lastly, our benchmarking analysis of the acn.sh tool showed that it has remarkable accuracy and outperforms the ACN computation approaches based on the copy number predictions of PICRUSt, CopyRighter, and PAPRICA. The fact that acn.sh is exclusively dedicated to the computation of the ACN allows to considerably simplify the analysis workflow.

The exploratory analyses performed on TARA Oceans metagenomes demonstrate the applicability of our tools to compute the AGS and ACN traits on unassembled metagenomic data, and predict the dominant ecological strategies taking place within microbial communities.

We note that the results presented here, show that the AGS and ACN can be derived analytically based on the annotation of single-copy and 16S rRNA genes. Accordingly,
future implementations of the ags.sh and acn.sh tools have the potential to improve in speed and accuracy, as gene notation tools continue to advance. Additionally, in future implementations, it will be of particular interest to include the computation of the AGS and ACN variances [55].

Availability and requirements
Project name: AGS-and-ACN tools
Project home page: https://github.com/pereiramemo/AGS-and-ACN-tools
Operating system(s): Platform independent.
Programming language: AWK, Bash, and R.
Other requirements: Docker.
License: GNU General Public License v3.0.
Any restrictions to use by non-academics: none.

Additional files

Additional file 1: Single-copy genes universally present in prokaryotes. Table showing the COG number of the 35 single-copy genes identified by Raes et al. (2007), and the number of sequences included in each COG. (XLSX 42 kb)

Additional file 2: Reference genome datasets. Excel sheets with the accession, taxid, organism name, and FTP URL of the genomes used to generate the simulated metagenomic datasets. (XLSX 105 kb)

Additional file 3: Simulated metagenomic datasets. Table showing main characteristics describing the simulated metagenomic datasets. (XLSX 42 kb)

Additional file 4: Benchmarking the running time of ags.sh against MicrobeCensus. Table showing the mean running time of ags.sh and MicrobeCensus using 4, 8, and 16 threads, for the estimation of the AGS in five TARA Oceans metagenomes subsampled to two million paired-end reads. (XLSX 35 kb)

Additional file 5: Benchmarking the accuracy of ags.sh against MicrobeCensus: figure illustration. Plots of the Pearson’s correlation coefficients (upper panel) and absolute percentage error (APE) value distributions (lower panel) of the AGS computed by ags.sh and MicrobeCensus, with respect to the reference AGS. The comparisons were performed using the simulated metagenomes of different read length of the General, Infant Gut, and Marine datasets. For the sake of clarity, 70 outlier APE values (2.3% of the total data) were not included in the plot. (PDF 60 kb)

Additional file 6: Benchmarking the accuracy of ags.sh against MicrobeCensus: summary statistics. Table showing the Medians, Means and Standard Deviations of the Average Percentage Error (i.e., MdAPE, MAPE, and SDAPE, respectively), and Pearson’s r correlation coefficients obtained in the analysis benchmarking the accuracy of ags.sh against MicrobeCensus. (XLSX 41 kb)

Additional file 7: Evaluating the running time of acn.sh. Table showing the mean running time of acn.sh and acn.sh plus acn.sh, using 4, 8, and 16 threads, for the estimation of the ACN in five TARA Oceans metagenomes subsampled to two million paired-end reads. (XLSX 35 kb)

Additional file 8: Benchmarking the accuracy of acn.sh against PICRUSt, CopyRighter, and PAPRICA: figure illustration. Plots of the Pearson’s correlation coefficient (upper panel) and the absolute percentage error (APE) value distributions (lower panel) of the ACN computed by acn.sh, PICRUSt, CopyRighter, and PAPRICA, with respect to the reference ACN. As mentioned above, we compared these tools using simulated metagenomes of different read length of the General, Infant Gut, and Marine datasets. For the sake of clarity, 100 outlier APE values (4.2% of the total data) were not included in the plot. (PDF 62 kb)

Additional file 9: Benchmarking the accuracy of acn.sh against PICRUSt, CopyRighter, and PAPRICA: summary statistics. Table showing the Medians, Means and Standard Deviations of the Average Percentage Error (i.e., MdAPE, MAPE, and SDAPE, respectively), and Pearson’s r correlation coefficients obtained in the analysis benchmarking the accuracy of acn.sh against PICRUSt, CopyRighter, and PAPRICA. (XLSX 38 kb)

Additional file 10: Comparing the AGS and ACN of TARA Oceans metagenomes between water layers. Table showing details of the Wilcoxon rank-sum tests, performed to evaluate the differentiation of TARA Oceans metagenomes from different water layers according to their AGS, and ACN. The comparisons were made on a matching subset of 63 metagenomes, obtained from the surface, deep chlorophyll maximum, and mesopelagic water layers (SRF, DCM, and MES, respectively), at 21 sampling sites. (XLSX 44 kb)

Additional file 11: Exploring the variability of the AGS and ACN of TARA Oceans metagenomes. Table with details of the Pearson correlation tests between and among the AGS and ACN, and the functional and taxonomic richness of TARA Ocean metagenomes, and the water depth and temperature of the sampling sites. (XLSX 34 kb)

Abbreviations
AGS: Average Gene Copy Number; ACN: Average Genome Size; APE: Absolute Percentage Error; DCM: Deep Chlorophyll Maximum; MAGs: Metagenome Assembled Genomes; MdAPE: Median of the Absolute Percentage Errors; MES: Mesopelagic; NGs: Number of Genomes; ORFs: Open Reading Frames; OTUs: Operational Taxonomic Units; SRF: Surface

Acknowledgements
This work was accomplished using computational facilities provided by the Max Planck Society.

Authors’ contributions
EP-F wrote the software, performed the analyses and drafted the manuscript. EP-F and AF-G conceived and designed the experiments. FOG, AF-G, and EP-F discussed the results. All authors contributed to writing the manuscript and approved the final version.

Funding
EP-F was supported by a Ph.D. fellowship of the German Academic Exchange Service (DAAD) and the Uruguayan National Research and Innovation Agency (ANII). AF-G received funding from the European Union’s Horizon 2020 research and innovation program [Blue Growth: Unlocking the potential of Seas and Oceans] under grant agreement no. 634486 436 (project acronym INMARE). None of the funding bodies played a role in the design of the study, collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
ags.sh and acn.sh are free software distributed under the GNU General Public License v3.0. These tools are available at https://github.com/pereiramemo/AGS-and-ACN-tools, where we also provide their source code. In the Additional file 2, we provide the RefSeq (https://www.ncbi.nlm.nih.gov/refseq) assembly accession number of the genome sequences used to simulate the metagenomic datasets. The metagenomic data analyzed here can be downloaded from https://www.ebi.ac.uk/ena (study accession PRJEB1787).

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
1Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359 Bremen, Germany.
References


Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
2.4 Mg-Traits pipeline and database: advancing functional trait-based approaches in metagenomics
Mg-Traits pipeline and database: advancing functional trait-based approaches in metagenomics

Emiliano Pereira-Flores\textsuperscript{1,2}, Renzo Kottmann\textsuperscript{1}, Albert Barberan\textsuperscript{3}, Frank Oliver Glöckner\textsuperscript{1,2} and Antonio Fernández-Guerra\textsuperscript{1,2,4}

1. Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, Germany
2. Department of Life Sciences & Chemistry, Jacobs University Bremen gGmbH, Bremen, Germany
3. Department of Soil, Water and Environmental Science, The University of Arizona, Arizona, USA
4. Oxford e-Research Centre (OeRC), University of Oxford, Oxford, UK

Contact: epereira@mpi-bremen.de
Abstract

Functional traits (FTs) are defined as measurable properties of an organism that influence its fitness. As such, FTs provide useful information to address ecological research questions. In particular, FT-based approaches in metagenomics are of great value to study microorganisms’ ecology. Metagenomic data can be used to compute FTs at the genomic level on a random sample of individuals from a community. These, as community-aggregated FTs, can be used to predict community ecological processes and identify environmental drivers.

In this work, we present the Metagenomic Traits PipeLine and DataBase: Mg-Traits-PL and Mg-Traits-DB, respectively. Mg-Traits-PL is dedicated to the computation of 15 metagenomic traits on short-read metagenomic data, ranging from GC content to functional diversity. Mg-Traits-DB is a PostgreSQL database, which contains characterized metagenomic traits in 356 marine metagenomes of three major datasets: Global Ocean Sampling, TARA Oceans, and Ocean Sampling Day 2104. We used the 139 prokaryotic metagenomes of TARA Oceans to investigate the metagenomic traits variability, and found significant differences between microbial communities inhabiting different water layers, which denote microorganisms’ environmental adaptations. Taken together, in these analyses we demonstrate that Mg-Traits-PL and Mg-Traits-DB contribute to improve the exploitation of metagenomic data and facilitate comparative and quantitative studies.
Introduction

In the last years, there has been increased recognition of the value of functional trait (FT) based approaches in microbial ecology (McGill et al. 2006; Green, Bohannan, and Whitaker 2008; Krause et al. 2014). FTs are measurable properties of an organism that influence its fitness, and as such, these provide complementary information to the ecological studies traditionally based on the taxonomic composition (Violle et al. 2007). Given the overwhelming diversity of microbes and the fact that these are subjected to high rates of horizontal gene transfer processes (Cordero and Polz 2014), the knowledge gained from the FT data becomes particularly useful to study many aspects of microbes’ ecology. For example, questions related to central microbial ecology research areas like biogeography, and diversity and ecosystem functioning can benefit when incorporating FT-based approaches (Green, Bohannan, and Whitaker 2008; Krause et al. 2014).

The application of FT-based approaches in microbial ecology can be particularly enhanced when coupled with metagenomics which, as a culture-independent method, allows us to obtain the genetic material of microorganisms inhabiting a wide range of environments (Kumar et al. 2015). Although the computation of FTs of specific species or individuals using metagenomic data can be highly challenging or impossible (it requires the metagenome binning and assembly), the metagenomic data is very well suited for the computation of FTs at the community level. In other words, unassembled metagenomic data can be used to readily compute community-aggregated FTs at the genome level from a random sample of individuals, independently of their taxonomic affiliation (hereafter referred to as mg-traits). Mg-traits (as community-aggregated traits) represent a conceptual strategy initially developed in plant ecology that can help us to predict ecological processes at the community level and identify environmental drivers (Fierer, Barberán, and Laughlin 2014).

Based on the work of Barberan and coworkers (Barberán et al. 2012), who analyzed several mg-traits and showed their value as ecological markers, we developed the Metagenomic Traits PipeLine and DataBase: Mg-Traits-PL and Mg-Traits-DB. Mg-
Traits-PL is dedicated to the computation of 15 mg-traits based on short-read metagenomic data, ranging from GC content to functional diversity (see Fig. 2.4-1 and Table 2.4-1). Mg-Traits-DB is a PostgreSQL database containing mg-trait-characterized publicly available metagenomes. Currently, it comprises 356 metagenomes from three major metagenomic projects: Global Ocean Sampling (GOS) (Rusch et al. 2007), Ocean Sampling Day 2014 (Kopf et al. 2015), and TARA Oceans (Karsenti et al. 2011).

To exemplify the applicability of Mg-Traits-PL, we performed an exploratory analysis on the 139 TARA Oceans prokaryotic metagenomes. First, we tested the spatial autocorrelation of each mg-trait and their capacity to capture the functional and taxonomic compositional differences between metagenomes. Secondly, we analyzed the mg-trait interrelationships, and the environmental factors influencing their variability. Lastly, we assessed the discrimination power of the taxonomic and functional composition to differentiate types of environments and geographic locations.

The mg-traits included in Mg-Traits-PL deliver a valuable set of ecological markers that can be used to characterize habitats types, geographic locations, or temporal samples. Furthermore, inter-trait relationships can be used as indicators of artifacts during sample processing. The combination of these metagenomic traits in a single pipeline facilitates a consistent and comprehensive interpretation of metagenomic data. In addition, the previously mg-trait-characterized metagenomes can contribute to metagenomic comparative analyses.

The information provided by each of the 15 mg-traits is described below. For the ease of communication, the mg-traits are numbered and grouped into four different categories.

1. **Direct mg-traits.** This category includes mg-traits computed directly on the metagenomic data, namely the GC content (1) and variance (2), and the dinucleotide frequency (Dinuc freq) (3). These provide nucleotide signatures known to be useful for the taxonomic assignments (Mrázek 2009) and environmental differentiation (Foerstner et al. 2005; Willner, Thurber, and Rohwer 2009).
2. **Open reading frame-derived mg-traits.** This second group includes the codon (4) and amino acid frequency (AA freq) (5) distributions, and the acidic to basic amino acid ratio (AB ratio) (6); all computed based on the predicted Open Reading Frame (ORF) sequences. Codon usage preferences are known to be shared among microorganisms inhabiting similar ecological conditions, reflecting the translational optimization of highly expressed and lifestyle-specific genes important for the adaptation to the environment (Roller et al. 2013). Likewise, the amino acid frequency of metagenomes has been shown to reflect adaptations of the encoded proteomes to different environmental conditions (Moura, Savageau, and Alves 2013). In particular, the salinity concentration appears as a significant factor impairing a selective pressure on the acidic to basic amino acid ratio (Rhodes et al. 2010).

3. **Functional annotation-derived mg-traits.** This category comprises six mg-traits, which are the functional composition (Fun comp) (7) and diversity (Fun div) (8), the percentage of functionally annotated metagenomic reads (Perc fun annot) (9) and transcription factors (Perc TF) (10), the Biosynthetic Gene Cluster (BGC) domain composition (BGC comp) (11), and the Average Genome Size (AGS) (12). The functional composition is a strong predictor of the environmental conditions and microbial ecological processes (e.g., Dinsdale et al. 2008; Raes et al. 2011; Sunagawa et al. 2015). The functional diversity and percentage of transcription factors are related to the heterogeneity and variability of the environment (Parter, Kashtan, and Alon 2007; Sunagawa et al. 2015), while the percentage of functionally annotated reads is inversely related to the functional novelty. On the other hand, the BGC domain composition (computed using BiG-MEx (Pereira-Flores et al. in prep.) (see section 2.2) covers 150 domains, which are part of BGC sequences responsible for the production of 44 different chemical classes. Many of the compounds produced by the BGC encoded proteins participate in microbes’ ecological processes, including competition, nutrient acquisition, and quorum sensing (Fischbach and Voigt 2010).
and class profiles have been shown to discriminate between types of environments. In addition, this mg-trait can be used as an input for BiG-MEx to explore the BGC domain diversity and class composition. Lastly, the AGS (computed using the ags.sh tool (Pereira-Flores, Glöckner, and Fernández-Guerra 2019) (see section 2.3)) serves as a proxy to assess microorganisms’ ecological strategies. Larger genomes tend to contain more diverse metabolic repertories and are associated with organisms inhabiting heterogeneous environments (Guieysse and Wuertz 2012).

4. **16S rRNA gene-derived mg-traits.** This last category includes traits computed based on the identification of metagenomic 16S rRNA gene sequences. These are the taxonomic composition (Taxa comp) (13) and diversity (Taxa div) (14), and the 16S rRNA gene Average Copy Number (ACN) (15) (the latter computed using the acn.sh tool (Pereira-Flores, Glöckner, and Fernández-Guerra 2019) (see section 2.3)). Assessing the taxonomic composition (i.e., community identities) and diversity is one of the main goals in metagenomic analyses, and standardly used to describe microbial communities. The 16S rRNA gene copy number is highly correlated with the maximum growth rate (Vieira-Silva and Rocha 2010) and characterizes microorganisms’ responses to nutrient availability (Roller, Stoddard, and Schmidt 2016).

In order to create Mg-Traits-DB, we characterized 356 marine metagenomic samples obtained from different geographic locations worldwide. This dataset includes 68, 149 and 139 metagenomes from the Global Ocean Sampling expedition (GOS) (Rusch et al. 2007), Ocean Sampling Day 2014 (OSD) (Kopf et al. 2015) and TARA Oceans expedition (Karsenti et al. 2011), respectively (see Table 2.4-2). These sequencing initiatives are among the most prominent global surveys of marine environments. Although each project was designed differently, all share the common goal of developing a better understanding of marine microbial communities. Together, these three datasets cover a wide range of sampled ecosystems across the oceans.
including all main oceanic regions, tropical, temperate and polar climates, and coastal, estuarine and open ocean environments. GOS was the first marine metagenomic survey on a global scale, and it was designed to address questions related to the genetic and biochemical microbial diversity in marine environments. It contains surface-water samples taken along a transect from the northwest Atlantic Ocean to the eastern tropical Pacific Ocean. OSD is a sequencing campaign that aims to generate the largest standardized marine microbial dataset in a single day. It consists of an orchestrated sampling on the summer solstice in the northern hemisphere and includes participants from across the globe. The majority of the samples were taken from surface waters in coastal areas within exclusive economic zones. Lastly, the TARA Oceans expedition, considerably larger than the other two projects, includes metagenomic samples taken at multiple water layers from all main oceanic regions.

To improve the consistency within and between the datasets, we selected a subset of metagenomes from each project to be included in Mg-Traits-DB. From OSD and TARA, we analyzed the samples representing the prokaryote-enriched fractions. In the case of the GOS expedition, we included the metagenomes sequenced with Sanger sequencing technology.

The biological interpretation of the mg-traits is strongly dependent on the quality and completeness of the contextual data (metadata). Accordingly, we supplemented Mg-Traits-DB with the corresponding metadata of the metagenomic samples, including the geographic coordinates, water depth, date, salinity, temperature, and the biome, material and feature classes of the Environmental Ontology (Buttigieg et al. 2016).

**Results and discussion**

We followed a similar exploratory analysis as in Barberan et al. (2012) to investigate the variability of the 15 mg-traits in the 139 prokaryotic TARA Oceans metagenomes and assess their utility as ecological descriptors. Firstly, we performed a Principal Component Analysis (PCA) on each of the multi-valued mg-traits (i.e., dinucleotide, codon and amino acid frequencies, and taxonomic, BGC domain and functional com-
position), and selected the first Principal Component (PC1) as a summary variable, to be used in following analyses. The PC1s of the dinucleotide, codon, and amino acid frequencies captured 87% or more of the total variability, while the PC1s of the functional, BGC domain and taxonomic composition captured 45%, 64%, and 49% of the total variability, respectively (see Table 2.4-3).

We applied the Moran’s I test to assess the spatial autocorrelation of each of the 15 mg-traits. For this task, we separately tested the autocorrelation on the metagenomes from the surface, deep chlorophyll maximum and mesopelagic water layers (SRF, DCM, and MES, respectively). In agreement with the work of Barberan et al. (2012), the majority of the mg-traits quantified in the SRF water layer metagenomes, were significantly and positively spatially autocorrelated (p-value < 0.05). On the contrary, communities from the DCM and MES water layers had lower autocorrelation values, and only four and three mg-traits, respectively, were significantly autocorrelated (p-value < 0.05) (see Table 2.4-4). A possible explanation for this observation is that, while the SRF metagenomes were all obtained at 5 meters depth, the DCM and MES metagenomes were sampled at different depths, covering a broader range of environmental conditions. In other words, geographically close metagenomes sampled at different depths, and consequently from highly dissimilar environments, are likely to obscure the mg-traits spatial autocorrelations. The TARA_037_MES_0.22-1.6 and TARA_038_MES_0.22 metagenomes are an example of such a scenario. These metagenomes are separated by a relatively small distance (i.e., 225 kilometers), while their sampling depth differs by 260 meters.

We performed a series of partial Mantel tests to analyze the association between the mg-traits and the functional and taxonomic compositions. This analysis consisted of testing whether the mg-trait and functional dissimilarities between samples are correlated when the taxonomic dissimilarities are considered and vice versa (i.e., mg-trait vs. taxonomic dissimilarities, given the functional dissimilarities). We observed that all mg-trait dissimilarities were significantly correlated with the functional dissimilarities, and non-significantly correlated with the taxonomic dissimilarities (p-value < 0.05; see
Table 2.4-5) (functional and taxonomic PC1s were not included in the former and latter comparisons, respectively). In correspondence with Barberan et al. (2012), these results show that the mg-traits included in Mg-Traits-PL have functional implications, and are independent of the taxonomic composition.

To get an overview of TARA Oceans metagenomes’ mg-trait profiles, we generated a two-way clustering heat map (see Fig. 2.4-2). This analysis showed that the metagenomes from the MES water layer had remarked differences with respect to the other two water layers. The sample clustering, performed on the Euclidean distances of the scaled mg-trait profiles using Ward’s method, revealed a subcluster formed almost exclusively of metagenomes from the MES water layer. Furthermore, only two MES metagenomes were placed outside this subcluster. The clustering of the variables was performed using Ward’s method on a correlation-based dissimilarity measure. As expected, the GC content and PC1s of the dinucleotide, codon, and amino acid frequencies formed a tight cluster, which in turn, was grouped with the Average Genome Size (AGS). The GC content is known to be the main factor influencing the dinucleotide, codon, and amino acid frequencies (Barberán et al. 2012; Zhou et al. 2014) and correlates with the genome size (Almpanis et al. 2018). The PC1s of the BGC domain, functional and taxonomic composition, and the taxonomic diversity formed a separate cluster. The latter three mg-traits describe the main characteristics of microbial communities and, as such, are subject to environmental changes that occur along the water column, as described by Sunagawa et al. (2015) for the same dataset. The PC1s of the BGC domain and functional composition had a particularly strong correlation (Pearson’s $r = -0.98$), and consequently were grouped within this cluster. The functional diversity and GC variance clustered together. This is likely to be explained by the fact that a higher functional diversity implies a higher GC content variability within and between the constituent genomes of a metagenome. In addition, the 16S rRNA gene Average Copy Number (ACN) was clustered together with the percentage of transcription factors (Perc TF). This finding agrees with the ecological strategies associated with these two mg-traits. Both the 16S rRNA gene average copy
number and percentage of transcription factors are associated with the capacity of microorganisms to respond to fluctuating environmental conditions. The Acidic to Basic amino acid ratio (AB ratio) and percentage of functionally annotated reads (Perc fun annot), were placed outside the cluster formed by the 16S rRNA gene average copy number and percentage of transcription factors. The acidic to basic amino acid ratio and percentage of functionally annotated reads showed a somewhat distinct variability pattern in relation to the other mg-traits (see Supplementary Table 6.2-1).

The environmental variability that occurs along the water column appears as the most important factor influencing the mg-trait profiles in this dataset. To further investigate the mg-traits variability, we compared the mg-trait value distributions between the SRF, DCM and MES water layers. For this comparison, we used a matching subset of 63 samples, representing the three water layers in 21 sampling stations (hereafter Layers-Subset). The results showed that nine of the 15 mg-traits differed significantly between the three water layers (paired Wilcoxon rank sum tests; all p-values < 0.05) (see Fig. 2.4-3 and Supplementary Tables 6.2-2). The PC1s of the BGC domain and functional composition were the two variables that showed the most significant differentiation. This result further confirms the value of the BGC domain composition as an mg-trait to characterize the environmental drivers of microbial communities’ composition.

The dramatic changes in the environmental conditions along the water column (e.g., light, temperature, pressure, oxygen concentration) are expected to impact considerably the microbial communities that inhabit these environments. The ecological implications of the mg-traits analyzed here, allow us to gain insight into the ecology of these microbial communities. For example, in agreement with Pereira-Flores et al. (2019) (see section 2.3), an increase in the average genome size, 16S rRNA gene average copy number, percentage of transcription factors, and taxonomic diversity in the MES water layer denote ecological strategies associated with more complex and heterogeneous environments. That is, organisms that form the MES water layer communities tend to participate in more complex ecological interactions, and have the capacity to
grow faster and exploit a more diverse range of nutrients.

We then assessed the discriminative power of the taxonomic and functional composition to differentiate types of environments and geographic locations. We compared the metagenome taxonomic and functional compositions between the SRF, DCM and MES water layers on the one hand, and the Indian Ocean, North Pacific Ocean, South Pacific Ocean, North Atlantic Ocean, South Atlantic Ocean, and the Mediterranean Sea on the other. To assess the discrimination power between water layers and ocean regions, we used the metagenomes of the Layers-Subset, and 57 surface water metagenomes representing the six ocean regions mentioned above (hereafter Regions-Subset), respectively. We applied a Permutational Multivariate Analysis of Variance (PERMANOVA) on the functional and taxonomic composition matrices (see Supplementary Tables 6.2-3, 6.2-4 and 6.2-5) of each dataset (i.e., four PERMANOVA in total: 1) taxonomic composition vs. water layers; 2) functional composition vs. water layers; 3) taxonomic composition vs. ocean regions; 4) functional composition vs. ocean regions). In line with the work of Barberan et al. (2012), the PERMANOVA tests performed on the Layers-Subset showed that the functional profile differentiates types of environments better than the taxonomic composition (R\textsuperscript{2} of 51 and 41 for the functional and taxonomic PERMANOVA, respectively). However, in contrast with the results obtained by these authors, both the taxonomic and functional composition revealed a weaker separation between ocean regions (PERMANOVA R\textsuperscript{2} of 31). This latter finding is in agreement with Sunagawa et al. (2015), who also observed a low discrimination power at the taxonomic level.

To visualize the separation (or lack of) between water layers and ocean regions, we performed a Principal Coordinate Analysis (PCoA) on the taxonomic and functional Bray-Curtis dissimilarity matrices (after applying the Hellinger transformation) (see Fig. 2.4-4). In agreement with the PERMANOVA tests, the taxonomic and functional PCoAs performed on the Layers-Subset showed a clear separation of the MES water layer from the other two water layers, while the PCoAs performed on the Regions-Subset, exhibited a high overlapping between many of the ocean regions.
Conclusions

Based on the work of Barberan et al. (2012) we developed the Mg-Traits-PL, dedicated to the computation of 15 mg-traits. This pipeline allows a comprehensive analysis of metagenomic data and contributes to the advancement of quantitative and predictive ecological analyses. In addition, we created the Mg-Traits-DB, currently composed of pre-computed mg-traits for 356 marine metagenomes, to help the contextualization and ecological interpretation of metagenomic data.

We used Mg-Traits-PL to explore the mg-trait variability in TARA Oceans dataset, where we performed similar analyses as in Barberan et al. (2012). In these analyses, we observed that the mg-traits track community and environmental changes that occur along the water column. Such findings, which are based on a considerable larger dataset compared to the work of Barberan and colleagues, further demonstrate the value of the 15 mg-traits for the characterization of microbial communities and as indicators of environmental adaptation.

The modular design of the Mg-Traits-PL will allow us to easily add new mg-traits in the future. For example, specific metabolic capacities and phylogenetic diversity indices. Additionally, by applying metagenomic binning methods (e.g., Strous et al. 2012; Alneberg et al. 2014) we will be able to compute traits for different taxonomic groups. Lastly, the inclusion of more metagenomic datasets (e.g., Malaspina (Pernice et al. 2014), and OSD 2015 and 2016) will contribute to expand Mg-Traits-DB and cover more environmental conditions.
Chapter 2. Results and Discussion

Materials and Methods

1. Data acquisition and data processing

Ocean Sampling Day (OSD) 2014 and TARA Oceans prokaryotic metagenomes were downloaded from the European Nucleotide Archive (ENA) (Harrison et al. 2019) (accessions PRJEB1787 and PRJEB8682, respectively). The metagenomes of the Global Ocean Sampling (GOS) expedition were downloaded from the iMicrobe database (https://www.imicrobe.us) (accession CAM_PROJ_GOS).

To pre-process the Illumina short-read metagenomic data (i.e., TARA Oceans and OSD metagenomes), we applied a methodology consisting of the following tasks: 1) Clipping the adapter sequences with BBduk tool from the BBMap 35.00 suite (Bushnell 2014); 2) Merging the paired-end reads with VSEARCH 2.3.4 (Rognes et al. 2016); 3) Quality trimming all reads at Q20 and filtering out sequences shorter than 45bp using BBduk; 4) De-replicating the quality-controlled sequences with VSEARCH.

2. Analysis of mg-trait profiles in TARA Oceans metagenomes

We applied Mg-Traits-PL on the previously pre-processed 139 prokaryotic metagenomes of TARA Oceans to compute the 15 mg-traits. The mg-trait values were stored in Mg-Traits-DB. We applied a Principal Component Analysis (PCA) on each of the multi-valued mg-traits (i.e., dinucleotide, codon, amino acid relative frequencies, and functional, BGC domain and taxonomic composition), and used the first Principal Component (PC1) as a summary variable in posterior analyses. The functional, BGC domain and taxonomic composition matrices were transformed with the Hellinger function, before applying the PCAs. To assess the spatial autocorrelation of each of the 15 mg-traits, we applied the Moran’s I test on the metagenomes from the surface, deep chlorophyll maximum and mesopelagic water layers (SRF, DCM, and MES respectively) independently. We computed the geographic distances between metagenomes with the earth.dist function from the fossil R package (Vavrek 2011), and performed
the Moran’s I test with the Moran.I function from the ape R package (Paradis, Claude, and Strimmer 2004).

We computed the taxonomic and functional Bray-Curtis dissimilarity matrices to be used in the partial Mantel tests, using the Hellinger transformed functional and taxonomic composition. Additionally, for each mg-trait, we computed a Euclidean distance matrix. We then performed a series of partial Mantel tests to compare the mg-trait dissimilarities with the functional (and taxonomic) dissimilarities controlling by the taxonomic (and functional) dissimilarities. To compute the partial Mantel tests, we used the function mantel.partial from the vegan R package (Oksanen et al. 2017).

To compute the two-way clustering heat map, we first scaled all mg-traits (i.e., mean = 0 and standard deviation = 1). The sample clustering was performed on the Euclidean distances of the scaled mg-traits using Ward’s method. The mg-trait clustering was computed using a correlation-based dissimilarity measure derived as $d_{i,j} = 1 - |r_{i,j}|$, where $d_{i,j}$ is the dissimilarity between mg-trait ”i” and ”j”, and $|r_{i,j}|$ is the absolute value of the Pearson’s r correlation between mg-trait ”i” and ”j”.

To test the differentiation of the mg-trait value distributions between the SRF, DCM and MES water layers, we applied a paired Wilcoxon rank sum test on each pair of water layers, using the function wilcox.test from the vegan R package. In this analysis, we used a matching subset composed of 63 metagenomes, obtained from the three water layers in 21 sampling stations (Layers-Subset). We performed a Permutational Multivariate Analysis of Variance (PERMANOVA) to evaluate the functional and taxonomic compositional differences between water layers and ocean regions. To test the differentiation between water layers, we used the compositional data of the metagenomes from the Layers-Subset. To test the differentiation between geographic locations, we used a subset of 57 surface metagenomes from six ocean regions: Indian Ocean, North Pacific Ocean, South Pacific Ocean, North Atlantic Ocean, South Atlantic Ocean, and
the Mediterranean Sea (i.e., Regions-Subset). The PERMANOVA tests were applied on the (Hellinger transformed) taxonomic and functional composition matrices of each dataset, using the adonis function of the vegan R package with 999 permutations. To account for multiple comparisons, we applied the Bonferroni correction to the p-values obtained from the pairwise PERMANOVA tests.

Lastly, based on the Hellinger transformed functional and taxonomic composition of the Layers- and Regions-Subset, we computed the respective Bray-Curtis dissimilarity matrices, and performed a Principal Coordinate Analysis (PCoA) on each matrix.

3. Implementation

The workflow implemented in Mg-Traits-PL consists of the following tasks: 1) Computing the mean and variance of the GC content, and the dinucleotide frequency using the infoseq and compseq tools of the EMBOSS package (Rice, Longden, and Bleasby 2000); 2) Predicting the Open Reading Frames (ORFs) with FragGeneScan-plus (Kim et al. 2015) to compute the codon and amino acid usage frequencies using the cusp tool of the EMBOSS package, and the acidic to basic ratio following Rhodes et al. (2010) approach; 3) Annotating the Pfam (Finn et al. 2016) and Biosynthetic Gene Cluster domains in the ORF amino acid sequences with the Ultra-fast Protein Classification tool UProC (Meinicke 2015), using the Pfam UProC database version 28.0 obtained from http://uproc.gobics.de and the BGC domain UProC database created in Pereira-Flores et al. (in prep.) (see section 2.2). The Pfam annotation is used to predict the functional composition and the Shannon diversity index, and the percentage of transcription factors and functionally annotated reads. The latter two mg-traits are calculated based on the total number of ORF sequences; 4) Computing the Average Genomes Size (AGS) using the ags.sh tool (Pereira-Flores, Glöckner, and Fernández-Guerra 2019) (see section 2.3); 5) Identifying
the 16S rRNA gene sequences with SortMeRNA (Kopylova, Noé, and Touzet 2012) to perform a taxonomic annotation using the SILVA Incremental Aligner (SINA) (Pruesse, Peplies, and Glöckner 2012), based on the non-redundant (NR99) reference multiple sequence alignment of the SILVA database release 128 (Quast et al. 2013). The taxonomic annotation is used to estimate the taxonomic composition at the order level and the Shannon diversity index. In addition, based on the identified 16S rRNA gene sequences, the acn.sh (Pereira-Flores, Glöckner, and Fernández-Guerra 2019) (see section 2.3) is used to estimate the 16S rRNA gene Average Copy Number per genome (ACN).

The mg-trait values computed by Mg-Traits-PL are stored in Mg-Traits-DB, which is based on the PostgreSQL management system (https://www.postgresql.org). In addition, this database contains the metadata associated to the metagenomes, including the geographic coordinates, water depth, date, salinity, temperature, and the biome, material and feature classes of the Environmental Ontology (Buttigieg et al. 2016).
Chapter 2. Results and Discussion

References


Chapter 2. Results and Discussion


Chapter 2. Results and Discussion


Chapter 2. Results and Discussion

Figures

**Mg-Traits-PL**

![Diagram of Mg-Traits-PL](image)

**Figure 2.4-1: Mg-Trait-PL.** The Mg-Traits-PL computes 15 mg-trait that are grouped into four different categories. 1) **Direct mg-trait**: the GC content and variance, and the dinucleotide frequency are computed based on the metagenomic short-read sequence data. 2) **ORF-derived mg-trait**: the Open Reading Frames (ORFs) are predicted in the metagenomic short-read sequence data, and used to compute the codon and amino acid frequencies, and the Acidic to Basic ratio (AB ratio). 3) **Functional annotation-derived mg-trait**: the Pfam and BGC domains are annotated in the ORF amino acid sequences with the UProC tool (Meinicke 2015), to compute the functional and BGC domain composition, the functional diversity, and the percentage of functionally annotated reads (Perc fun annot) and transcription factors (Perc TF). In addition, the Average Genome Size (AGS) is computed using the ags.sh tool (Pereira-Flores, Glöckner, and Fernández-Guerra 2019). 4) **16S rRNA gene-derived mg-trait**: the 16S rRNA genes are identified with the SortmerRNA tool (Kopylova, Noé, and Touzet 2012), and taxonomically annotated with the SILVA Incremental Aligner (SINA) (Pruesse, Peplies, and Glöckner 2012), to estimate the taxonomic composition and diversity. Lastly, the identified ribosomal gene sequences are used to compute the 16S rRNA gene Average Copy Number (ACN) with the acn.sh tool (Pereira-Flores, Glöckner, and Fernández-Guerra 2019).
Figure 2.4-2: Two-way clustering heat map of the 139 prokaryotic TARA Oceans metagenomes based on the mg-trait profiles. The 15 mg-trait were scaled (mean = 0 and standard deviation = 1) to generate the two-way clustering heat map based on the mg-trait metagenome profiles. To compute the sample clustering, we applied Ward’s method on the Euclidean distances computed based on the scaled mg-trait profiles. To cluster the variables, we applied Ward’s method on the correlation-based dissimilarities computed as $d_{i,j} = 1 - |r_{i,j}|$, where $d_{i,j}$ is the dissimilarity between mg-trait "i" and "j", and $|r_{i,j}|$ is the absolute value of the Pearson’s r correlation between mg-trait "i" and "j".

---

Chapter 2. Results and Discussion
Chapter 2. Results and Discussion

Figure 2.4-3: Mg-Traits boxplots. The 63 metagenomes of the Layers-Subset were used to compare the mg-trait value distributions between the surface, deep chlorophyll maximum and mesopelagic water layers (SRF, DCM, and MES, respectively). The mg-trait boxplots marked with an asterisk differed significantly between the three water layers (paired Wilcoxon rank sum test; p-value < 0.05; see Supplementary Table 6.2-2).
Chapter 2. Results and Discussion

Figure 2.4-4: Differentiation of water layers and ocean regions based on their taxonomic and functional composition. (a) Principal Coordinate Analyses (PCoAs) performed on the functional and taxonomic (Bray-Curtis) composition dissimilarities of the 63 metagenomes of the Layers-Subset. (b) PCoAs performed on the functional and taxonomic composition dissimilarities of the 57 surface metagenomes of the Regions-Subset. To compute the dissimilarity matrices, the functional and taxonomic composition matrices were transformed with the Hellinger function. Metagenomes are colored and grouped according to the water layer or ocean region they were sampled from. The acronyms IO, MS, NAO, NPO, SAO, and SPO, stand for Indian Ocean, Mediterranean Sea, North Atlantic Ocean, North Pacific Ocean, South Atlantic Ocean, and South Pacific Ocean, respectively.
Chapter 2. Results and Discussion

Tables

Table 2.4-1: Computed mg-traits. The 15 mg-traits computed by Mg-Traits-PL are classified into four categories: direct mg-traits, ORF-derived mg-traits, functional annotation-derived mg-traits, and 16S rRNA gene-derived mg-traits.

<table>
<thead>
<tr>
<th>Category</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Direct mg-traits</td>
<td>GC content</td>
</tr>
<tr>
<td>2 Direct mg-traits</td>
<td>GC variance</td>
</tr>
<tr>
<td>3 Direct mg-traits</td>
<td>Dinucleotide frequency (Dinuc freq)</td>
</tr>
<tr>
<td>4 ORF-derived mg-traits</td>
<td>Codon frequency (Codon freq)</td>
</tr>
<tr>
<td>5 ORF-derived mg-traits</td>
<td>Amino acid frequency (AA freq)</td>
</tr>
<tr>
<td>6 ORF-derived mg-traits</td>
<td>Acid to basic amino acid ratio (AB ratio)</td>
</tr>
<tr>
<td>7 Functional annotation-derived mg-traits</td>
<td>Functional composition (Fun comp)</td>
</tr>
<tr>
<td>8 Functional annotation-derived mg-traits</td>
<td>Functional diversity (Fun div)</td>
</tr>
<tr>
<td>9 Functional annotation-derived mg-traits</td>
<td>Percentage functionally annotated reads (Perc fun annot)</td>
</tr>
<tr>
<td>10 Functional annotation-derived mg-traits</td>
<td>Percentage of transcription factors (Perc TF)</td>
</tr>
<tr>
<td>11 Functional annotation-derived mg-traits</td>
<td>BGC domain composition (BGC comp)</td>
</tr>
<tr>
<td>12 Functional annotation-derived mg-traits</td>
<td>Average Genome Size (AGS)</td>
</tr>
<tr>
<td>13 16S rRNA gene-derived mg-traits</td>
<td>Taxonomic composition (Taxa comp)</td>
</tr>
<tr>
<td>14 16S rRNA gene-derived mg-traits</td>
<td>Taxonomic diversity (Taxa div)</td>
</tr>
<tr>
<td>15 16S rRNA gene-derived mg-traits</td>
<td>16S rRNA gene Average Copy Number (ACN)</td>
</tr>
</tbody>
</table>
Chapter 2. Results and Discussion

Table 2.4-2: Mg-Traits-DB datasets. Details of the three metagenomic datasets included in Mg-Traits-DB: Global Ocean Sampling expedition (GOS), Ocean Sampling Day 2014 (OSD) and TARA Oceans expedition.

<table>
<thead>
<tr>
<th></th>
<th>GOS</th>
<th>OSD</th>
<th>TARA Oceans</th>
</tr>
</thead>
<tbody>
<tr>
<td># of metagenomes</td>
<td>83</td>
<td>150</td>
<td>243</td>
</tr>
<tr>
<td># of metagenomes included in MG-traits-DB</td>
<td>68</td>
<td>149</td>
<td>139</td>
</tr>
<tr>
<td>Sequencing technology</td>
<td>Sanger</td>
<td>Illumina MiSeq</td>
<td>Illumina HiSeq 2000</td>
</tr>
<tr>
<td>ENA study accession</td>
<td>PRJNA13694</td>
<td>PRJEB8682</td>
<td>PRJEB7988</td>
</tr>
<tr>
<td>Fraction size (µm)</td>
<td>0.1 - 0.8</td>
<td>0 - 0.22</td>
<td>0.22 - 1.6; 0.22 - 3</td>
</tr>
</tbody>
</table>

Table 2.4-3: Multi-valued traits summary variables. Details of the first principal components (PC1s) obtained from the Principal Component Analysis (PCA) performed on the multi-valued traits.

<table>
<thead>
<tr>
<th>Mg-Trait</th>
<th>PC1 captured variability</th>
<th># of variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dinuc freq</td>
<td>96.0%</td>
<td>16</td>
</tr>
<tr>
<td>Codon freq</td>
<td>91.3%</td>
<td>61</td>
</tr>
<tr>
<td>AA freq</td>
<td>87.18%</td>
<td>20</td>
</tr>
<tr>
<td>Fun comp</td>
<td>44.9%</td>
<td>16,230</td>
</tr>
<tr>
<td>BGC comp</td>
<td>63.87%</td>
<td>150</td>
</tr>
<tr>
<td>Taxa comp</td>
<td>49.35%</td>
<td>499</td>
</tr>
</tbody>
</table>
Table 2.4-4: Spatial autocorrelation tests. Details of the Moran’s I spatial autocorrelation tests performed on each of the 15 mg-trait. The surface, deep chlorophyll maximum and mesopelagic water layers (SRF, DCM, and MES, respectively), were analyzed independently. P-values lower than 0.05 are shown in boldface.

<table>
<thead>
<tr>
<th>Mg-trait</th>
<th>Moran’s I (SRF samples)</th>
<th>Moran’s p-value (SRF samples)</th>
<th>Moran’s I (DCM samples)</th>
<th>Moran’s p-value (DCM samples)</th>
<th>Moran’s I (MES samples)</th>
<th>Moran’s p-value (MES samples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC content</td>
<td>0.078</td>
<td>9.64e-03</td>
<td>0.003</td>
<td>5.16e-01</td>
<td>0.09</td>
<td>4.57e-02</td>
</tr>
<tr>
<td>GC variance</td>
<td>0.117</td>
<td>4.64e-04</td>
<td>0.118</td>
<td>9.48e-04</td>
<td>0.026</td>
<td>2.57e-01</td>
</tr>
<tr>
<td>Dinuc PC1</td>
<td>0.078</td>
<td>1.02e-02</td>
<td>-0.01</td>
<td>7.31e-01</td>
<td>0.063</td>
<td>1.18e-01</td>
</tr>
<tr>
<td>Codon PC1</td>
<td>0.072</td>
<td>1.57e-02</td>
<td>-0.007</td>
<td>6.81e-01</td>
<td>0.066</td>
<td>1.06e-01</td>
</tr>
<tr>
<td>AA PC1</td>
<td>0.068</td>
<td>2.16e-02</td>
<td>-0.038</td>
<td>7.56e-01</td>
<td>0.031</td>
<td>2.92e-01</td>
</tr>
<tr>
<td>AB ratio</td>
<td>0.161</td>
<td>3.23e-06</td>
<td>0.163</td>
<td>1.1e-05</td>
<td>-0.028</td>
<td>9.18e-01</td>
</tr>
<tr>
<td>Fun PC1</td>
<td>0.232</td>
<td>6.91e-11</td>
<td>0.141</td>
<td>9.14e-05</td>
<td>-0.063</td>
<td>6.05e-01</td>
</tr>
<tr>
<td>Fun div</td>
<td>0.091</td>
<td>4.49e-03</td>
<td>0.01</td>
<td>4.28e-01</td>
<td>0.154</td>
<td>2.36e-03</td>
</tr>
<tr>
<td>Perc fun annot</td>
<td>0.056</td>
<td>5.28e-02</td>
<td>0.008</td>
<td>4.54e-01</td>
<td>-0.047</td>
<td>8.38e-01</td>
</tr>
<tr>
<td>Perc TF</td>
<td>0.19</td>
<td>2.95e-08</td>
<td>0.146</td>
<td>4.72e-05</td>
<td>-0.042</td>
<td>8.88e-01</td>
</tr>
<tr>
<td>BGC PC1</td>
<td>0.16</td>
<td>3.61e-06</td>
<td>0.037</td>
<td>1.42e-01</td>
<td>-0.077</td>
<td>4.50e-01</td>
</tr>
<tr>
<td>AGS</td>
<td>0.053</td>
<td>5.04e-02</td>
<td>0.022</td>
<td>1.78e-01</td>
<td>-0.011</td>
<td>6.95e-01</td>
</tr>
<tr>
<td>Taxa PC1</td>
<td>-0.007</td>
<td>8.17e-01</td>
<td>-0.016</td>
<td>8.40e-01</td>
<td>-0.029</td>
<td>9.27e-01</td>
</tr>
<tr>
<td>Taxa div</td>
<td>0.114</td>
<td>5.27e-04</td>
<td>0.05</td>
<td>7.87e-02</td>
<td>0.2</td>
<td>1.03e-04</td>
</tr>
<tr>
<td>ACN</td>
<td>0.066</td>
<td>2.51e-02</td>
<td>-0.041</td>
<td>6.60e-01</td>
<td>-0.05</td>
<td>7.39e-01</td>
</tr>
</tbody>
</table>
Table 2.4-5: Partial Mantel tests. For each of the 15 mg-traits, we performed a partial Mantel test to compare the mg-traits dissimilarities (Euclidean distance) with the functional (or taxonomic) dissimilarities (Bray-Curtis) controlling for the taxonomic (or functional) dissimilarities. P-values lower than 0.05 are shown in boldface.

<table>
<thead>
<tr>
<th>Mg-trait</th>
<th>Mantel r Fun</th>
<th>Taxa</th>
<th>Mantel p-value Fun</th>
<th>Taxa</th>
<th>Mantel r Taxa</th>
<th>Fun</th>
<th>Mantel p-value Taxa</th>
<th>Fun</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC content</td>
<td>0.441</td>
<td>1.00e-03</td>
<td>-0.178</td>
<td>1.00e+00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GC variance</td>
<td>0.105</td>
<td>6.00e-03</td>
<td>-0.065</td>
<td>8.92e-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dinuc PC1</td>
<td>0.478</td>
<td>1.00e-03</td>
<td>-0.197</td>
<td>1.00e+00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Codon PC1</td>
<td>0.528</td>
<td>1.00e-03</td>
<td>-0.203</td>
<td>1.00e+00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AA PC1</td>
<td>0.476</td>
<td>1.00e-03</td>
<td>-0.197</td>
<td>1.00e+00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB ratio</td>
<td>0.075</td>
<td>4.10e-02</td>
<td>0.062</td>
<td>1.36e-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fun PC1</td>
<td>0.741</td>
<td>1.00e-03</td>
<td>NA</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fun div</td>
<td>0.241</td>
<td>1.00e-03</td>
<td>-0.12</td>
<td>9.95e-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perc fun annot</td>
<td>0.222</td>
<td>1.00e-03</td>
<td>-0.108</td>
<td>9.87e-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perc TF</td>
<td>0.459</td>
<td>1.00e-03</td>
<td>-0.184</td>
<td>1.00e+00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BGC PC1</td>
<td>0.672</td>
<td>1.00e-03</td>
<td>-0.235</td>
<td>1.00e+00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGS</td>
<td>0.265</td>
<td>1.00e-03</td>
<td>-0.083</td>
<td>9.42e-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taxa PC1</td>
<td>NA</td>
<td>NA</td>
<td>-0.11</td>
<td>9.89e-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taxa div</td>
<td>0.311</td>
<td>1.00e-03</td>
<td>-0.004</td>
<td>5.28e-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACN</td>
<td>0.28</td>
<td>1.00e-03</td>
<td>-0.144</td>
<td>9.98e-01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 3

Final remarks and conclusions

This thesis contributed to two lines of research: metagenomic mining of natural product biosynthetic gene clusters and the application of functional trait-based approaches in metagenomics. This chapter provides final remarks and draws general conclusions on the main findings of these two lines of research.

3.1 Metagenomic mining of natural product biosynthetic gene clusters

In section 2.2, we present the study titled "Mining metagenomes for natural product biosynthetic gene clusters: unlocking new potential with ultrafast techniques". This work aims to contribute to the exploration of Biosynthetic Gene Clusters (BGCs) in metagenomic data. BGCs are responsible for the production of a large number of natural products with biotechnological applications, and are also interesting from an ecological perspective, as these participate in organisms’ interactions and adaptation to the environment (Fischbach and Voigt 2010). Although metagenomics offers unique advantages to study the BGCs encoded by microorganisms, it also poses significant challenges. These can be summarized as follows: 1) Metagenomes consist of a very
large number of short-read sequences, commonly obtained from a diverse community, and for which their taxonomic origin and genome coordinates are unknown; 2) BGC sequences can extend more than 100 kbp, making their identification very difficult (or impossible) in short-read sequence data; 3) Commonly, a tiny fraction of the metagenomic data can be assembled in contigs long enough to allow the identification of BGCs. As a consequence, current tools dedicated to the BGC chemical class exploration can only analyze large (meta)genomic contigs. On the other hand, although BGC domains can be annotated in short-read sequences, tools dedicated to the analysis of these domains only focus on Polyketide and Non-ribosomal Peptide Synthetases (PKS and NRPS, respectively) (Weber and Kim 2016). In conclusion, although metagenomic data allows us to access the coding potential of microorganism in nature, it remains mostly underexploited in terms of BGC research. To address this challenge, we developed the Biosynthetic Gene Cluster Metagenomic Exploration toolbox (BiG-MEx). The main contribution of BiG-MEx is that it performs ultra-fast profiling of BGC domains and chemical classes in unassembled metagenomic data, endowing this toolbox with an unprecedented capacity to explore the untapped diversity of BGCs encoded by microorganisms. In addition, BiG-MEx extends previous approaches dedicated to the PKS and NRPS domain diversity and phylogenomic analysis, to several domains of other BGC chemical classes, while also scaling the methodology to cope with large amounts of data. Namely, BiG-MEx can be used in metagenomic studies to estimate the BGC domain and class composition and diversity, generate Operational Domain Unit (ODU) abundance tables and determine domain phylogenetic associations. The interrogation of this data is highly valuable to understand the ecological roles of BGCs in Nature and delineate the environmental factors that associate with changes in the BGC composition. Additionally, such analyses can contribute to natural product discovery efforts. Besides the exploratory applications, BiG-MEx can be used to screen and prioritize (meta)genomic samples, when looking for specific BGC natural products of interest. To this end, taxon- or environment-specific BGC class abundance models can be used
to generate ultra-fast BGC class profiles in swallow sequencing depth (meta)genomic data.

There is a growing number of studies using amplicon data of BGC domains to explore the biosynthetic capacity of microbial communities (e.g., Charlop-Powers et al. 2014; Charlop-Powers et al. 2015; Lemetre et al. 2017; Wei et al. 2018). In these works, the BGC domains are amplified from environmental DNA, allowing fine-grained analyses of their diversity and distribution in the environment. However, it is important to stress that compared with amplicon data, metagenomic data has certain advantages to explore the biosynthetic capacity of microbial communities. Metagenomic data can be used to simultaneously analyze a large number of domains (currently 150 with BiG-MEx), which represent the BGC composition of a microbial community. Also, although metagenomic data can suffer from different biases (e.g., associated to library preparation procedures (Jones et al. 2015)), it is not subjected to the PCR bias introduced when amplifying the target sequences from environmental DNA (Lee et al. 2012). Furthermore, metagenomic data contains functional and taxonomic information, which is fundamental to perform a comprehensive study of the ecology of these genes. On top of these advantages, the dramatic reduction in high-throughput sequencing costs makes the generation of metagenomic data economically feasible for any standard laboratory (Logares et al. 2014).

An interesting finding obtained when we applied BiG-MEx to explore metagenomes of the TARA Oceans (Sunagawa et al. 2015) and Human Microbiome Project datasets (The Human Microbiome Project Consortium 2012), is that the BGC domain and class composition can significantly differentiate types of environments. This result highlights the importance of BGCs in the adaptation of microorganisms to the environmental conditions which they inhabit and stresses the role of BGCs as functional traits (i.e., organisms’ characteristics linked to their performance). In this context, BiG-MEx integrates into a common thread of research connecting the different studies of this thesis, as an effort to contribute to the development of tools and the exploration of functional
traits in metagenomic data.

3.2 Functional trait-based approaches in metagenomics

In sections 2.3 and 2.4, we present the thesis work dedicated to the application of functional trait-based approaches in metagenomics. Functional trait data can be of particular advantage to study microorganisms’ ecology. The quantification of functional traits can be implemented as integrative functions, as is the case of community aggregated functional traits (i.e., traits measured in a random sample of individuals in a community irrespective of their taxonomic affiliation). These traits are remarkably well suited to study community microbial ecological processes based on metagenomic data. Indeed, functional traits quantified in a complete metagenomic sample, are community aggregated functional traits measured at the genome level. These provide an ecological framework to exploit metagenomic data further to gain insights into ecosystem-level processes and identify the factors that determine microorganisms’ biogeography, diversity patterns, and community dynamics (Violle et al. 2012; Fierer, Barberán, and Laughlin 2014).

Within this line of work, we include the study titled “Fast and accurate average genome size and 16S rRNA gene average copy number computation in metagenomic data”. This work presents the ags.sh and acn.sh tools dedicated to the analytical computation of the Average Genome Size (AGS) and 16S rRNA gene Average Copy Number (ACN), respectively. These metagenomic traits provide a rich source of information to study the dominant ecological strategies in a microbial community: the AGS and ACN traits can be utilized to position microbial communities in axes of variation representing ecological strategies ranging from specialist to generalist and from copiotroph to oligotroph (Cobo-Simón and Tamames 2017). The ags.sh and acn.sh tools are based on recent advances in gene annotation approaches, which allow fast and accurate
computation of these traits in unassembled metagenomic data, and in turn the character-
ization of the ecological strategies within a microbial community. Using these tools, we
analyzed TARA Oceans metagenomes and revealed the ecological strategies asso-
ciated with the different environmental conditions that occur in the water column.

In addition, we present the study titled “Mg-Traits pipeline and database: advancing
functional trait-based approaches in metagenomics”. In this study, the ags.sh and
acn.sh tools, together with BiG-MEx domain profiling module and the scripts dedicated
to the computation of other 12 traits (ranging from GC content to functional diversity),
were integrated into a single pipeline named Mg-Traits-PL. As such, Mg-Traits-PL fa-
cilitates the computation of a comprehensive set of metagenomic traits consistently.

We used Mg-Traits-PL to compute the 15 metagenomic traits in metagenomes of the
TARA Oceans, Ocean Sampling Day 2014 and Global Ocean Sampling expedition
datasets to create the database Mg-Traits-DB. Further, we used the metagenomic trait
data of TARA Oceans metagenomes, to characterize the microbial communities asso-
ciated with different water layers. The combined analysis of metagenomic traits like
the BGC domain and functional composition, average genome size, 16S rRNA gene
average copy number, percentage of transcription factors, and GC content and var-
iance, revealed marked differences between microbial communities from the surface,
deep chlorophyll maximum, and mesopelagic water layers. Such mg-trait analysis ex-
pands the results obtained in sections 2.2 and 2.3 to provide a more comprehensive
picture of the ecology of microorganisms along the water column. In essence, all the
analyses performed in TARA Oceans metagenomes showed differences between mi-
crobial communities, that indicate contrasting life history strategies between surface
and deep-sea environments (i.e., K- and r-selected, respectively).

It is evident that accurate and comprehensive contextual data is fundamental to the
analysis and discoverability of metagenomic data. In particular, high-quality environ-
mental data of the metagenomic sampling sites is critical to fully realize the potential
of the tools developed in this thesis. For these reasons, this Ph.D. project also con-
tributed to the improvement of contextual data associated with sequence data, by col-
laborating in the performance assessment of the tool EXTRACT (Pafilis et al. 2016), and participating in the Sample Record Annotation Workshop (Hoopen et al. 2016). The tool EXTRACT is designed to assist the identification of standard-compliant terms (e.g., ENVO (Buttigieg et al. 2016) and NCBI Taxonomy terms (Sayers et al. 2019)) to annotate biological samples. The Sample Record Annotation Workshop consisted of curating contextual data of metagenomic samples from the European Nucleotide Archive (ENA) (Harrison et al. 2019) to assess the benefits of this work.

In sum, the work of this thesis contributed to the ecological analyses of metagenomic data, by developing tools for the computation of different metagenomic traits, and assisting the curation of contextual data. The interrelationships between metagenomic traits and with the environmental parameters constitute a powerful approach particularly well suited for quantitative and predictive studies, which are currently needed to advance our understanding of the ecology of microorganisms.
Chapter 4

Outlook

Current metagenomic projects can generate large amounts of sequence data and are expected to generate even more in the near future. Large and diverse volumes of metagenomic data are fundamental to improve our understanding of the microbial world. In particular, these offer an excellent opportunity to mine natural environments for Biosynthetic Gene Clusters (BGCs).

In this thesis, we contributed to the metagenomic mining of BGCs by developing the toolbox BiG-MEx, which allows a rapid and comprehensive exploratory analysis of large volumes of data. Some of the most promising applications of BiG-MEx are studying the biogeography and ecology of BGCs, and exploring the biosynthetic capacity of microbial communities from underexploited environments. Marine environments are particularly attractive for such applications. Marine microbial communities represent a rich source of chemical diversity, which remains largely underexploited (Trindade et al. 2015). Given that some of the biggest sequencing efforts are from marine environments (e.g., Ocean Sampling Day, TARA Oceans, Malaspina (Duarte 2015), BiOGEOTRACES (Biller et al. 2018)), the exploration of BGCs in marine microbial communities is very timely. For example, metagenomes of the TARA Oceans, Malaspina and Ocean Sampling Day datasets, could be used to compare the BGC composition between open-ocean, coastal, surface and deep-sea environments. Moreover, other chemically rich environments like host-associated and soil would be of particular in-
terest to perform further comparative analyses. For such applications, BiG-MEx could be used to analyze the distribution, co-occurrence, and diversity of ODUs, and BGC domains and classes. These analyses could then be integrated with the taxonomic and environmental data to identify associations between BGCs at different levels (i.e., ODU, domain, and class levels) and with specific environmental conditions and taxonomic groups, to characterize the ecology of these genomic elements. Additionally, BiG-MEx could be used to perform more fine-grained analyses of specific BGC domains and classes. For instance, extending the biogeographic analysis of polyketide and nonribosomal peptide synthetase domains performed by Lemetre et al. (2017) in soils, to other environments and BGC domains.

Lastly, BiG-MEx, by providing a rapid BGC characterization of metagenomic samples, could be valuable in approaches combing low-cost short-read sequencing with more precise but also more expensive technologies. For example, BiG-MEx could be used to prioritize samples, BGC sequences, or taxonomic groups for a more fine-grained examination using single-molecule long-read technologies or single-cell genomics.

It is important to stress that before BiG-MEx there was no technology available to systematically and comprehensively analyze large volumes of metagenomic data. Accordingly, we envision that by complementing other powerful tools like the Biosynthetic Gene Similarity Clustering and Prospecting Engine (BiG-SCAPE), the CORe Analysis of Syntenic Orthologues to prioritize Natural products biosynthetic gene clusters (CORASON) software (Navarro-Muñoz et al. 2019), the BGC assembler BiosyntheticSPAdes (Meleshko et al. 2019), and the pipeline antiSMASH, BiG-MEx will become part of a central toolbox for BGCs research.

On the other hand, in this thesis, we contributed to microorganisms’ ecology studies based on metagenomic data, by developing the Average Genome Size (AGS) and 16S rRNA gene Average Copy Number (ACN) computation tools and the Mg-Traits pipeline and database. Follow-up studies within this line of work will include the delineation of ecological strategies associated with contrasting environments based on the computation of the AGS and ACN traits (e.g., Cobo-Simón and Tamames 2017). Also, an
interesting future application of Mg-Traits pipeline is the identification of metagenomic traits relevant to the classification of specific environmental attributes. For example, the application of machine learning methods to rank metagenomic traits according to their capacity to differentiate healthy vs. ill-associated human gut microbiomes or polluted vs. healthy environments. Further, the Mg-Traits pipeline could be of value to monitor changes and study microbial community successions in time-series studies, and genomic observatories. For instance, the Mg-Traits pipeline could be integrated into the analysis workflows of the metagenomic data generated by the OSD sequencing campaign, and the bacterioplankton succession study being carried out on the coast of Uruguay (http://marinegenomics.interior.edu.uy). These studies are critical to advance our understanding of microbial communities’ responses to environmental changes. This is a highly relevant topic, considering the implications of microorganisms on the ecosystem, and the fact that the environmental degradation and disruption of planetary ecosystem processes derived from human activities are likely to worsen in the near future (De Saedeleer 2016; Molotoks et al. 2018).

Regarding future improvements in the Mg-Traits pipeline, we plan to implement the computation of taxon-specific traits (based on taxonomically binned metagenomic data), which in turn, could be used to estimate trait value distributions and informative trait statistics like the median, mode, skewness, and standard deviation (Violle et al. 2012; Loranger et al. 2018).

Currently, there is a need to develop new bioinformatics algorithms and tools to efficiently analyze the large amounts of short-read metagenomic data generated by NGS technologies. The work in this thesis responded to such demand, by developing a series of tools that improve the exploitation of metagenomic data. In a broader context, advances in bioinformatics and metagenomic approaches will yield new possibilities to address challenges in diverse areas, ranging from human health to environmental pollution. Furthermore, developments in this area will allow large-scale integration studies to advance our understanding of global ecosystem processes. The future will certainly be a highly exciting time to work in bioinformatics.
Chapter 5

Additional scientific publications

1. **EXTRACT**: interactive extraction of environment metadata and term suggestion for metagenomic sample annotation

   **Authors:** Evangelos Pafilis, Pier Luigi Buttigieg, Barbara Ferrell, Emiliano Pereira, Julia Schnetzer, Christos Arvanitidis and Lars Juhl Jensen

   **Published in Database; Volume 2016; Year 2016.**

   **Contribution:** Performed the user evaluation of EXTRACT to assess the improvements in terms of speed and accuracy when curating metadata.

   **Relevance:** The metadata of a biological sample is essential to perform comparative analyses and improve its discoverability. EXTRACT is an interactive annotation tool designed to help curators identify and extract standard-compliant terms to annotate metadata using online texts, associated with biological samples.
2. **Value, but high costs in post-deposition data curation**  
   **Authors:** Petra ten Hoopen, Clara Amid, Pier Luigi Buttigieg, Evangelos Pafilis, Panos Bravakos, Ana M. Cerdeño-Tarraga, Richard Gibson, Tim Kahlke, Aglaia Legaki, Kada Narayana Murthy, Gabriella Papastefanou, Emiliano Pereira, Marc Rossello, Ana Luisa Toribio and Guy Cochrane.  
   **Published in Database; Volume 2016; Year 2016.**  
   **Contribution:** Participated in the Sample Record Annotation Workshop as a metadata curator of metagenomic samples, to estimate the effort needed to improve the quality of the metadata.  
   **Relevance:** Complete and accurate metadata of biological samples is frequently lacking in publicly available datasets. Metadata is essential for the interpretation and re-utilization of biological data. This work describes an exercise, which consisted of curating metadata associated with metagenomic samples available in the European Nucleotide Archive, to assess the costs and time needed to improve the quality of the metadata.

3. **The North Sea goes viral: Occurrence and distribution of North Sea bacteriophages**  
   **Authors:** Alexa Garin-Fernández, Emiliano Pereira-Flores, Frank Oliver Glöckner, Antje Wichels  
   **Published in Marine Genomics; Volume 41; Year 2018.**  
   **Contribution:** Performed the bioinformatics analyses (together with AGF).  
   **Relevance:** Viruses have a significant influence on marine microbial population dynamics, participating in cell lysis and horizontal gene transfer processes. This study utilizes virus metagenomic data to analyze bacteriophages populations in the North Sea and gains insight into the environmental factors influencing their biogeography.
Chapter 6

Appendix: supplementary materials

6.1 Supplementary material of the work “Mining metagenomes for natural product biosynthetic gene clusters: unlocking new potential with ultra-fast techniques"

6.1.1 Supplementary figures
Supplementary Figure 6.1-1: Benchmarking UProC running time and evaluating the ODU diversity estimation approach.

(a) Comparison of the wall-clock running time between uproc-prot and hmmsearch for the annotation of 150 BGC domains in nine TARA Oceans metagenomes, three from each of the Surface (SRF), Deep Chlorophyll Maximum (DCM) and Mesopelagic (MES) water layers. The annotations were performed four times, increasing the thread number in each run (i.e., 4, 8, 16 and 32 threads). Uproc-prot was on average 18 times faster than hmmsearch, and scaled better with increasing thread number. The largest difference was observed for 32 threads, where uproc-prot was 25 times faster than hmmsearch. (b) Comparison of the Operational Domain Unit (ODU) diversity estimated with BiG-MEx vs. the ODU diversity computed based on complete domain sequences (i.e., reference diversity). In this analysis, we estimated the ODU Shannon diversity of the adenylation (AMP-binding), Condensation, Ketosynthase (KS) and Acyltransferase (AT) domains in the simulated Marine-TM metagenomic dataset. Complete domain sequences were annotated and extracted from the genome sequences used to simulate the Marine-TM metagenomes.
Supplementary Figure 6.1-2: Validation of BiG-MEx BGC class abundance predictions. Scatter plots comparing the reference and predicted abundances for 23 BGC classes. MdAPE stands for Median Absolute Percentage Error. The black line represents the one-to-one relationship between the reference and predicted BGC class abundances. The BGC class abundance models trained with the Marine-RM metagenomic dataset were used to predict the abundance in the Marine-TM metagenomic dataset. The reference abundance was computed from the annotation of the BGC classes with antiSMASH on the complete genome sequences from which the Marine-TM metagenomes were simulated.
Supplementary Figure 6.1-3: Benchmarking BiG-MEx BGC class abundance predictions. Plots of the Pearson correlation coefficients (upper panel) and the absolute percentage error distributions (lower panel) of the BGC class abundances predicted by the models and estimated by the assembly approach, with respect to the reference abundances. For the sake of clarity, 41 metagenomes (2.05% of the total data) with an outlier absolute percentage error value were excluded from the plot. For this analysis, we assembled 50 metagenomes of the Marine-TM dataset to estimate the BGC class abundance (assembly approach).
Supplementary Figure 6.1-4: TARA Oceans BGC domain composition analysis. Principal Coordinates Analysis (PCoA) performed on a Bray-Curtis dissimilarity matrix of the BGC domain relative abundance profiles of the 139 prokaryotic metagenomes of TARA Oceans. The boxplot in the bottom section of the panel shows the PCo1 value distributions for the SRF, DCM and MES water layers. The PCo1 axis differentiated the MES water layer from the other two layers (Wilcoxon rank sum test: all p-values < 0.0001; see Supplementary Table 6.1-2).
Supplementary Figure 6.1-5: Barplots showing the BGC class relative abundance composition. BiG-MEx models were used to predict the abundance of 24 BGC classes in the matching subset of 63 metagenomes of TARA Oceans. The metagenomes from Surface, Deep Chlorophyll Maximum and Mesopelagic water layers are displayed in three different plots (upper, middle and lower panels, respectively).
Supplementary Figure 6.1-6: Geographically mapped BGC class abundances. Relative abundances are represented as coxcomb plots. BiG-MEx modes were used to predict the abundances of 24 BGC classes in TARA Oceans metagenomes. The metagenomes from Surface, Deep Chlorophyll Maximum and Mesopelagic water layers are displayed in three different plots (upper, middle and lower panels, respectively). Only sampling sites separated by more than 1000 km are shown.
**Supplementary Figure 6.1-7: Sequence length comparison.** Boxplot showing the sequence length distribution of BGCs from MIBiG database, the Metagenome Assembled Genomes (MAGs) of TARA Oceans, and 500,000 randomly subsampled contigs from the assemblies of the nine TARA Oceans metagenomes shown in Supplementary Table 6.1-8.
Supplementary Figure 6.1-8: HMP BGC domain and class composition analysis. (a) Principal Coordinate Analysis (PCoA) performed on the dissimilarity matrix (Bray-Curtis) computed from the BGC class relative abundance profiles of 491 metagenomes of the HMP dataset (including 118 supragingival plaque, 128 tongue dorsum, 107 buccal mucosa, and 138 stool metagenomes). (b) PCoA performed on the dissimilarity matrix (Bray-Curtis) computed from the BGC domain relative abundance profiles of the same HMP metagenomes. Boxplots in the left and lower panels, show the distribution of the metagenomes from the different body sites in the PCo1 and PCo2 ordination axes, respectively. The metagenomes were grouped according to the body sites in the first two ordination axes in both PCoAs.
Supplementary Figure 6.1-9: BiG-MEx domain-based diversity analysis. Box plots comparing the distributions of the AMP-binding and condensation ODU diversity in the supragingival plaque, tongue dorsum, buccal mucosa, and stool body sites. For this analysis, we used a subset of 216 metagenomes that were obtained from the four body sites in the same individuals. The diversity of both domains was significantly higher in the supragingival plaque body site (pairwise Wilcoxon rank sum test; all p-values < 0.0001; see Supplementary Table 6.1-6).
Supplementary Figure 6.1-10: Evaluation of BiG-MEx BGC class abundance predictions in shallow metagenomes. Comparison of the BGC class abundance predictions between metagenomes of 100 and 5 million reads. We selected 30 TARA Oceans metagenomes, randomly subsampled these to 100 and 5 million reads, and used them as an input to predict the BGC class abundances with BiG-MEx BGC class abundance module. We selected the 13 BGC classes predicted as present in more than 10 of the metagenomes of 100 million reads.
6.1.2 Supplementary tables

Supplementary Table 6.1-1: Details of the Marine-RM, Marine-TM, Human-Oral, and Human-Stool simulated metagenomic datasets.

<table>
<thead>
<tr>
<th></th>
<th>Marine-RM</th>
<th>Marine-TM</th>
<th>Human-Oral</th>
<th>Human-Stool</th>
</tr>
</thead>
<tbody>
<tr>
<td># of genomes</td>
<td>378</td>
<td>344</td>
<td>209</td>
<td>479</td>
</tr>
<tr>
<td># of species</td>
<td>363</td>
<td>308</td>
<td>140</td>
<td>338</td>
</tr>
<tr>
<td># of genera</td>
<td>153</td>
<td>172</td>
<td>59</td>
<td>106</td>
</tr>
<tr>
<td># of metagenomes</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td># of PE reads</td>
<td>600,000,000</td>
<td>600,000,000</td>
<td>600,000,000</td>
<td>600,000,000</td>
</tr>
<tr>
<td># of HQ sequences</td>
<td>300,009,308</td>
<td>300,009,384</td>
<td>300,010,948</td>
<td>300,017,357</td>
</tr>
<tr>
<td># of predicted ORFs</td>
<td>288,385,700</td>
<td>290,575,789</td>
<td>287,083,275</td>
<td>286,809,051</td>
</tr>
<tr>
<td># of ORFs annotated with a BGC domain</td>
<td>5,068,930</td>
<td>5,607,322</td>
<td>3,407,877</td>
<td>3,249,576</td>
</tr>
</tbody>
</table>

Supplementary Table 6.1-2: Wilcoxon rank sum tests (two-sided) performed to evaluate the significance of the differentiation between TARA Oceans metagenomes of the different water layers according to their ordination in the first axis of the Principal Coordinate Analyses (PCoAs), and the Operational Domain Unit (ODU) diversity estimates. SRF, DCM, and MES correspond to surface, deep chlorophyll maximum and mesopelagic water layers, respectively.

<table>
<thead>
<tr>
<th>Compariso</th>
<th>Pairwise</th>
<th>N</th>
<th>P-value</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGC domain-based PCoA first axis</td>
<td>SRF vs. DCM</td>
<td>No 105 (63 + 42)</td>
<td>0.001911</td>
<td>1798</td>
</tr>
<tr>
<td>BGC domain-based PCoA first axis</td>
<td>SRF vs. MES</td>
<td>No 93 (63 + 30)</td>
<td>8.889e-15</td>
<td>1889</td>
</tr>
<tr>
<td>BGC domain-based PCoA first axis</td>
<td>DCM vs. MES</td>
<td>No 72 (42 + 30)</td>
<td>&lt; 2.2e-16</td>
<td>7</td>
</tr>
<tr>
<td>BGC class-based PCoA first axis</td>
<td>SRF vs. DCM</td>
<td>No 105 (63 + 42)</td>
<td>0.004043</td>
<td>1763</td>
</tr>
<tr>
<td>BGC class-based PCoA first axis</td>
<td>SRF vs. MES</td>
<td>No 93 (63 + 30)</td>
<td>1.538e-14</td>
<td>1880</td>
</tr>
<tr>
<td>BGC class-based PCoA first axis</td>
<td>DCM vs. MES</td>
<td>No 72 (42 + 30)</td>
<td>&lt; 2.2e-16</td>
<td>23</td>
</tr>
<tr>
<td>AMP-binding ODU alpha diversity</td>
<td>SRF vs. DCM</td>
<td>Yes 42 (21 + 21)</td>
<td>6.387e-05</td>
<td>0</td>
</tr>
<tr>
<td>AMP-binding ODU alpha diversity</td>
<td>SRF vs. MES</td>
<td>Yes 42 (21 + 21)</td>
<td>9.537e-07</td>
<td>0</td>
</tr>
<tr>
<td>AMP-binding ODU alpha diversity</td>
<td>DCM vs. MES</td>
<td>Yes 42 (21 + 21)</td>
<td>0.0002928</td>
<td>19</td>
</tr>
<tr>
<td>Condensation ODU alpha diversity</td>
<td>SRF vs. DCM</td>
<td>Yes 42 (21 + 21)</td>
<td>0.001176</td>
<td>27</td>
</tr>
<tr>
<td>Condensation ODU alpha diversity</td>
<td>SRF vs. MES</td>
<td>Yes 42 (21 + 21)</td>
<td>1.907e-06</td>
<td>1</td>
</tr>
<tr>
<td>Condensation ODU alpha diversity</td>
<td>DCM vs. MES</td>
<td>Yes 42 (21 + 21)</td>
<td>0.0003538</td>
<td>20</td>
</tr>
</tbody>
</table>
**Supplementary Table 6.1-3:** Permutational Multivariate Analysis of Variance (PERMANOVA) to test if the BGC class composition predicted in TARA Oceans metagenomes differs between water layers. The PERMANOVA was performed on the Bray-Curtis dissimilarity matrix obtained from the BGC class relative abundance profiles, for a balanced subset of 63 metagenomes (i.e., 21 from each water layer). P-values were adjusted with the Bonferroni correction for multiple comparisons. SRF, DCM, and MES correspond to surface, deep chlorophyll maximum and mesopelagic water layers, respectively.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>P-value</th>
<th>P-value adjusted</th>
<th>Pseudo-F statistic</th>
<th>R²</th>
<th># of perm.</th>
<th>N</th>
<th>Degrees of freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>All water layers</td>
<td>0.0001</td>
<td>NA</td>
<td>79.68</td>
<td>0.726</td>
<td>999</td>
<td>63</td>
<td>2</td>
</tr>
<tr>
<td>SRF vs. DCM</td>
<td>0.007</td>
<td>0.021</td>
<td>3.69</td>
<td>0.084</td>
<td>999</td>
<td>42</td>
<td>1</td>
</tr>
<tr>
<td>SRF vs. MES</td>
<td>0.001</td>
<td>0.003</td>
<td>153.12</td>
<td>0.793</td>
<td>999</td>
<td>42</td>
<td>1</td>
</tr>
<tr>
<td>DCM vs. MES</td>
<td>0.001</td>
<td>0.003</td>
<td>78.06</td>
<td>0.661</td>
<td>999</td>
<td>42</td>
<td>1</td>
</tr>
</tbody>
</table>
**Supplementary Table 6.1-4:** Student's t-tests performed to evaluate the significance of the correlations described in TARA Oceans dataset analyses. The taxonomic and functional richness and diversity were computed by Sunagawa et al. (2015). ODU: Operational Domain Units.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Pearson's r</th>
<th>P-value</th>
<th>Statistic</th>
<th>N</th>
<th>Degrees of freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGC domain-based PCoA first axis vs. temperature</td>
<td>-0.73</td>
<td>&lt; 2.2e-16</td>
<td>-12.289</td>
<td>135</td>
<td>133</td>
</tr>
<tr>
<td>BGC class-based PCoA first axis vs. temperature</td>
<td>-0.73</td>
<td>&lt; 2.2e-16</td>
<td>-12.221</td>
<td>135</td>
<td>133</td>
</tr>
<tr>
<td>AMP-binding ODU alpha diversity vs. taxonomic richness</td>
<td>0.80</td>
<td>1.155e-14</td>
<td>10.68</td>
<td>58</td>
<td>60</td>
</tr>
<tr>
<td>Condensation ODU alpha diversity vs. taxonomic richness</td>
<td>0.79</td>
<td>3.802e-14</td>
<td>9.945</td>
<td>58</td>
<td>60</td>
</tr>
<tr>
<td>AMP-binding ODU alpha diversity vs. taxonomic diversity</td>
<td>0.66</td>
<td>8.688e-09</td>
<td>6.719</td>
<td>58</td>
<td>60</td>
</tr>
<tr>
<td>Condensation ODU alpha diversity vs. taxonomic diversity</td>
<td>0.49</td>
<td>7.008e-05</td>
<td>4.284</td>
<td>58</td>
<td>60</td>
</tr>
<tr>
<td>AMP-binding ODU alpha diversity vs. functional richness</td>
<td>0.87</td>
<td>1.979e-19</td>
<td>13.416</td>
<td>58</td>
<td>60</td>
</tr>
<tr>
<td>Condensation ODU alpha diversity vs. functional richness</td>
<td>0.76</td>
<td>2.508e-12</td>
<td>8.833</td>
<td>58</td>
<td>60</td>
</tr>
<tr>
<td>AMP-binding ODU alpha diversity vs. functional diversity</td>
<td>0.54</td>
<td>6.747e-06</td>
<td>4.949</td>
<td>58</td>
<td>60</td>
</tr>
<tr>
<td>Condensation ODU alpha diversity vs. functional diversity</td>
<td>0.59</td>
<td>7.601e-07</td>
<td>5.544</td>
<td>58</td>
<td>60</td>
</tr>
</tbody>
</table>
Supplementary Table 6.1-5: PERMANOVA to test if the predicted BGC class composition in the HMP metagenomes differs between body sites. The PERMANOVA was performed on the Bray-Curtis dissimilarity matrices obtained from the relative abundance of BGC classes, for a balanced subset of 216 metagenomes (i.e., 54 from each body site). P-values were adjusted with the Bonferroni correction for multiple comparisons.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>P-value</th>
<th>P-value adjusted</th>
<th>Pseudo-F statistic</th>
<th>R²</th>
<th># of perm</th>
<th>N</th>
<th>Degrees of freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>All body sites</td>
<td>0.001</td>
<td>NA</td>
<td>140.29</td>
<td>0.665</td>
<td>999</td>
<td>216</td>
<td>3</td>
</tr>
<tr>
<td>Supragingival plaque vs. buccal mucosa</td>
<td>0.001</td>
<td>0.006</td>
<td>146.64</td>
<td>0.58</td>
<td>999</td>
<td>108</td>
<td>1</td>
</tr>
<tr>
<td>Supragingival plaque vs. tongue dorsum</td>
<td>0.001</td>
<td>0.006</td>
<td>57.31</td>
<td>0.35</td>
<td>999</td>
<td>108</td>
<td>1</td>
</tr>
<tr>
<td>Supragingival plaque vs. stool</td>
<td>0.001</td>
<td>0.006</td>
<td>324.9</td>
<td>0.75</td>
<td>999</td>
<td>108</td>
<td>1</td>
</tr>
<tr>
<td>Buccal mucosa vs. tongue dorsum</td>
<td>0.001</td>
<td>0.006</td>
<td>45.24</td>
<td>0.3</td>
<td>999</td>
<td>108</td>
<td>1</td>
</tr>
<tr>
<td>Buccal mucosa vs. stool</td>
<td>0.001</td>
<td>0.006</td>
<td>134.65</td>
<td>0.56</td>
<td>999</td>
<td>108</td>
<td>1</td>
</tr>
<tr>
<td>Tongue dorsum vs. stool</td>
<td>0.001</td>
<td>0.006</td>
<td>192.15</td>
<td>0.64</td>
<td>999</td>
<td>108</td>
<td>1</td>
</tr>
</tbody>
</table>
Chapter 6. Appendix: supplementary materials

**Supplementary Table 6.1-6**: Wilcoxon rank sum tests (two-sided) performed to evaluate the significance of the differentiation between the supragingival plaque, and the buccal mucosa, tongue dorsum and stool body sites, according to the Operational Domain Unit (ODU) diversity.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Pairwise</th>
<th>N</th>
<th>P-value</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMP-binding ODU alpha diversity</td>
<td>Supragingival plaque vs. buccal mucosa</td>
<td>Yes</td>
<td>108 (54 + 54)</td>
<td>1.769e-10</td>
</tr>
<tr>
<td>AMP-binding ODU alpha diversity</td>
<td>Supragingival plaque vs. tongue dorsum</td>
<td>Yes</td>
<td>108 (54 + 54)</td>
<td>4.649e-09</td>
</tr>
<tr>
<td>AMP-binding ODU alpha diversity</td>
<td>Supragingival plaque vs. stool</td>
<td>Yes</td>
<td>108 (54 + 54)</td>
<td>4.81e-10</td>
</tr>
<tr>
<td>Condensation ODU alpha diversity</td>
<td>Supragingival plaque vs. buccal mucosa</td>
<td>Yes</td>
<td>108 (54 + 54)</td>
<td>1.672e-10</td>
</tr>
<tr>
<td>Condensation ODU alpha diversity</td>
<td>Supragingival plaque vs. tongue dorsum</td>
<td>Yes</td>
<td>108 (54 + 54)</td>
<td>1.871e-10</td>
</tr>
<tr>
<td>Condensation ODU alpha diversity</td>
<td>Supragingival plaque vs. stool</td>
<td>Yes</td>
<td>108 (54 + 54)</td>
<td>2.767e-10</td>
</tr>
</tbody>
</table>

**Supplementary Table 6.1-7**: Details of the TARA Oceans and HMP analyzed metagenomes.

<table>
<thead>
<tr>
<th></th>
<th>TARA Oceans</th>
<th>HMP</th>
</tr>
</thead>
<tbody>
<tr>
<td># of metagenomes</td>
<td>139</td>
<td>491</td>
</tr>
<tr>
<td># of reads</td>
<td>51,220,320,380</td>
<td>33,893,000,957</td>
</tr>
<tr>
<td># of HQ sequences</td>
<td>25,458,208,431</td>
<td>26,997,667,248</td>
</tr>
<tr>
<td># of predicted ORFs</td>
<td>23,281,706,611</td>
<td>23,184,977,824</td>
</tr>
<tr>
<td># of ORFs annotated with a BGC domain</td>
<td>260,490,101</td>
<td>235,798,413</td>
</tr>
</tbody>
</table>
Chapter 6. Appendix: supplementary materials

Supplementary Table 6.1-8: Details of the nine TARA Oceans metagenomes used in the evaluation of BiG-MEx BGC domain identification module. SRF, DCM, and MES abbreviations correspond to surface, deep chlorophyll maximum and mesopelagic water layers, respectively.

<table>
<thead>
<tr>
<th>Sample name</th>
<th>Water layer</th>
<th># of PE reads</th>
<th># of HQ sequences</th>
<th># of predicted ORFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>TARA_004_SRF_0.22-1</td>
<td>SRF</td>
<td>404,962,514</td>
<td>219,479,553</td>
<td>203,123,857</td>
</tr>
<tr>
<td>TARA_007_SRF_0.22-1</td>
<td>SRF</td>
<td>221,166,612</td>
<td>138,645,938</td>
<td>119,212,194</td>
</tr>
<tr>
<td>TARA_009_SRF_0.22-1</td>
<td>SRF</td>
<td>489,617,426</td>
<td>232,980,853</td>
<td>210,248,922</td>
</tr>
<tr>
<td>TARA_004_DCM_0.22-1</td>
<td>DCM</td>
<td>476,921,334</td>
<td>220,441,035</td>
<td>202,942,150</td>
</tr>
<tr>
<td>TARA_007_DCM_0.22-1</td>
<td>DCM</td>
<td>178,519,830</td>
<td>107,088,525</td>
<td>96,658,295</td>
</tr>
<tr>
<td>TARA_009_DCM_0.22-1</td>
<td>DCM</td>
<td>416,553,274</td>
<td>217,526,916</td>
<td>202,183,924</td>
</tr>
<tr>
<td>TARA_037_MES_0.22-1</td>
<td>MES</td>
<td>399,390,808</td>
<td>193,596,707</td>
<td>179,380,880</td>
</tr>
<tr>
<td>TARA_038_MES_0.22-1</td>
<td>MES</td>
<td>260,588,896</td>
<td>139,878,638</td>
<td>126,264,027</td>
</tr>
<tr>
<td>TARA_039_MES_0.22-1</td>
<td>MES</td>
<td>277,574,946</td>
<td>137,114,641</td>
<td>124,762,241</td>
</tr>
</tbody>
</table>

Supplementary Table 6.1-9: Details of the assembled metagenomes of the Marine-TM dataset.

| # of assembled metagenomes | 50             |
| # of PE reads              | 200,000,000    |
| # of contigs               | 4,479,150      |
| Average N50                | 755.52         |
| # of contigs annotated with a BGC domain | 749,575 |
| # of contigs annotated with a BGC class | 465 (0.06%)* |

*Percentage of contigs annotated with a BGC class taking the contigs annotated with a BGC domain as the total.
### Supplementary Table 6.1-10: Statistics of the BGC domain reference phylogenetic reconstructions.

<table>
<thead>
<tr>
<th>BGC domain</th>
<th># of seqs in MIBiG</th>
<th># of seqs in phylo</th>
<th>Clust threshold</th>
<th>Subst model</th>
<th>Coph dist mean</th>
<th>Coph dist sd</th>
<th>Coph dist max</th>
<th>Coph dist min</th>
<th>Phylo Div</th>
<th>BGC class</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMP-binding</td>
<td>2265</td>
<td>1355</td>
<td>0.7</td>
<td>LG</td>
<td>3.228</td>
<td>0.981</td>
<td>7.057</td>
<td>0.269</td>
<td>2982.379</td>
<td>nrps,other</td>
</tr>
<tr>
<td>ATd</td>
<td>538</td>
<td>445</td>
<td>0.7</td>
<td>LG</td>
<td>2.027</td>
<td>0.505</td>
<td>4.111</td>
<td>0.134</td>
<td>2732.548</td>
<td>transatpks</td>
</tr>
<tr>
<td>BLS</td>
<td>58</td>
<td>49</td>
<td>0.9</td>
<td>LG</td>
<td>4.932</td>
<td>0.907</td>
<td>8.412</td>
<td>0.629</td>
<td>1554.448</td>
<td>blactam</td>
</tr>
<tr>
<td>CaiA</td>
<td>83</td>
<td>68</td>
<td>0.9</td>
<td>LG</td>
<td>4.932</td>
<td>0.907</td>
<td>8.412</td>
<td>0.629</td>
<td>1554.448</td>
<td>other</td>
</tr>
<tr>
<td>Condensation</td>
<td>2180</td>
<td>1381</td>
<td>0.9</td>
<td>LG</td>
<td>3.116</td>
<td>0.883</td>
<td>6.425</td>
<td>0.271</td>
<td>3107.288</td>
<td>nrps</td>
</tr>
<tr>
<td>DOIS</td>
<td>45</td>
<td>30</td>
<td>0.9</td>
<td>LG</td>
<td>3.116</td>
<td>0.883</td>
<td>6.425</td>
<td>0.271</td>
<td>3107.288</td>
<td>nrps</td>
</tr>
<tr>
<td>DUF1205</td>
<td>167</td>
<td>133</td>
<td>0.9</td>
<td>LG</td>
<td>3.116</td>
<td>0.883</td>
<td>6.425</td>
<td>0.271</td>
<td>3107.288</td>
<td>nrps</td>
</tr>
<tr>
<td>Glyco_transf_1</td>
<td>108</td>
<td>85</td>
<td>0.9</td>
<td>LG</td>
<td>4.932</td>
<td>0.907</td>
<td>8.412</td>
<td>0.629</td>
<td>1554.448</td>
<td>blactam</td>
</tr>
<tr>
<td>Glyco_transf_2</td>
<td>157</td>
<td>121</td>
<td>0.9</td>
<td>LG</td>
<td>4.932</td>
<td>0.907</td>
<td>8.412</td>
<td>0.629</td>
<td>1554.448</td>
<td>other</td>
</tr>
<tr>
<td>LANc_like</td>
<td>76</td>
<td>68</td>
<td>0.9</td>
<td>LG</td>
<td>4.932</td>
<td>0.907</td>
<td>8.412</td>
<td>0.629</td>
<td>1554.448</td>
<td>nrps</td>
</tr>
<tr>
<td>Lant_dehyd_C</td>
<td>52</td>
<td>46</td>
<td>0.9</td>
<td>LG</td>
<td>4.932</td>
<td>0.907</td>
<td>8.412</td>
<td>0.629</td>
<td>1554.448</td>
<td>transatpks</td>
</tr>
<tr>
<td>Lant_dehyd_N</td>
<td>26</td>
<td>22</td>
<td>0.9</td>
<td>LG</td>
<td>4.932</td>
<td>0.907</td>
<td>8.412</td>
<td>0.629</td>
<td>1554.448</td>
<td>transatpks</td>
</tr>
<tr>
<td>LipV</td>
<td>46</td>
<td>30</td>
<td>0.9</td>
<td>LG</td>
<td>3.116</td>
<td>0.883</td>
<td>6.425</td>
<td>0.271</td>
<td>3107.288</td>
<td>nrps</td>
</tr>
<tr>
<td>MGT</td>
<td>58</td>
<td>49</td>
<td>0.9</td>
<td>LG</td>
<td>4.932</td>
<td>0.907</td>
<td>8.412</td>
<td>0.629</td>
<td>1554.448</td>
<td>blactam</td>
</tr>
<tr>
<td>MGT2</td>
<td>58</td>
<td>49</td>
<td>0.9</td>
<td>LG</td>
<td>4.932</td>
<td>0.907</td>
<td>8.412</td>
<td>0.629</td>
<td>1554.448</td>
<td>blactam</td>
</tr>
<tr>
<td>NAD_binding_4</td>
<td>189</td>
<td>155</td>
<td>0.9</td>
<td>LG</td>
<td>4.932</td>
<td>0.907</td>
<td>8.412</td>
<td>0.629</td>
<td>1554.448</td>
<td>blactam</td>
</tr>
<tr>
<td>PKS_AT</td>
<td>38</td>
<td>33</td>
<td>0.9</td>
<td>LG</td>
<td>4.932</td>
<td>0.907</td>
<td>8.412</td>
<td>0.629</td>
<td>1554.448</td>
<td>blactam</td>
</tr>
<tr>
<td>PKS_AT2</td>
<td>32</td>
<td>26</td>
<td>0.9</td>
<td>LG</td>
<td>3.116</td>
<td>0.883</td>
<td>6.425</td>
<td>0.271</td>
<td>3107.288</td>
<td>nrps</td>
</tr>
<tr>
<td>TIGR03604</td>
<td>46</td>
<td>37</td>
<td>0.9</td>
<td>LG</td>
<td>3.116</td>
<td>0.883</td>
<td>6.425</td>
<td>0.271</td>
<td>3107.288</td>
<td>nrps</td>
</tr>
<tr>
<td>TIGR03604</td>
<td>46</td>
<td>37</td>
<td>0.9</td>
<td>LG</td>
<td>3.116</td>
<td>0.883</td>
<td>6.425</td>
<td>0.271</td>
<td>3107.288</td>
<td>nrps</td>
</tr>
<tr>
<td>TIGR03975</td>
<td>44</td>
<td>33</td>
<td>0.9</td>
<td>LG</td>
<td>3.116</td>
<td>0.883</td>
<td>6.425</td>
<td>0.271</td>
<td>3107.288</td>
<td>nrps</td>
</tr>
<tr>
<td>TIGR03604</td>
<td>46</td>
<td>37</td>
<td>0.9</td>
<td>LG</td>
<td>3.116</td>
<td>0.883</td>
<td>6.425</td>
<td>0.271</td>
<td>3107.288</td>
<td>nrps</td>
</tr>
<tr>
<td>TIGR03604</td>
<td>46</td>
<td>37</td>
<td>0.9</td>
<td>LG</td>
<td>3.116</td>
<td>0.883</td>
<td>6.425</td>
<td>0.271</td>
<td>3107.288</td>
<td>nrps</td>
</tr>
<tr>
<td>TIGR03604</td>
<td>46</td>
<td>37</td>
<td>0.9</td>
<td>LG</td>
<td>3.116</td>
<td>0.883</td>
<td>6.425</td>
<td>0.271</td>
<td>3107.288</td>
<td>nrps</td>
</tr>
</tbody>
</table>

Chapter 6. Appendix: supplementary materials
6.2 Supplementary material of the work “Mg-Traits pipeline and database: advancing functional trait-based approaches in metagenomics”

6.2.1 Supplementary tables
Supplementary Table 6.2-1: Mg-Traits correlation matrix. Lower triangular of the Pearson’s r correlation matrix of the 15 mg-traits.

<table>
<thead>
<tr>
<th></th>
<th>ACN</th>
<th>AA PC1</th>
<th>AB ratio</th>
<th>AGS</th>
<th>BGC PC1</th>
<th>Codon PC1</th>
<th>Dinuc PC1</th>
<th>Fun PC1</th>
<th>Fun div</th>
<th>GC cont</th>
<th>GC var</th>
<th>Perc</th>
<th>Perc fun annot</th>
<th>Taxa PC1</th>
<th>Taxa div</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA PC1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB ratio</td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGS</td>
<td>0.35</td>
<td>0.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BGC PC1</td>
<td>0.31</td>
<td>0.79</td>
<td>0.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Codon PC1</td>
<td>-0.42</td>
<td>-0.74</td>
<td>-0.46</td>
<td>-0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dinuc PC1</td>
<td>0.57</td>
<td>0.98</td>
<td>0.3</td>
<td>0.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fun PC1</td>
<td>0.59</td>
<td>0.99</td>
<td>0.27</td>
<td>0.78</td>
<td>-0.73</td>
<td>0.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fun div</td>
<td>0.44</td>
<td>0.72</td>
<td>0.38</td>
<td>0.45</td>
<td>-0.98</td>
<td>0.77</td>
<td>0.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GC cont</td>
<td>0.39</td>
<td>0.48</td>
<td>-0.02</td>
<td>0.46</td>
<td>-0.1</td>
<td>0.46</td>
<td>0.49</td>
<td>0.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GC var</td>
<td>0.58</td>
<td>0.97</td>
<td>0.23</td>
<td>0.77</td>
<td>-0.69</td>
<td>0.98</td>
<td>0.99</td>
<td>0.68</td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perc TF</td>
<td>0.23</td>
<td>0.59</td>
<td>-0.28</td>
<td>0.43</td>
<td>-0.19</td>
<td>0.61</td>
<td>0.66</td>
<td>0.21</td>
<td>0.58</td>
<td>0.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perc fun annot</td>
<td>0.54</td>
<td>0.47</td>
<td>0.14</td>
<td>0.28</td>
<td>-0.64</td>
<td>0.49</td>
<td>0.46</td>
<td>0.71</td>
<td>0.35</td>
<td>0.43</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taxa PC1</td>
<td>0.016</td>
<td>-0.03</td>
<td>0.17</td>
<td>-0.29</td>
<td>-0.29</td>
<td>0.04</td>
<td>-0.02</td>
<td>0.35</td>
<td>-0.03</td>
<td>-0.07</td>
<td>-0.23</td>
<td>0.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taxa div</td>
<td>0.23</td>
<td>0.67</td>
<td>0.34</td>
<td>0.63</td>
<td>-0.72</td>
<td>0.7</td>
<td>0.66</td>
<td>0.7</td>
<td>0.31</td>
<td>0.64</td>
<td>0.29</td>
<td>0.48</td>
<td>0.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.21</td>
<td>0.74</td>
<td>0.33</td>
<td>0.67</td>
<td>-0.72</td>
<td>0.76</td>
<td>0.74</td>
<td>0.7</td>
<td>0.49</td>
<td>0.73</td>
<td>0.48</td>
<td>0.38</td>
<td>-0.1</td>
<td>0.7</td>
<td></td>
</tr>
</tbody>
</table>
**Supplementary Table 6.2-2:** Paired Wilcoxon rank sum tests. Details of the paired Wilcoxon rank sum tests comparing the mg-trait value distributions between surface and deep chlorophyll maximum, and mesopelagic water layers (SRF, DCM and MES respectively). To perform the tests we used the mg-trait profiles of the 63 matching metagenomes from the Layers-Subset (i.e., 21 metagenomes from each water layer). P-values lower than 0.05 are shown in boldface.

<table>
<thead>
<tr>
<th>Mg_Trait</th>
<th>N</th>
<th>Statistic (SRF vs. DCM)</th>
<th>P-value</th>
<th>Statistic (SRF vs. MES)</th>
<th>P-value</th>
<th>Statistic (DCM vs. MES)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC content</td>
<td>42</td>
<td>34</td>
<td>3.28e-03</td>
<td>5</td>
<td>9.54e-06</td>
<td>25</td>
<td>8.52e-04</td>
</tr>
<tr>
<td>GC variance</td>
<td>42</td>
<td>90</td>
<td>3.93e-01</td>
<td>106</td>
<td>7.59e-01</td>
<td>138</td>
<td>4.52e-01</td>
</tr>
<tr>
<td>Dinuc PC1</td>
<td>42</td>
<td>25</td>
<td>8.52e-04</td>
<td>2</td>
<td>2.86e-06</td>
<td>13</td>
<td>8.39e-05</td>
</tr>
<tr>
<td>Codon PC1</td>
<td>42</td>
<td>26</td>
<td>1.00e-03</td>
<td>1</td>
<td>1.91e-06</td>
<td>6</td>
<td>1.34e-05</td>
</tr>
<tr>
<td>AA PC1</td>
<td>42</td>
<td>15</td>
<td>1.31e-04</td>
<td>1</td>
<td>1.91e-06</td>
<td>14</td>
<td>1.05e-04</td>
</tr>
<tr>
<td>AB ratio</td>
<td>42</td>
<td>3</td>
<td>4.77e-06</td>
<td>1</td>
<td>1.91e-06</td>
<td>31</td>
<td>2.15e-03</td>
</tr>
<tr>
<td>Fun PC1</td>
<td>42</td>
<td>7</td>
<td>1.81e-05</td>
<td>0</td>
<td>9.54e-07</td>
<td>0</td>
<td>9.54e-07</td>
</tr>
<tr>
<td>Fun div</td>
<td>42</td>
<td>47</td>
<td>1.56e-02</td>
<td>58</td>
<td>4.60e-02</td>
<td>125</td>
<td>7.59e-01</td>
</tr>
<tr>
<td>Perc fun annot</td>
<td>42</td>
<td>150</td>
<td>2.43e-01</td>
<td>36</td>
<td>4.28e-03</td>
<td>17</td>
<td>1.97e-04</td>
</tr>
<tr>
<td>Perc TF</td>
<td>42</td>
<td>74</td>
<td>1.57e-01</td>
<td>6</td>
<td>1.34e-05</td>
<td>16</td>
<td>1.61e-04</td>
</tr>
<tr>
<td>BGC PC1</td>
<td>42</td>
<td>226</td>
<td>9.54e-06</td>
<td>231</td>
<td>9.54e-07</td>
<td>231</td>
<td>9.54e-07</td>
</tr>
<tr>
<td>AGS</td>
<td>42</td>
<td>22</td>
<td>5.10e-04</td>
<td>18</td>
<td>2.41e-04</td>
<td>32</td>
<td>2.48e-03</td>
</tr>
<tr>
<td>Taxa PC1</td>
<td>42</td>
<td>69</td>
<td>1.11e-01</td>
<td>0</td>
<td>9.54e-07</td>
<td>0</td>
<td>9.54e-07</td>
</tr>
<tr>
<td>Taxa div</td>
<td>42</td>
<td>17</td>
<td>1.97e-04</td>
<td>0</td>
<td>9.54e-07</td>
<td>7</td>
<td>1.81e-05</td>
</tr>
<tr>
<td>ACN</td>
<td>42</td>
<td>85</td>
<td>3.04e-01</td>
<td>1</td>
<td>1.91e-06</td>
<td>46</td>
<td>1.42e-02</td>
</tr>
</tbody>
</table>
Supplementary Table 6.2-3: Permutational Multivariate Analysis of Variance (PERMANOVA): differentiation of water layers based on the functional and taxonomic compositions. PERMANOVA tests comparing the Surface, Deep Chlorophyll Maximum and Mesopelagic water layers (SRF, DCM, and MES, respectively) based on the functional and taxonomic compositions. For these analyses, we used the 63 metagenomes of the Layers-Subset. P-values lower than 0.05 are shown in boldface.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Test</th>
<th>Sample size</th>
<th>Statistic</th>
<th>R²</th>
<th>p-value</th>
<th>p-value adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Functional</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>composition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>vs. Water layers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>All water layers</td>
<td>63</td>
<td>30.764</td>
<td>0.506</td>
<td>1.00e-03</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>SRF vs. DCM</td>
<td>42</td>
<td>3.708</td>
<td>0.085</td>
<td>1.00e-03</td>
<td>3.00e-03</td>
</tr>
<tr>
<td></td>
<td>SRF vs. MES</td>
<td>42</td>
<td>62.124</td>
<td>0.608</td>
<td>1.00e-03</td>
<td>3.00e-03</td>
</tr>
<tr>
<td></td>
<td>DCM vs. MES</td>
<td>42</td>
<td>3.708</td>
<td>0.085</td>
<td>1.00e-03</td>
<td>3.00e-03</td>
</tr>
<tr>
<td></td>
<td>Taxa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>composition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>vs. Water layers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>All water layers</td>
<td>63</td>
<td>20.429</td>
<td>0.405</td>
<td>1.00e-03</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>SRF vs. DCM</td>
<td>42</td>
<td>3.755</td>
<td>0.086</td>
<td>2.00e-03</td>
<td>6.00e-03</td>
</tr>
<tr>
<td></td>
<td>SRF vs. MES</td>
<td>42</td>
<td>39.487</td>
<td>0.497</td>
<td>1.00e-03</td>
<td>3.00e-03</td>
</tr>
<tr>
<td></td>
<td>DCM vs. MES</td>
<td>42</td>
<td>20.313</td>
<td>0.337</td>
<td>1.00e-03</td>
<td>3.00e-03</td>
</tr>
</tbody>
</table>
### Supplementary Table 6.2-4: Permutational Multivariate Analysis of Variance (PERMANOVA): differentiation of ocean regions based on the functional composition.

The PERMANOVA tests comparing the six ocean regions were performed on the functional composition of the 57 surface metagenomes of the Regions-Subset. The ocean regions acronyms IO, MS, NAO, NPO, SAO, and SPO, stand for Indian Ocean, Mediterranean Sea, North Atlantic Ocean, North Pacific Ocean, South Atlantic Ocean, and South Pacific Ocean, respectively. P-values lower than 0.05 are shown in boldface.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Test</th>
<th>Sample size</th>
<th>Statistic</th>
<th>$R^2$</th>
<th>p-value</th>
<th>p-value adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>All ocean regions</td>
<td>NAO vs. MS</td>
<td>16 (10+6)</td>
<td>5.841</td>
<td>0.294</td>
<td>3.00e-03</td>
<td>4.50e-02</td>
</tr>
<tr>
<td></td>
<td>NAO vs. IO</td>
<td>22 (10+12)</td>
<td>8.044</td>
<td>0.287</td>
<td>1.00e-03</td>
<td>1.50e-02</td>
</tr>
<tr>
<td></td>
<td>NAO vs. SAO</td>
<td>18 (10+8)</td>
<td>2.107</td>
<td>0.116</td>
<td>3.70e-02</td>
<td>5.55e-01</td>
</tr>
<tr>
<td></td>
<td>NAO vs. SPO</td>
<td>25 (10+15)</td>
<td>4.323</td>
<td>0.158</td>
<td>2.00e-03</td>
<td>3.00e-02</td>
</tr>
<tr>
<td></td>
<td>NAO vs. NPO</td>
<td>16 (10+6)</td>
<td>3.93</td>
<td>0.219</td>
<td>1.30e-02</td>
<td>1.95e-01</td>
</tr>
<tr>
<td></td>
<td>MS vs. IO</td>
<td>18 (6+12)</td>
<td>7.917</td>
<td>0.331</td>
<td>1.00e-03</td>
<td>1.50e-02</td>
</tr>
<tr>
<td></td>
<td>MS vs. SAO</td>
<td>14 (6+8)</td>
<td>4.988</td>
<td>0.294</td>
<td>2.00e-03</td>
<td>3.00e-02</td>
</tr>
<tr>
<td></td>
<td>MS vs. SPO</td>
<td>21 (6+15)</td>
<td>3.451</td>
<td>0.154</td>
<td>1.00e-03</td>
<td>1.50e-02</td>
</tr>
<tr>
<td></td>
<td>MS vs. NPO</td>
<td>12 (6+6)</td>
<td>3.773</td>
<td>0.274</td>
<td>1.00e-03</td>
<td>1.50e-02</td>
</tr>
<tr>
<td></td>
<td>IO vs. SAO</td>
<td>20 (12+8)</td>
<td>6.601</td>
<td>0.268</td>
<td>1.00e-03</td>
<td>1.50e-02</td>
</tr>
<tr>
<td></td>
<td>IO vs. SPO</td>
<td>27 (12+15)</td>
<td>6.618</td>
<td>0.209</td>
<td>1.00e-03</td>
<td>1.50e-02</td>
</tr>
<tr>
<td></td>
<td>IO vs. NPO</td>
<td>18 (12+6)</td>
<td>1.973</td>
<td>0.11</td>
<td>5.30e-02</td>
<td>7.95e-01</td>
</tr>
<tr>
<td></td>
<td>SAO vs. SPO</td>
<td>23 (8+15)</td>
<td>3.975</td>
<td>0.159</td>
<td>1.00e-03</td>
<td>1.50e-02</td>
</tr>
<tr>
<td></td>
<td>SAO vs. NPO</td>
<td>14 (8+6)</td>
<td>2.949</td>
<td>0.197</td>
<td>9.00e-03</td>
<td>1.35e-01</td>
</tr>
<tr>
<td></td>
<td>SPO vs. NPO</td>
<td>21 (15+6)</td>
<td>1.186</td>
<td>0.059</td>
<td>2.76e-01</td>
<td>1.00e+00</td>
</tr>
</tbody>
</table>
Supplementary Table 6.2-5: Permutational Multivariate Analysis of Variance (PERMANOVA): differentiation of ocean regions based on the taxonomic composition. The PERMANOVA tests comparing the six ocean regions were performed on the taxonomic composition of the 57 surface metagenomes of the Regions-Subset. The ocean regions acronyms IO, MS, NAO, NPO, SAO, and SPO, stand for Indian Ocean, Mediterranean Sea, North Atlantic Ocean, North Pacific Ocean, South Atlantic Ocean, and South Pacific Ocean, respectively. P-values lower than 0.05 are shown in boldface.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Test</th>
<th>Sample size</th>
<th>Statistic</th>
<th>R²</th>
<th>p-value</th>
<th>p-value adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>All ocean regions</td>
<td>NAO vs MS</td>
<td>16 (10+6)</td>
<td>6.987</td>
<td>0.333</td>
<td>1.00e-03</td>
<td>1.50e-02</td>
</tr>
<tr>
<td></td>
<td>NAO vs IO</td>
<td>22 (10+12)</td>
<td>6.136</td>
<td>0.235</td>
<td>1.00e-03</td>
<td>1.50e-02</td>
</tr>
<tr>
<td></td>
<td>NAO vs SAO</td>
<td>18 (10+8)</td>
<td>3.884</td>
<td>0.195</td>
<td>1.00e-03</td>
<td>1.50e-02</td>
</tr>
<tr>
<td></td>
<td>NAO vs SPO</td>
<td>25 (10+15)</td>
<td>4.799</td>
<td>0.173</td>
<td>3.00e-03</td>
<td>4.50e-02</td>
</tr>
<tr>
<td></td>
<td>NAO vs NPO</td>
<td>16 (10+6)</td>
<td>4.043</td>
<td>0.224</td>
<td>1.00e-02</td>
<td>1.50e-01</td>
</tr>
<tr>
<td></td>
<td>MS vs IO</td>
<td>18 (6+12)</td>
<td>6.452</td>
<td>0.287</td>
<td>2.00e-03</td>
<td>3.00e-02</td>
</tr>
<tr>
<td>Taxonomic composition</td>
<td>MS vs SAO</td>
<td>14 (6+8)</td>
<td>5.565</td>
<td>0.317</td>
<td>1.00e-03</td>
<td>1.50e-02</td>
</tr>
<tr>
<td>vs. Ocean regions</td>
<td>MS vs SPO</td>
<td>21 (6+15)</td>
<td>3.584</td>
<td>0.159</td>
<td>2.00e-03</td>
<td>3.00e-02</td>
</tr>
<tr>
<td></td>
<td>MS vs NPO</td>
<td>12 (6+6)</td>
<td>3.98</td>
<td>0.285</td>
<td>7.00e-03</td>
<td>1.05e-01</td>
</tr>
<tr>
<td></td>
<td>IO vs SAO</td>
<td>20 (12+8)</td>
<td>7.102</td>
<td>0.283</td>
<td>1.00e-03</td>
<td>1.50e-02</td>
</tr>
<tr>
<td></td>
<td>IO vs SPO</td>
<td>27 (12+15)</td>
<td>4.439</td>
<td>0.151</td>
<td>1.00e-03</td>
<td>1.50e-02</td>
</tr>
<tr>
<td></td>
<td>IO vs NPO</td>
<td>18 (12+6)</td>
<td>2.31</td>
<td>0.126</td>
<td>1.00e-02</td>
<td>1.50e-01</td>
</tr>
<tr>
<td></td>
<td>SAO vs SPO</td>
<td>23 (8+15)</td>
<td>5.092</td>
<td>0.195</td>
<td>1.00e-03</td>
<td>1.50e-02</td>
</tr>
<tr>
<td></td>
<td>SAO vs NPO</td>
<td>14 (8+6)</td>
<td>3.825</td>
<td>0.242</td>
<td>3.00e-03</td>
<td>4.50e-02</td>
</tr>
<tr>
<td></td>
<td>SPO vs NPO</td>
<td>21 (15+6)</td>
<td>1.184</td>
<td>0.059</td>
<td>2.85e-01</td>
<td>1.00e+00</td>
</tr>
</tbody>
</table>
Acknowledgements

First, I have to thank Prof. Dr. Frank Oliver Glöckner, Dr. Renzo Kottmann, and Dr. Antonio Fernández-Guerra for giving me the opportunity to do my Ph.D. at the Microbial Genomics and Bioinformatics Research Group.

Also, many thanks to Dr. Marnix Medema, Dr. Renzo Kottmann, Dr. Pier Luigi Buttigieg and Prof. Dr. Frank Oliver Glöckner for their advice, guidance, and interest in this project.

Special thanks to Dr. Antonio Fernández-Guerra for his direct supervision, and all the time he took to review, discuss and improve my research. All the shared code and publications, meetings and advice made a significant difference in my doctoral training.

Thanks to the members of the thesis committee Prof. Dr. Frank Oliver Glöckner, Prof. Dr. Matthias S. Ullrich, Prof. Dr. Marc-Thorsten Hütt, Dr. Marnix Medema and Prof. Dr. Matias Arim for their interest and advice on my work.

Many thanks to the co-authors Dr. Peter Meinicke, and Dr. Albert Barberan for reviewing the work presented here.

I have to thank the MarMic program, especially Dr. Christiane Glöckner, for helping me and facilitating the paperwork during the Ph.D.

Many thanks to Dr. Christian Quast for his help and advice on computational issues.
I also want to thank Dr. Viola Krukenberg, Dr. Pier Luigi Buttigieg, Dr. Gerdhard Jessen, Dr. Ivaylo Kostadinov, MSc. Alexa Garin-Fernández, and all the Microbial Genomics and Bioinformatics Research Group for creating a very friendly environment.

Last but not least, my family for their support during this time.
Bibliography


Babilonia, Joany et al. (2018). “Comparative Metagenomics Provides Insight into the Ecosystem Functioning of the Shark Bay Stromatolites, Western Australia”. In: Frontiers in Microbiology 9, p. 1359. DOI: 10.3389/fmicb.2018.01359.


Bibliography


Logares, Ramiro et al. (2014). “Metagenomic 16S rDNA Illumina Tags Are a Powerful Alternative to Amplicon Sequencing to Explore Diversity and Structure of Microbial Communities”. In: *Environmental Microbiology* 16 (9), pp. 2659–71. DOI: 10.1111/1462-2920.12250.


Niehus, Rene et al. (2015). “Migration and Horizontal Gene Transfer Divide Microbial Genomes into Multiple Niches”. In: *Nature Communications* 6, p. 8924. DOI: 10.1038/ncomms9924.


Parks, Donovan H. et al. (2017). “Recovery of Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life”. In: *Nature Microbiology* 2 (11), pp. 1533–42. DOI: 10.1038/s41564-017-0012-7.


Quince, Christopher et al. (2017). “Shotgun Metagenomics, from Sampling to Analysis”. In: *Nature Biotechnology* 35 (9), pp. 833–44. DOI: 10.1038/nbt.3935.


Yarza, Pablo et al. (2014). “Uniting the Classification of Cultured and Uncultured Bacteria and Archaea Using 16S rRNA Gene Sequences”. In: *Nature Reviews Microbiology* 12 (9), pp. 635–45. DOI: 10.1038/nrmicro3330.