Max Planck Institute for Marine Microbiology

Universität Bremen

Master of Science Thesis:

Ecological implications and characterization of genes of unknown function in the marine environment

University of Bremen
Max Planck Institute for Marine Microbiology
Microbial Genomics and Bioinformatics Group

Matthew S. Schechter

March, 15, 2018

The work of this thesis fulfills requirements
for the degree of Master of Science in Marine Microbiology
at the University of Bremen by Matthew S. Schechter.

First evaluator: Prof. Dr. Frank Oliver Glöckner
Second evaluator: Dr. Pier Luigi Buttigieg
Supervisor: Dr. Antonio Fernández-Guerra

Abstract

The era of environmental metagenomic sequencing has added terabytes of sequencing data to databases. To explore this immense diversity to genes, clustering techniques have been used to group protein coding sequences into protein families (clusters). This has lead to the discovery of novel families and redefined protein diversity. Unfortunately, due to the lack of isolated microbial genomes from the environment and incorrect protein annotation management in databases, many protein families are left without a known function. This constrains microbial functional research and often the “unknowns" are discounted during microbial functional profiling. There have been many approaches to expand the annotations of known proteins to the unknowns, but still a large proportion of protein families remain uncharacterized. Additionally, there have been attempts to categorize and classify the unknown, but sufficient categories have not been described to include all unknown environmental protein coding sequences. The Vanni *et. al.* (in prep) workflow takes a novel approach to protein clustering, focusing on cluster quality and categorization of the unknown clusters. In this thesis, the resulting clusters from Vanni *et. al.* workflow were used to investigate the distribution and biogeography of the unknown in the TARA Oceans dataset. This investigation demonstrates that when utilizing the whole metagenomic sample (knowns and unknowns) there is increased variation in ordinations which leads to clearer sample site separation. It is also found that the unknown fraction of protein clusters have distinct characteristics of being adaptive proteins. Finally, evidence is presented of an ubiquitous unknown fraction of protein clusters throughout the world’s oceans.

Abbreviations

DUF Domain of Unknown Function

EU Environmental Unknowns

FAIR Findable Accessible Interoperable Reproducible

GU Genomic Unknowns

HMM Hidden Markov Model

Kwp Knowns without PFAM annotation

ORF Open Reading Frame

OTU Operational Taxonomic Unit

PCA Principal Component Analysis

TP TARA Ocean prokaryotic metagenomes (all depths)

TPS TARA Ocean prokaryotic, surface metagenomes

Glossary

Cluster A group of protein coding sequences that were grouped together by an unsupervised sequence clustering algorithm.

Component A group of clusters from the Vanni et. al. workflow that were aggregated together due to PFAM domain architecture redundancy.

Environmental Unknowns ORF clusters with an unknown function and are not found in sequenced or draft- genomes.

Genomic Unknowns ORF clusters that have an unknown function (e.g. DUF, hypothetical protein) but are found in sequenced or draft-genomes.

Knowns ORF clusters that have been annotated with a PFAM domains of known function.

Knowns without PFAMs ORF clusters that have a known function but do not contain PFAM annotations.

Contents

List of Figures

List of Tables

Introduction

Evolution has yielded an immense diversity of microbial functions. Today, many of those functions remain uncharacterized and exploring the unknown taxa and functions of the world's microbiomes should be a “central priority for biologists" (Bernard et al., 2018). Even after a century of functional characterization of the model organism *Saccharomyces cerevisiae*, greater than 30% of its protein coding sequences do not have a clear function (Ellens et al., 2017). To shed light on the unknown functions of the microbial world, an important first step is to categorize and quantify how much is already known and how much is not known. With modern sequencing technologies and large scale sampling projects, this is now possible.

The emerging fields of metagenomics and shotgun sequencing have allowed for the description of the total genetic content in environmental samples. Rather than amplifying specific target DNA sequences (i.e. amplicon studies), metagenomics has the potential to simultaneously unveil both function and taxonomy of the microorganisms present in a sample. This allows for the comparison and quantification of differences in microbial physiological traits between sites. With the onset of next generation sequencing (NGS), cost per sequence has dramatically decreased down to 1 cent per megabase (https://www.genome.gov/sequencingcostsdata/, accessed 12.03.2018). Owing to this, large scale metagenomics sampling projects such as Global Ocean Sampling (GOS) (Rusch et al., 2007), TARA Oceans (Sunagawa et al., 2015), and Ocean Sampling Day (OSD) (Kopf et al., 2015) have sampled the world's oceans and added terabytes of novel sequence data to databases.

GOS was the first attempt to create a snapshot of ocean genetic diversity with some of the largest sampling transects and collections of metadata ever accomplished (Rusch et al., 2007). Its dataset generated 6.12 million open reading frames (ORFs) from its Sanger sequencing long reads (Yooseph et al., 2007). Additionally, the

GOS dataset lead to fascinating insights into marine microbial ecology, including the elucidation of the ocean's dominant microbial genomes (i.e. *Prochlorococcus* and SAR-11) and uncovering more microbial diversity in the ocean than previously thought (Nealson & Venter, 2007, Rusch et al., 2007, Venter, 2004, Yooseph et al., 2007).

Inspired by the GOS dataset, The TARA Oceans Expedition went a step further, not only increasing the breadth of sampling, but adding a systematic exploration of the sunlit ocean with the addition of more filter fractions and depths, and unprecedented sequencing depth. The first analysis of the prokaryotic fraction of the TARA dataset yielded a 40 million non-redundant reference gene catalogue (Sunagawa et al., 2015). Later, a metatranscriptome analysis of the eukaryotic fraction added a separate 116 million non-redundant reference gene catalogue (Carradec et al., 2018). Investigation into the prokaryotic dataset uncovered temperature as the main driver of microbial function and diversity, and revealed that the core functionality of the ocean microbiome compared to the human microbiome is 73% similar (Sunagawa et al., 2015).

From data to knowledge

Microbial knowledge discovery from GOS and TARA would not have been possible without the ability for predicted genes (i.e. open reading frames - ORFs) to be annotated. With annotated ORFs, overall community functional fingerprints can be compared between samples. Analysis of the TARA ocean data revealed that upon aggregation of functions within metagenomes, there is functional redundancy throughout the world's ocean (Louca et al., 2016). Yet, due to the sheer number of ORFs generated from these datasets, in the recent years, a variety of tools have been developed to efficiently compare and annotate environmental sequences with characterized sequences in databases. Algorithmic improvements to the original BLAST program (Altschul et al., 1990), like DIAMOND (Buchfink et al., 2014) or MMSEQS2 (Steinegger & Söding, 2017) have allowed for local alignments of query sequences against a target sequence database in a high-throughput manner. However, sequence similarity based annotation methods have limitations if the scientific question at hand is detecting remote homologies or accurately assigning function. In fact, sequence similarity does not inherently imply homology or function, but is merely a metric to describe nucleotide resemblance (States & Boguski, 1991). To detect functional similarity and/or investigate evolutionary based questions of proteins (e.g. homologues, paralogues, and conserved domains), profile based tools should additionally be used.

A common way to create protein profiles is by calculating Hidden Markov Models (HMM) from multiple sequence alignments (MSAs) of protein families. Pfam (Finn et al., 2015) is an example of a profile database that manually curates protein families of highly conserved functional regions within proteins called domains. Other examples of protein family profile databases are Clusters of Orthologous Genes (COG) (Tatusov, 2000), SFams (Sharpton et al., 2012), and eggNOG (Huerta-Cepas et al., 2015). Protein profile databases differ in their curation methodology, but manual curation of protein families is the best way to ensure high quality. This is because an expert can add biological context to a protein family's MSA and identify the spurious members. Unfortunately, due to the size of nucleotide databases and large metagenome projects, the amount of sequences now far outweighs the time constraints of manual definition and curation of protein families.

To handle this amount of data, unsupervised clustering algorithms are being used to define an initial set of related proteins. These algorithms follow two main approaches, graph based and sequence similarity based (Pavlopoulos, 2017). Graph based clustering algorithms start by performing an all-versus-all alignment using tools like DIAMOND or BLAST. Next, they create a sequence similarity network (SSN) using on the criterion of the alignment (e.g. e-value) as edges, then perform a clustering of the SSN to extract components as protein families (MCL is an example of a popular method (Enright et al., 2002)). On the other hand, tools that are based on sequence similarity grouping use k-mer frequencies (MMSEQS2 (Steinegger & Söding, 2017), CD-HIT (Li & Godzik, 2006), UCLUST (Edgar, 2010)). One characteristic of the latter methods is their increased speed. With this, they are well suited to be applied to the massive metagenomic data sets.

In fact, one way to explore protein family diversity of environmental metagenomic ORFs is to cluster them with already recovered ORFs sequences in databases. This can lead to previously characterized protein families being expanded, shrunk, or the creation of novel families. GOS took this approach and yielded 1,700 clusters (grouped sequences after clustering) with no detectable homology or function (Yooseph et al., 2007). Although new families may emerge when including new sequences, some may have a known function, but a vast majority will remain unknown.

De-bunking the unknown

There have been different strategies to expand annotations to unknown protein families including protein structure/sequence similarity analysis and domain co-occurrence networks. Jaroszewski et al. (2009) structurally characterized 250 PFAM domains of unknown function (DUFs) using structural genomics. After comparing DUFs to known proteins on a structural level, it was determined that most of the DUFs were distant variants and not completely novel. Mudgal et al. (2015) expanded upon this work with increased sensitivity and higher quality structural databases which lead to 20% of DUFs being annotated as distantly related domains. Both studies discuss that the protein universe may soon be saturated from a structural perspective and most “novel" sequences are due to protein diversification from environmental niche selection. Based on the assumption that DUFs are products of niche adaptation, Buttigieg et al. (2013) postulated that DUFs may co-occur with domains of known function between metagenomic samples. To explore this, they calculated the co-occurrence of DUFs with characterized domains in the GOS dataset and used correlation and network exploration to infer function, i.e. “guilty by association" method.

Even after deploying different strategies to annotate functions to new protein families, some families remain unknown. It is then an important step to confirm clusters that do not have detectable homology or function, are indeed novel and legitimate biological entities. This is a unique challenge in metagenomic data due to short reads, insufficient sequencing depth (Barrientos-Somarribas et al., 2018, Parks et al., 2011). Moreover, gene prediction algorithms have limitations and can yield inaccurate ORFs predictions leading to spurious proteins, which can lead to spurious protein families. Dedicated databases have been curated to detect and remove spurious proteins (e.g. AntiFAM (Eberhardt et al., 2012)). Additionally, specific tools like Spurio (Höps et al., 2018) have been designed to score query ORFs on their likelihood of being spurious.

Another strategy to confirm protein family novelty is to determine if the members of the cluster are under positive or negative selective pressure by calculating non-synonymous to synonymous substitutions (Ka/Ks) (Barrientos-Somarribas et al., 2018, Yooseph et al., 2007). If ORFs have a Ka/Ks value close to 1, there is a strong indication for no selective pressure, thus a predicted ORF may not be a real coding sequence (Nekrutenko, 2001). One issue with this validation strategy is, it assumes that all proteins in a metagenomic sample are under the same environmental pressures. When filtering out novel clusters with the criteria of Ka/Ks being closed to one, an assumption is made that all proteins are under purifying selection.

Categorizing the unknown

Even after various strategies to expand annotations and debunk unknown clusters, numerous protein families still indeed remain with unknown functions. The concept of “unknowns" in microbiology is a recognized problem that has implications for both fundamental and applied biological research. The term “biological dark matter" was first used by Marcy et al. (2007) to describe the uncultured majority of microbes that could only be indirectly detected via environmental sequencing methods. Later, the term would also be used to refer to protein families with no detectable homology to characterized proteins from sequenced genomes (Ellens et al., 2017, Fischer & Eisenberg, 1999, Perdigão et al., 2017, Roux et al., 2015).

When categorizing and defining the unknown, it is important to quantify in reality how much is already known. Perdigão et al. (2017) approached the unknowns from a structural perspective by making the “Dark Proteome Database". Protein sequences are ranked on a scale of “darkness" based on how confidently parts or the whole sequence can have of an inferred 3D structure. This is done by calculating MSAs to protein structural databases. Wyman et al., 2017 categorized SFams (clusters created via a graph-based approach from sequences of genomes) as FUnkFams (protein families of unknown function) if the clusters could not be annotated by the Pfam or the NCBI Conserved Domain database.

For this thesis, I used the clusters of ORFs produced from the work of Vanni et. al. (in prep.)(https://orcid.org/0000-e2-1124-1147). This workflow builds upon the motivations of the methods mentioned above but focuses on ORF cluster validation and deeper categorization of the unknown. Vanni et. al. used MMSEQS2 to cluster the 322,248,552 ORFs from The Human Microbiome Project (HMP), TARA Oceans (Sunagawa et al., 2015), Ocean Sampling Day (OSD) (Kopf et al., 2015), Global Ocean Sampling Expedition (GOS) (Yooseph et al., 2007), and Malaspina (Duarte, 2015). This dataset yielded 32,465,074 clusters that were then validated downstream. In parallel to clustering, the ORF dataset received Pfam annotations which were used to assist in cluster validation and unknown categorization.

Cluster validation is a critical step to protein clustering because unsupervised clustering algorithms can include non-biologically relevant ORFs into clusters due to incomplete or spurious ORFs gene duplications, and genome rearrangements. Vanni et. al. validated clusters for compositional and functional homogeneity. At the end of the validation

steps, 2,953,903 clusters were successfully passed on to be categorized in terms of their level of unknown.

The goal of the categorization was to quantify the amount of novel versus characterized clusters based on the ability to annotate their function. For this purpose, the ORF cluster space was classified in the following categories:

Knowns with PFAM (Knowns)
ORF clusters that have been annotated with a PFAM domains of known function.
Knowns without PFAMs (Kwp)
ORF clusters that have a known function but do not contain PFAM annotations.
Genomic Unknowns (GUs)
ORF clusters that have an unknown function (e.g. DUF, hypothetical protein) but are found in sequenced or draft-genomes.
Environmental Unknowns (EUs)
ORF clusters with an unknown function and are not found in sequenced or draft-genomes.

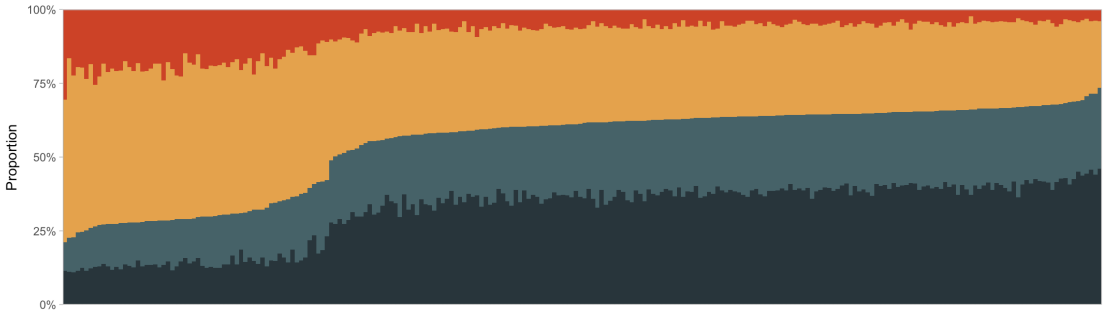

Figure 1.1: **Proportions of Vanni et. al. cluster categories in the TARA Ocean metagenomes.** The X axis denotes each of the TARA metagenomes while the Y axis is the relative proportion of the different Vanni et. al. cluster categories. Some of the metagenomes have greater than 70% Unknowns. (Red = EUs; Yellow = GUs; Blue=Knowns; Black = Kwp)

In 16S ribosomal DNA (rDNA) amplicon studies, amplified sequences are clustered together to form operational taxonomic units (OTU). This is advantageous because it reduces the data set for downstream computations and allows for annotation of the cluster representative with taxonomy. If each individual amplicon was treated as an OTU, alpha and beta diversity measurement computations would be very expensive thus making low resolution hypotheses about diversity difficult to make. A similar idea was applied by Vanni et. al. to their dataset to aggregate clusters into larger components due to the Known clusters exhibiting domain architecture redundancy.

This may have been caused by the limitations of the clustering method used (based on sequence similarity) to detect distant homologies and the threshold selected (30% of similarity), resulting in multiple clusters with the same domain architecture.

In total, there were 9,687 clusters with more than one PFAM annotation and 18,368 PFAM annotation combinations in the final clusters. To address this, they used the consensus sequence of the Knowns to create a sequence similarity network (SSN) and extracted components.

The term “components" will now be used for the rest of this thesis to describe aggregated clusters occurring in the TARA ocean dataset.

Aim of this thesis

The cluster components are ideal to address the main objective of this thesis - the exploration of the ecological significance and functional biogeography of the unknown functional diversity in the world's oceans.

To echo back to the 16S rDNA amplicon analogy discussed previously, OTUs that do not receive a taxonomical annotation can still be included in ecological analyses. In this thesis, I use this philosophy but with unannotated ORF components. In the TARA Ocean data set, some samples have greater than 70% of the metagenome categorized as unknown (Fig. 1.1). Until now a large proportion of this functionally uncharacterized fraction has not been used for the study of the prokaryotic fraction (Roux et al., 2015). In this thesis, it is demonstrate that including the whole metagenomic sample (Knowns and Unknowns) adds more variance to sample site ordinations. Also, it is shown that utilizing the unknowns in functional biogeography can lead to new insights into genetic differences between sites in different ocean regions. Finally, it is shown that there is a potential group of EU proteins that are ubiquitous throughout the ocean. To increase the evidence that the detected ubiquitous EUs are indeed real proteins in the environment, we localized them in the contigs (assembled sequence fragments from metagenomic data) of high quality, manually curated set of metagenomic assembled genomes (MAGs) extracted from the TARA ocean project (Delmont et al., 2017).

Materials and Methods

2.1 Initial dataset preparation

The data used in this thesis was produced from the Vanni et. al. ORF clustering workflow (https://orcid.org/0000-e2-1124-1147) which includes The Human Microbiome Project (HMP), TARA Oceans, Ocean Sampling Day (OSD), Global Ocean Sampling Expedition (GOS), and Malaspina (Table 2.1). This dataset of 1,837 metagenomes (322,248,552 ORFs) was clustered into 32,465,074 total clusters, of which 2,953,903 were considered as high-quality. These high quality clusters were aggregated into components, then subsetted for only ones that occurred in the TARA prokaryotic enriched data set. This yielded a total of 18,644 components for the final data set used in this thesis (Table 2.2).

Table 2.1: Vanni et. al. initial clustering dataset

Metagenomic Project	**Number of Metagenomes**	**Number of ORFs**
Human Microbiome Project	1249	162,687,295
TARA Oceans	242	111,903,261
Ocean Sampling Day	150	7,015,383
Global Ocean Sampling	80	20,068,580
Malaspina	116	20,574,033
Total	**1,837**	**322,248,522**

The TARA prokaryotic component abundance dataset (prokaryote-enriched fractions: 0.22 to 1.6 mm, 0.22 to 3 mm; n = 139) was filtered for low total count samples and low mean proportion components. Samples were removed from the analysis if their total component counts were less than the median total component count subtracted

from 1.5 * MAD (Median Absolute Deviation) of the total counts. Next, components with a mean proportion less than 1e-5 across all samples were removed from the dataset using the Tidyverse package in R (Wickham, 2017). The resulting filtered component abundance matrices were then stored as phyloseq objects (McMurdie & Holmes, 2013, Oksanen et al., 2017). R version 3.4.3 (2017-11-30) was used for all R libraries mentioned in this thesis (R Development Core Team, 2008).

Table 2.2: TARA subset of components used for this thesis

	Knowns	**GUs**	**EUs**	**Kwp**	**Total**
Components	3997	6497	1534	6616	18,644

2.2 Data standardization

For PCA (principal component analysis) ordinations the dataset was center log ratio (CLR) transformed (Piepel & Aitchison, 1988) using the R bioconductor package, microbiome (Lahti et al., 2012-2017). Cumulative sum scaling (CSS) normalization from the R package metagenomeSEQ (Paulson et al., 2013) was applied to the dataset for the distance-decay analysis and nMDS (non-Metric Dimensional Scaling) ordinations.

2.3 Indirect gradient analysis of TARA Ocean sample sites

The CLR transformed, component abundance matrix was visualized using PCA. These calculations were performed using the R package Vegan (Oksanen et al., 2017) and graphically arranged using the R package ggpubr (Kassambara, 2017). Sample sites in the ordinations were colored by sample depth category and the temperature to explore gradients and clustering patterns. Contextual data from the TARA Ocean project was used from Sunagawa et al. (2015).

Residuals of the PCA ordination were visualized to screen for normality and explore if the relationships between variables were linear. Additionally, a scree plot was plotted to visualize the variance captured by each principal component. Next, a NMDS plot was calculated, based on a Bray-Curtis dissimilarity matrix, to see if sample site clustering was similar to the PCA via a distance based ordination method. Additionally, a Shepard plot was visualized to explore how well the ordinated distances in the

NMDS were a good representation of the actual distances within the dissimilarity matrix.

Hypotheses about the clustering results observed in the ordinations were tested for significance. This was done by calculating a PERMANOVA using the adonis function in the VEGAN package in R. Additionally, clustering due to beta-dispersion was explored by using the betadisper function, also in the Vegan package in R (Oksanen et al., 2017).

2.4 Niche breadth analysis of components

To quantify scores of component theoretical niche and resource occupancy, Niche Breadth (B) was calculated (Levins, 1966).

$$B = 1/\sum_{i=1}^{N} P^2 ij$$

B is one divided by the sum of all proportions of a biological entity (P) from 1 to N sites of biological entity i through biological entity j. From a macro-ecological perspective, B is one divided by the sum of all proportions a species represents in all the samples measured. The fact that P is squared in the denominator of the equation removes some additive effect of the summed proportions.

To classify components as having a “wide" or “narrow", a null distribution was created of each component B score. The original component abundance matrix was randomized 100 times using the Vegan package with the quasiswap count method in the function *nullmodel* (Miklós & Podani, 2004, Oksanen et al., 2017). This method randomizes abundance matrices by mixing up numbers of 2x2 matrix subsets within the larger matrix. Additionally, the method maintains abundance matrix column and row sums to preserve original attributes of the matrix in the new randomized matrices. Once the distribution for each component is calculated, if a component score was in the top 2.5% of its distribution, it was classified as “wide". If it was in the bottom 2.5% of the distribution, it was classified as “narrow". The distributions of B categories were visualized using the R package ggplot2 (Wickham, 2016) and assembled using the R package ggpubr (Kassambara, 2017).

2.5 Screening beta diversity for geographic distance effects

For this method section the CSS transformed, component abundance matrix was used. First, Bray-Curtis dissimilarity was calculated using the Vegan package function vegdist (Oksanen et al., 2017). Distance-decay plots were determined by plotting the Haversine distance between TARA Ocean surface sample sites against the Bray-Curtis dissimilarities. Haversine distance was calculated using the R package geosphere (Hijmans, 2017). The resulting graph was visualized using the R package ggplot2 (Wickham, 2016) and assembled using the R package ggpubr (Kassambara, 2017). We used the function mantel and partial.mantel to test the correlation (spearman) between the Bray-Curtis dissimilarity and the geographic distance matrices. In the case of the partial Mantel test we use the deltaTemp matrix, the absolute temperature difference between TARA samples in degree Celsius. Mantel tests went through 9999 permutations.

TARA metagenomic sample sets defined by Delmont et al. (2017) were used to separate TARA samples into three oceanic regions: Pacific, Atlantic, and Indian. In the Atlantic subset, samples were removed that were below the equator to focus analysis on the transect of the Gulf Stream. Distance-decay analysis was done separately for each region to remove biases of continental divides. Finally, component categories were separated into ubiquitous and non-ubiquitous. Ubiquitous components are defined as components that have a mean proportion greater than 1e-5 and be found in every sample in the TARA ocean project. To be categorized as non-ubiquitous, components only had to have a mean proportion greater than 1e-5.

2.6 Analyzing the ubiquitous EUs

The ubiquitous EUs were tested to see if they were real protein clusters and not artifacts of metagenomic sequencing or assembly. First, spurious proteins (falsely predicted ORFs) were filtered out by searching the EU clusters consensus sequence against the antiFAM database (Eberhardt et al., 2012) using the hmmsearch program from the HMMER suite (Eddy, 1998) with the *–cut-ga* significance threshold. Results from the search were then parsed using e-value $>$ 1e-5 and coverage $>=$ 60% as additional thresholds.

The second step to legitimize the ubiquitous EU clusters was to detect remote homology using an iterative HMM-HMM profile search of the EU clusters against the Uniclust database (Mirdita et al., 2016). We used HHBlits from the HHsuite software package (Remmert et al., 2011) with two iterations. All queries with a probability larger than 90% to any target sequence in the database were discarded. Next, we attempted to assign taxonomy to the remaining ubiquitous EUs by running Kaiju in greedy mode to ensure sensitivity and precision (Menzel et al., 2016).

Finally, we mapped the ubiquitous EUs to high quality, manually curated MAGs from the TARA Ocean Project to see if they are found in populations of genomes (Delmont et al., 2017). We aligned the cluster members from the ubiquitous EUs with FAMSA (Deorowicz et al., 2016) and we used the program hhmake from the HHSUITE to create hidden Markov model profiles (HMM). All EU HMM were stored, indexed, and retrieved using the file based storage software ffindex (https://github.com/soedinglab/ffindex_soedinglab, accessed 12.03.2018). EU HMM were retrieved and converted to MMSEQ2 format using *convertprofiledb* from MMSEQS2. Next, the predicted ORFs of the TARA MAGs were converted to a MMSEQS2 database using the MMSEQS2 command *createdb*. Finally, each ORF was mapped to the profile with the MMSEQS2 command search with the parameters *-e* 1e-25, --cov-mode 2 and -c 0.8. The results were then converted to a BLAST-tab formatted database using *convertalis* program from MMSEQS2, then parsed and plotted with the ggplot2 package. Contigs containing the interesting ORFs were retrieved from the Anvi'o profiles using the program *anvi-export-gene-calls* from Anvi'o v4 (Eren et al., 2015). The functional annotation of the contigs was performed by Prokka Seemann (2014) in metagenomic mode. The gene plots were drawn with the R package genoPlotR (Guy et al., 2010). Muscle (Edgar, 2004) was used to create multiple sequence alignments of the components of interest, and the conserved consensus sequence logos were drawn using the WebLogo web server (Crooks, 2004). Nucleotide sequences of the clusters with hits in the MAGs were also searched against NCBI nt and Microbial genomes using blastn from the BLAST package Camacho et al. (2009).

2.7 Code and data availability

All source code is available in a public repository. Additionally, all data used in this thesis will be available as FAIR data to ensure open science and access (Wilkinson

et al., 2016).

Github: https://github.com/mschecht/Unknown_unknowns

Figshare: DOI - 10.6084/m9.figshare.5979658

Results and Discussion

3.1 Balancing noise removal and sparsity of component abundance dataset

The ecological analyses in this thesis started with two datasets of component counts from the TARA ocean prokaryotic metagenomes from all depths (TP) and only the surface samples (TPS). These two datasets were filtered first for low total count samples and second for low mean proportion components (Methods 2.1). This ensured that: (1) extremely low abundant components were filtered out; (2) sequencing noise was removed; and (3) overall size of the resulting abundance matrices was decreased to lower computational resources for calculations.

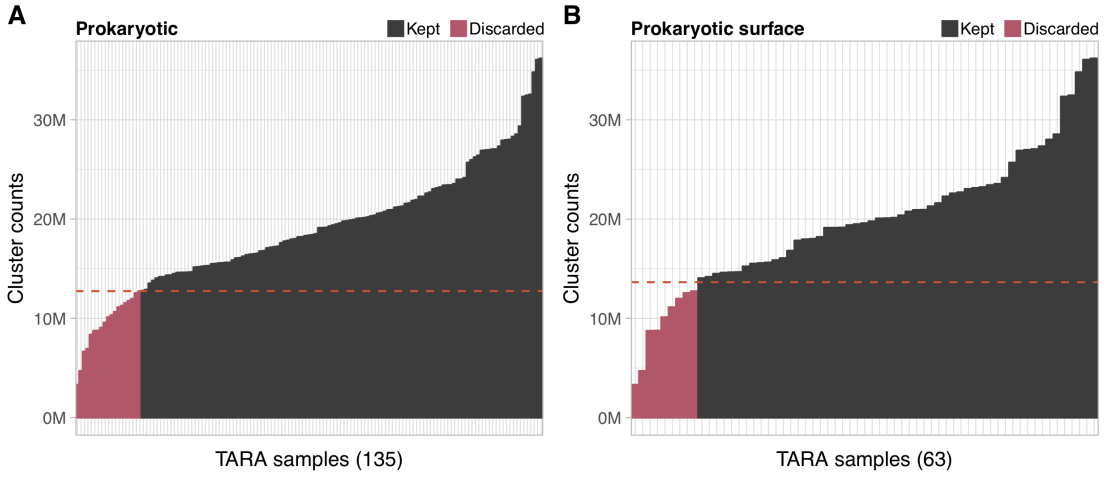

Figure 3.1: **Step 1 of dataset filtering - removing low count samples.** X axis denotes each sample in the dataset while the Y axis denotes total cluster counts for each sample. Samples with total counts below the value, median subtracted from 1.5*MAD, were removed from the analyses. (A) Total component counts for TARA prokaryotic metagenomes (TP). (B) Total component counts for TARA surface, prokaryotic metagenomes (TPS).

During the first filtering step, nine samples were removed from the TPS dataset and 19 samples were removed from TP dataset. Sample with low total counts, relative to

other samples, have the potential to distort beta-diversity analysis because their total counts could make them more similar to each other due to low counts rather a real beta-diversity signal. The cutoff threshold, Median - 1.5*MAD, was chosen because the MAD statistic was chosen it is a reliable measurement of the center of the data and is not sensitive to outliers (Figure 3.1).

The second filtering step removed all components that had a mean proportion across all samples lower than 1e-5. After this step, the final TPS dataset contained 18,644 components while the TP dataset contained 15,353 respectively. This was an effective cutoff for both the TP and TPS datasets because overall component content and counts were able to be maximized while sparsity was lowered. The less sparsity an abundance table has, the more power the abundance variables have in explaining the dataset. Yet, a side effect of decreasing sparsity is removing counts which decreases the total amount of information the dataset has. In both datasets, if the mean proportion requirement was increased to 1e-4, sparsity would have reached close to 0. This could have led to a large amount of total counts and components being sacrificed. For this analysis, we did not want to remove the low abundance signals in order to capture more functional diversity in the dataset. A summary of the 1e-5 mean proportion plot against sparsity, total counts, and proportions can be seen in Figure 3.2.

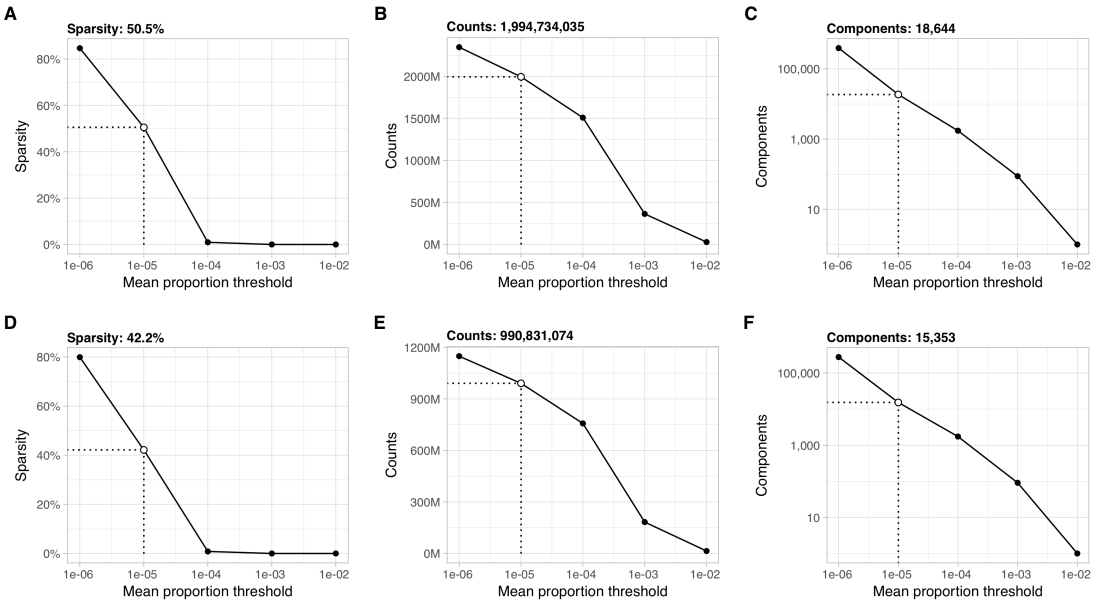

Figure 3.2: **Step 2 of dataset filtering - removing low abundance components.** Metagenome component mean proportion (1e-5 cutoff) is plotted against dataset sparsity, total cluster counts, and number of components. A-C represent the TARA prokaryotic, all depths dataset (TP). D-F represent the TARA prokaryotic, surface dataset (TPS).

3.2 Indirect gradient analysis of TARA Ocean sites based on component abundances

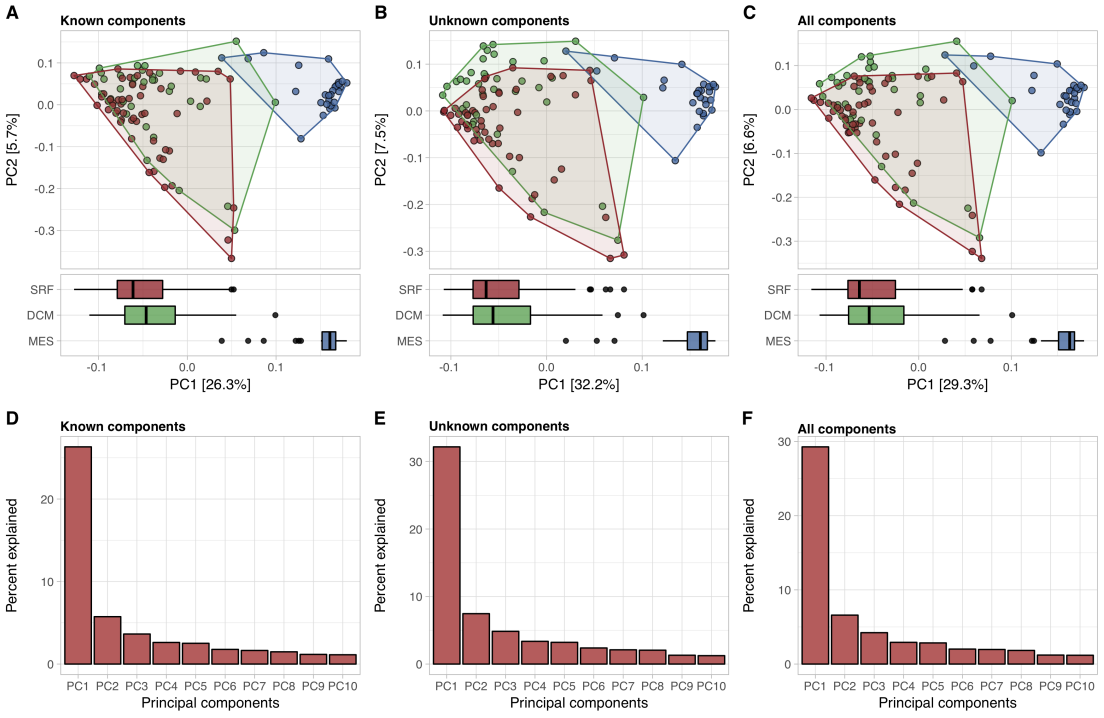

Figure 3.3: **Principal component analysis (PCA) of 135 TARA Ocean prokaryotic metagenome sampling sites from surface (SRF), deep chlorophyll maximum (DCM), and mesopelagic (MES) sites.** A-C) Sites are colored by depth of origin and are defined by proportions of component categories: Unknowns (EU + GU), Knowns (K + Kwp), All (EU + GU + K + Kwp). Below each ordination are bar plots of the distribution of each depth category across the first principal component axis. D-F) Scree plots of each PCA. The X axis denotes each principal axis while the Y axis denotes the percent variance explained by the axis. Residuals of each PCA ordination can be found in the Appendix A.2

Principal component analysis (PCA) of the TARA ocean prokaryotic sample sites produced a meaningful ordination (Figure 3.3 A-C). Due to the effect of compositionality often associated with omics datasets, a center log ratio transformation (CLR) was performed prior to calculated the PCA (Gloor et al., 2017). Scree plots indicated that in all three ordinations the majority of variance was captured in the first two principal component axes (PC) (Figure 3.3 D-F). Additionally, histograms of residuals were calculated and showed a bimodal distribution around 0 (Appendix A.1). NMDS ordination was calculated using pairwise site Bray-Curtis dissimilarity (Appendix A.2). The calculation was iterated 20 times with a stress around 0.8 for all categories (Table 3.1).

Table 3.1: Stress Values for NMDS ordination

	NMDS stress
Unknowns (EUs + GUs)	0.07997218
Knowns (K + Kwp)	0.08218745
All combined	0.07974366

3.3 TARA Ocean Sites cluster based on depth

PCA ordinations of the TARA ocean prokaryotic sample sites clustered based on their depth of origin (surface (SRF), deep chlorophyll maximum (DCM), and mesopelagic (MES) regardless of which component category sites are defined by. The SRF and DCM sites overlap, while the MES samples form their own cluster. Additionally, in all cases the first principal component axis (PC1) explains the most variance in the ordination. PERMANOVA of the ordination against the depth categories and temperature from the associated metadata was significant, allowing the rejection of the null hypothesis that the three depth categories have the same cluster centroids in the ordination (Table 3.2). Additionally, beta-dispersion was not significant in terms of the depth category (Table 3.2) and indicates that clustering was not due to dispersive effects. If beta-dispersion was significant, it would indicate that clustering of sites in the ordination was due to dispersion effects of the data and not entirely due to depth category. But since it was not significant, there is increased confidence in the PERMANOVA results. NMDS ordination based on Bray-Curtis dissimilarity also show the same clustering of sites based on depth category and temperature (Appendix A.2).

Previous ordinations of the TARA Ocean metagenomes have shown site clustering by depth of origin (Louca et al., 2016, Sunagawa et al., 2015). Louca et al. (2016), calculated a metric multidimensional scaling (MDS) of the TARA sampling sites, but instead, defined sites based on aggregated, biogeochemically relevant microbial functions. Additionally, Sunagawa et al. (2015) and Louca et al. (2016) calculated a principal coordinate analysis with sites defined by 16S rDNA gene tags (miTAG) extracted from the metagenomes. Both analyses concluded that prominent environmental factors (e.g. temperature, nutrients levels, access to light) from the different depth layers have a strong influence in shaping the structure and function of the microbial community. This explanation can be applied to this analysis as well.

Regardless of what cluster categories ORFs are from, the environmental factors from the depth layer are selecting components in a similar way. The SRF and DCM samples

Table 3.2: TARA ocean prokaryotic metagenome sampling site ordination PERMANOVA for depth category hypothesis testing

	PERMANOVA		Beta-dispersion
	R^2	p-value	p-value
Unknowns (EUs + GUs)	0.27237	0.001	0.073
Knowns (K + Kwp)	0.22649	0.001	0.121
All combined	0.25064	0.001	0.117

may overlap due to depth proximity. Both SRF and DCM are sunlit environments and have access to fresh primary production carbon sources. On the other hand, the MES samples clusters separately from the SRF and DCM. This is most likely owing to the different environmental factors in the MES, such as lack of solar radiation, lower temperatures, and higher dissolved oxygen and nutrients, which can select for different microbial functions than the epipelagic depths. Additionally, there is increased amounts of recalcitrant carbon sources in the MES because smaller/simple carbohydrate sources are most likely respired at high depths of the biological carbon pump (Azam & Malfatti, 2007). In an environment with more complex carbon sources, a diverse array of microbial functions may be selected for to metabolize a wider range of nutrients.

3.4 Greater variance observed in PCAs when all components are used to define sample sites

In Figure 3.3 A-C, more variance is captured in PC1 of the Unknowns (32.2%) while less variance is captured in the Knowns PC1 (26.3%). The Knowns + Unknowns PC1 variance falls in between (29.3%). The fact that PC1 of the Unknowns ordination has the most variance infers that sites are more dissimilar when they are defined by the unknown fraction of components. Also, by accounting for both the Known and Unknown fraction in the All ordination, the site dissimilarity is effectively increasing by including the Unknown fraction. Similar to defining sites with all detected OTUs regardless if taxonomy is assigned to them, ORF clusters, with or without annotation, can be tracked as a functional entity in metagenomes and help separate samples in higher resolution. Louca et al., 2016 ordinated the TARA Ocean sites based on biogeochemical relevant microbial functions and determined there was functional redundancy in the ocean. This conclusion may have been constrained by the fact that sites were only defined by known functions. If the Louca et al., 2016 sites were defined with the complete functional repertoire of the metagenomic sample, their results may have had increased dissimilarity. It is clear that when sites are defined by the knowns

Table 3.3: TARA ocean prokaryotic metagenome sampling site ordination PEMANOVA for temperature hypothesis testing

	PERMANOVA	
	R^2	p-value
Unknowns (EUs + GUs)	0.27237	0.001
Knowns (K + Kwp)	0.22649	0.001
All combined	0.25064	0.001

and the unknowns, more variance is added to the ordination. This may indicate that the unknown fraction could be site specific adaptive proteins.

3.5 Unique component composition in Southern Ocean Samples

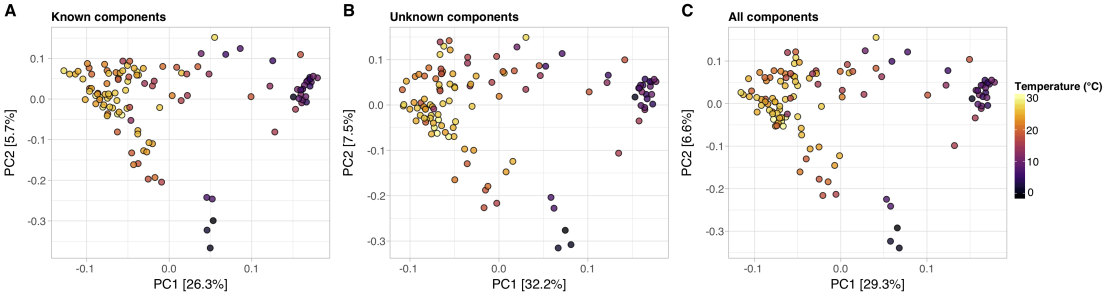

Figure 3.4: **Principal component analysis of the prokaryotic fraction of the 139 TARA ocean sampling sites from all depths colored by temperature.** Sites are defined by proportions of component categories: Unknowns (EU + GU), Knowns (K + Kwp), All (EU + GU + K + Kwp)

A temperature gradient across PC1 of Figure 3.4 was ordinated in the TARA sampling sites. Samples in the SRF and DCM cluster have warmer temperatures while the MES cluster has the coldest temperatures. Colder samples from the SRF and DCM ordinate closer to the MES cluster. These results are in line with Sunagawa et al., 2015 who found that temperature is the main driving factor for community composition in the ocean. The SRF and DCM have warmer temperatures due to increased solar radiation, while heat is dissipated in the MES. The clustering of the ordinations based on the origin depth of the sample may be due to temperature which could then select for different microbial functions (PERMANOVA for temperature found in Table 3.3).

An exception to the temperature gradient across PC1 is the three Southern Ocean samples (TARA site 84 SRF, DCM and site 85 SRF) at the bottom of the ordination.

Even though they are some of the coldest samples in the dataset, they do not follow the temperature gradient along PC1 and do not cluster with the rest of the cold MES samples. In fact, PC2 appears to drive their separation from the rest of the samples.

These three Southern Ocean samples may have a unique component composition due to their placement in the Antarctic Circumpolar Current (ACC). The ACC is a located in the Southern Ocean and presents a unique biome compared to other oceanic currents due to its strength, isolation, dramatic seasonal light flux and nutrient levels. Strong westerly winds around Antarctica drive the current due to pressure gradients caused by surface wind friction and density differences. Because no continents obstruct its flow, the ACC is the only ocean current system that circles the earth on a similar latitude range. This allows to be the fastest current in the ocean with speeds around 130e6 $\frac{m^3}{s^-1}$ through Drake passage (MarMic Physical Oceanography Lecture, 2017).

The ACC also has unique physicochemical properties that may also influence specialized microbial functions. Due to low altitude wind cells, Aeolian dust input is limited which makes mineral concentrations limited. On the other hand, strong upwelling increases concentrations of other nutrients, such as nitrate and phosphate. This makes for a unique environment because despite high nutrients, chlorophyll levels are low due to Fe limitations (Cavicchioli, 2015, Martin et al., 1990, Tagliabue et al., 2014). Additionally, the ACC has unique biogeography due to dramatic polar light cycles. During summer season, long periods of light generate high levels of primary production, while during low light winter's, the overall microbial community switches to different ecological strategies such as heterotrophy and predation (Milici et al., 2017, Wilkins et al., 2013). Due to the reasons listed above, unique functions for local adaptation in the Southern ocean may be driving the dissimilarity from the other TARA ocean samples.

Interestingly, TARA site 85 MES metagenome did not cluster with the Antarctic SRF and DCM metagenomes at the bottom of the ordination, but rather clustered with the rest of the MES metagenomes. This reinforces that the MES is a different environment compared to the SRF and DCM. Even though TARA site 85 MES was sampled in the ACC, the deep depth layer environmental parameters have a stronger influence on its component composition. To further explore this observation, direct gradient analysis should be applied to the Southern Ocean to resolve its unique microbial functional properties.

Using sites to ordinate the TARA ocean sites revealed a distinct functional difference between the Southern Ocean samples and the rest of the TARA samples. This was in part due to the use of components to define sites. No matter which category of components was used (Knowns, Unknowns, All), the Southern Ocean samples we separated. This is an indication that defining sites based off of component composition may help resolve fine differences between sites in ecological analyses in the future.

The ordination of sample sites based on the different component fractions preliminarily indicated that the Unknown fraction may be able to increase site dissimilarity. To investigate further, we performed a Levin's Niche Breadth analysis to explore component niche occupancy.

3.6 Niche Breadth analysis of component distribution

According to Levins (1966), Niche breadth (B) is the theoretical, multidimensional range of resources and habitats a species can occupy and access. Here we used the equation 1 to quantify the niche specialization of all prokaryotic components in the TARA oceans sampling sites. B was originally designed to classify macro-ecology abundance biogeography, i.e. birds and deer in different environments. Ocean microbial systems are different compared to macro-ecological systems for a few reasons. First, the definition of species in microbial ecology is still debate. Second, microbes are subjected to functional and physiological plasticity due to horizontal gene transfer. With these factors in mind, the Niche Breadth measurement may still be informative for analyzing the distribution of components. Recently, the measurement was used to describe the biogeography of operational taxonomic units (OTUs) in coastal Antarctic lakes (Logares et al., 2012). Logares et al. (2012) used B to denote “generalists" vs “specialists" to explore niche specialization of microbes within the domain of the sampling regime. It has also been hinted that the distributions and associations of genes with ecological niches may be an effective strategy for determining the role of accessory/adaptive genes in pangenomic diversity (Shapiro, 2017). Although more research needs to be done, these studies have laid groundwork for using Niche Breadth in the context of component distribution.

This niche breadth analysis revealed that greater than 99% of the EU components have a narrow niche breadth. Also, the majority of GUs are narrow while the Knowns are more evenly distributed between narrow and wide components. In the sample

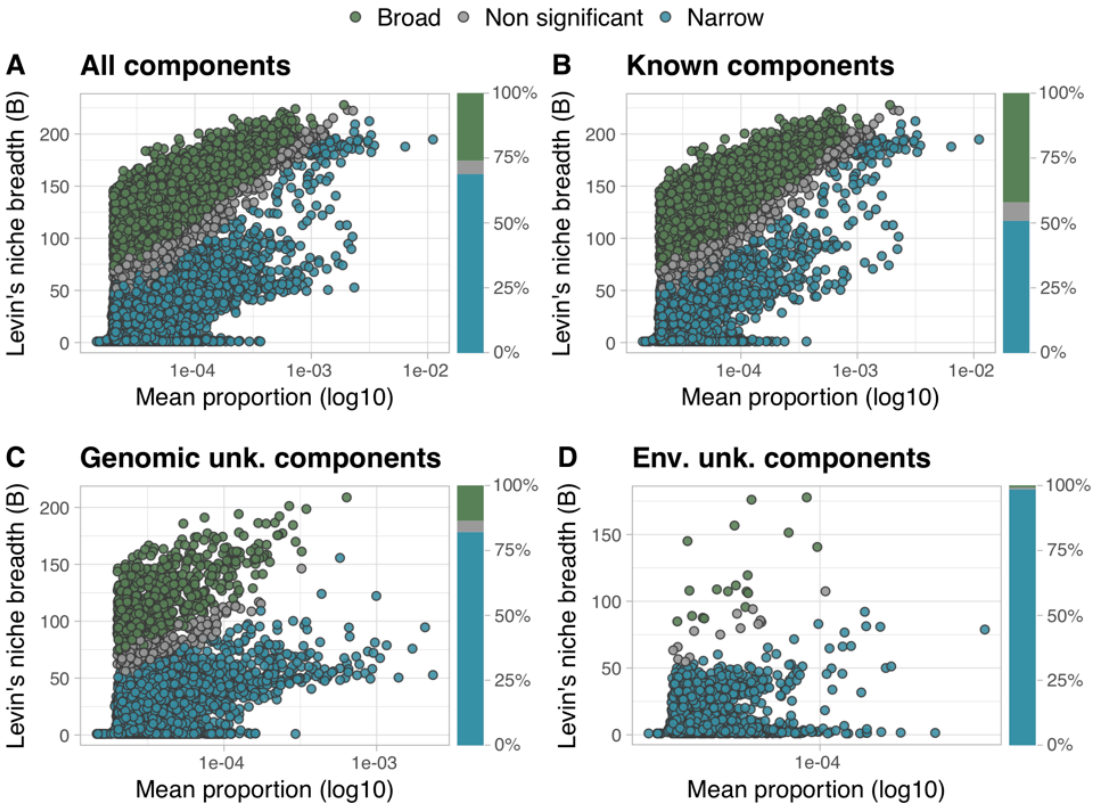

Figure 3.5: **Levin's Niche Breadth (B) scores of the component categories.** The x-axis denotes the average, relative abundance of a component across all TARA prokaryotic metagenomes. The y-axis denotes a component B scores. Components are colored by their classification due to their B score. The bar charts on the right of each plot represent the relative proportions each B classification in the component category. See section 2.4 for description of classification of the components based off their B score.

site ordination, it was found that including the Unknown fraction increases variance between sites. This result compliments the Niche Breadth analysis because the majority of the GU and EU components are categorized as having a narrow B (i.e. found in a few samples in higher proportions). EUs in general being associated with narrowness may indicate that they are selected in prokaryotic genomes due to specific environmental factors that occur in niche spaces (TARA ocean sampling sites). These components may be uncharacterized EUs because they are non-core, accessory functions from members of the rare biosphere (McInerney et al., 2017). Functions from the rare biosphere are difficult to characterize due the inability to culture members and assemble their genomes into metagenomic assembled genomes due to low read coverage.

In Figure 3.5 D, there is a group of EU components that are narrow and with mean proportions greater than 1e-4. This indicates that they are found in relatively high abundance in only a few samples. These components may be truly adaptive and could be providing the microbes with a selective advantage. Additionally, in Figure 3.5 B and

C, there are Known and GU narrow components with mean proportions greater than 1e-3. These components may also be adaptive components and could be immediately investigated due the Knowns having a Pfam annotation and the GUs being associated with taxonomy.

Though taxonomy and functional annotations are not available for the EUs by definition, the niche breadth analysis still provides more evidence that they are environmentally adaptive proteins in microbes that have be selected for by specific niche factors at the site. To increase evidence for this hypothesis we examined the component categories from a biogeographical point of view.

3.7 Beta diversity analyses

Microbial beta-diversity was explored in relation to geographical distance between sampling sites. The general hypothesis for this analysis is that sites will increase in genetic dissimilarity (beta-diversity) with increasing distance. Here we perform a distance-decay analysis while defining beta diversity by different component categories: Unknowns (EU + GU), Knowns (K + Kwp), and All (Known + Unknown).

The results of distance-decay analysis of the TARA ocean surface, prokaryotic metagenomes samples shows that beta-diversity between sites is the most dissimilar when sites are defined by the Unknowns in all three ocean regions (Figure 3.6 B, E). When sites are defined by the Known fraction, they are the most similar (Figure 3.6 D). It is also shown that when the Unknown fraction is included with the Known fraction that overall dissimilarity between sites is increased (Figure 3.6 B).

The fact that sites are more dissimilar when defined by the Unknowns is another indication that the Unknown fraction may represent proteins specific to niche adaptation of particular sites. The environments at each sample site may have a unique signature of adaptive proteins and thus increase overall site dissimilarity. Another aspect to consider is the Unknown fraction is around 33% of all clusters in this analysis. Because the abundance matrix of a site, when defined by the Unknowns, will always be smaller than then Known abundance matrix (due to the proportions of the component categories), the Unknowns matrix may be more sensitive to unique site abundance signals, thus

increasing beta dissimilarity.

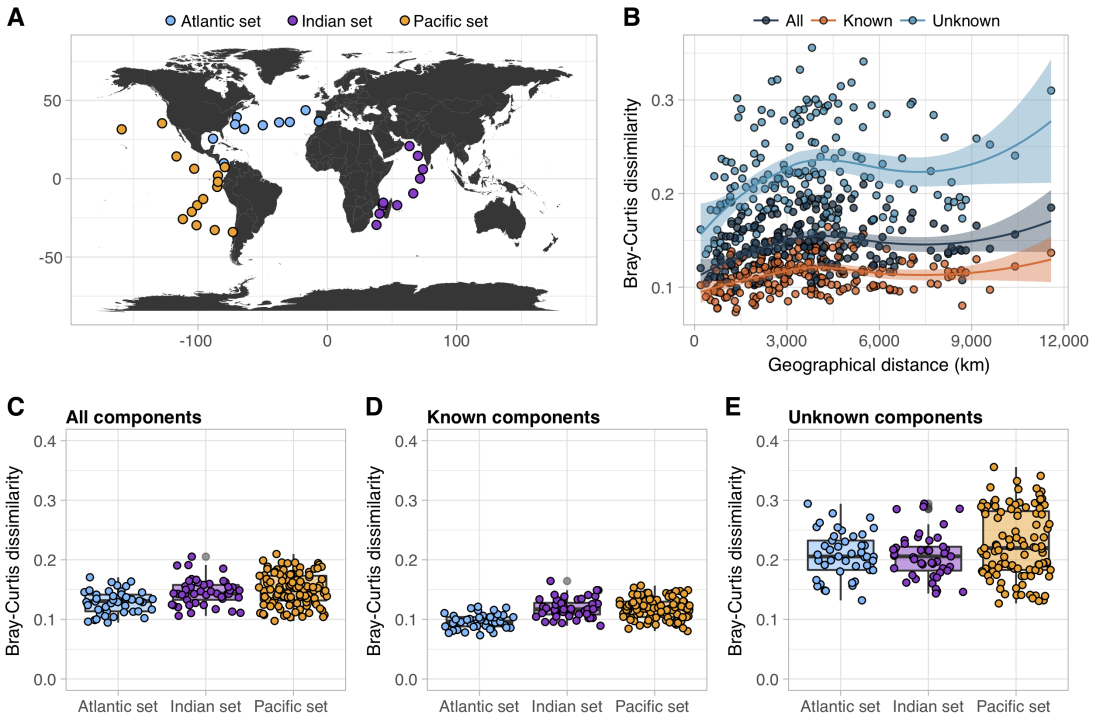

Figure 3.6: **Sample sites beta-diversity vs. geographic distance.** A) Map of TARA Ocean samples used for distance-decay analysis. Samples were divided into regions: Pacific, Atlantic, and Indian. B) Combined distance-decay of samples between all regions. Sites appear three times defined by the Unknowns (EU + GU), Knowns (K + Kwp), and All. The x-axis denotes the geographical distance between sampling sites, while the y-axis denotes the Bray-Curtis dissimilarity beta diversity measurement. (C-E) Box-plots of the pairwise Bray-Curtis dissimilarity beta diversity measurement for each component category, separated by ocean region.

On the other hand, the larger Known matrix could have more noise due to its increased component list and be less sensitive to the Known adaptive protein signal, thus decreasing dissimilarity between sites.

When sites are defined by the Knowns they are the most similar in all ocean regions. This may be explained by the bias that Known components (Knowns with PFAM annotations) are protein sequences that are shared between a large fraction of the microbial community (e.g. housekeeping genes). This was indicated in the niche breadth analysis because around 25% of the knowns had a wide niche breadth score.

The majority of our knowledge of Known microbial functions comes from essential/core gene experiments on cultured microbes that have been studied in the lab (Bernard et al., 2018). Since our knowledge of protein functions are constrained by this, the

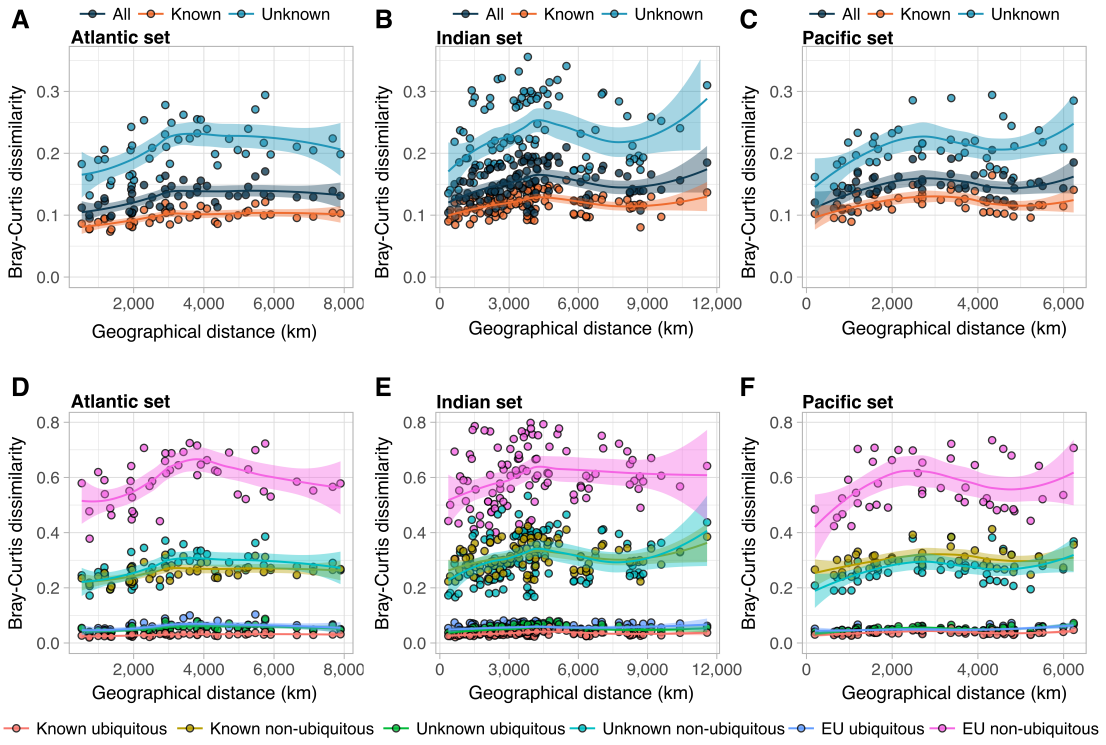

Figure 3.7: **Sample sites beta-diversity vs. geographic distance - ubiquitous and non-ubiquitous components.** (A-C). Distance-decay plot of TARA ocean surface, prokaryotic metagenomes by oceanic region defined in Figure 3.6 B. Sites appear six times defined by the EU (ubiquitous, non-ubiquitous), GU (ubiquitous, non-ubiquitous), and Knowns (ubiquitous, non-ubiquitous). The x-axis denotes the geographical distance between sampling sites, while the y-axis denotes the Bray-Curtis dissimilarity beta diversity measurement. (D-E) Box-plots of the Bray-Curtis dissimilarity beta diversity measurement for each component category.

Table 3.4: Mantel and Partial Mantel test to measure correlation (spearman) between beta-diversity and Haversine distance of TARA Ocean dataset.

	Mantel test		partial-Mantel test	
	ρ	**p-value**	ρ	**p-value**
Unknowns (EUs + GUs)	0.966	0.001	0.512	0.001
Knowns (K + Kwp)	0.968	0.001	0.509	0.001
All combined	0.965	0.001	0.515	0.001

Known fraction may be similar in all environments. Additionally, it has been shown that essential/core genes receive less purifying selection and are more conserved across the domains of life (Jordan et al., 2002). Thus when sites are defined by Known components, dissimilarity is lower due to the overlap of known components between microbes.

On a global scale, the Mantel test showed a significant positive correlation between genetic distance (Bray-Curtis) and geographical distance (Haversine) for the three

component classes (Table 3.4), indicating that closer samples are more similar than the distant. Those results are expected as the geographic isolation is related to the well-defined continental separation of the three oceanic regions used in the analyses: Atlantic, Pacific, and Indian. In all three ocean regions, when sampling sites are defined by All component categories, overall site dissimilarity increases in comparison to the Knowns. This is most likely because including the Unknown fraction in conjunction with the Known fraction increases overall site dissimilarity. The unique, site specific signal from the EUs and GUs may come from the rare biosphere and be environmentally adaptive proteins. This can add site unique component abundances which in turn increase dissimilarity between sites. Additionally, in all three ocean regions, each component category has a similar Bray-Curtis dissimilarity. This results hints that the relationship between the component categories may be a global ocean trend.

The All category represents the utilization of an entire metagenomic sample. In general, functional metagenomic analyses only compare ORFs that have known functions while the rest of the unannotated sequences are discarded. These analyses come in the form of heat maps (McMahon, 2015) or aggregated ordinations (Louca et al., 2016). Here, it is demonstrated that incorporating the whole metagenomic sample into beta diversity measurements is informative and helps to highlight the dissimilarities between the samples.

Separate analyses were then performed on each individual region (Table 3.5). This was done to remove the effect of continental masses distorting the distance decay-analysis, increase resolution of the analysis on the specific regions, and take advantage of the best transects the TARA Ocean sampling regime had to offer.

The Atlantic region had samples that transected the North Atlantic Ocean across the Gulf Stream (Figure 3.7 A, D). In Figure 3.7 A, there is a peak of dissimilarity in all three component categories at around 3,000 km. Increase in sample dissimilarity between 0 and 3,000 km is most likely due to the changing environment from coastal to open water. The dissimilarity then levels off from 3,000 km to 8,000 km. This could be due to the open water samples being in the Gulf stream. This ocean current pulls warm subtropical waters across the Atlantic towards Europe. The current may maintain similar environmental parameters as it crosses the Atlantic thus removing the distance-decay effect on microbial communities.

Table 3.5: Mantel and Partial Mantel test to measure correlation (spearman) between beta-diversity and Haversine distance in local in Atlantic, Pacific, and Indian ocean regions of TARA Oceans.

	Mantel test		partial-Mantel test	
	ρ	**p-value**	ρ	**p-value**
Unknowns (Atlantic)	0.385	0.0201	0.0201	0.0240
Knowns (Atlantic)	0.475	0.0066	0.387	0.0240
All (Atlantic)	0.433	0.0120	0.417	0.0173
Unknowns (Pacific)	0.128	0.2528	0.086	0.3160
Knowns (Pacific)	0.072	0.3384	0.046	0.3818
All (Pacific)	0.111	0.2630	0.076	0.3317
Unknowns (Indian)	0.068	0.3189	0.068	0.3070
Knowns (Indian)	-0.196	0.8849	-0.177	0.8595
All (Indian)	-0.013	0.5159	-0.013	0.5159

The Pacific and Indian sample regions had samples that followed a latitudinal North South transect (Figure 3.7 B, C, E, F). Both distance-decay plots follow a similar trend with an initial positive slope increase in dissimilarity, followed by a trough, then ending with dissimilarity increasing again. Previous findings have shown that environmental variation has more of an effect on functional group composition than dispersal limitation in the open ocean (Louca et al., 2016). Additionally, temperature has been shown to be the main driving force for community composition in surface ocean waters (Sunagawa et al., 2015). These observations are congruent with the curvature found in Pacific region distance-decay plot because similar environments can be found on either side of the equator in low or high latitudes. Latitudes equidistant from the equator receive similar solar radiation due to the curvature of the Earth and thus reflect similar temperatures (e.g. Arctic and Antarctic waters).

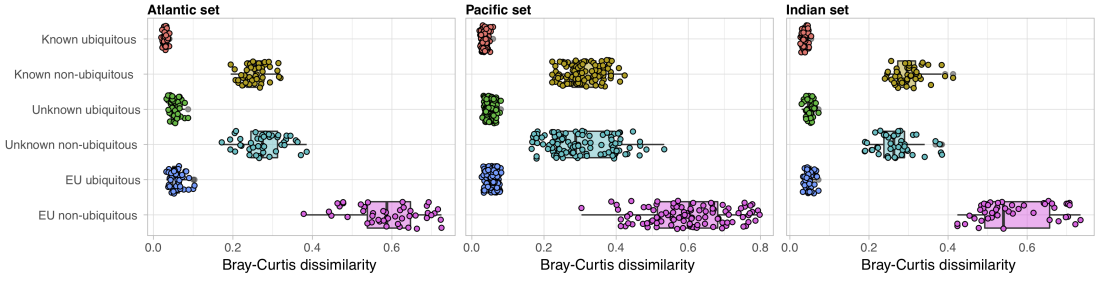

Figure 3.8: **Beta-diveristy of ubiquitous vs non-ubiquitous components.** Box plots representing the distribution of Bray-Curtis dissimilarity between sites in regions defined by Delmont et al. (2017).

Another explanation for the distance-decay curvature that can be applied to both the Indian and Pacific regions is the sampling regimes moving away and back towards the

coasts. Since the sampling paths start and return in similar environments across the gradient of open to coastal waters, component composition dissimilarity increases and decreases, responding to the environment. Though the Atlantic has a similar sampling structure, the Indian and Atlantic have a higher density of coastal samples. Both the Indian and Pacific distance-decay analysis have increases in dissimilarity at the farther geographic distances in the plot. This may be explained by the influence of the coastal environments selecting for unique component composition due to terrestrial runoff.

To further investigate if the Unknowns are site specific and adaptive proteins, component categories were separated and filtered into ubiquitous and non-ubiquitous fractions. The requirement for being ubiquitous was having a presence in every TARA ocean surface sample, anything less was considered non-ubiquitous. In all three ocean regions similar patterns emerge. When sites are defined by non-ubiquitous categories, EU defined sites are significantly more dissimilar than the other non-ubiquitous categories (Figure 3.7). The Unknown and Known non-ubiquitous categories fall within a similar range of dissimilarity. On the other hand, the categories defined by ubiquitous fraction all fall into a very low, similar dissimilarity range of little no distance decay (Figure 3.7).

The EU non-ubiquitous sites were overall the most dissimilar to each other. In light of the results from the ordination, and niche breadth analysis, this was an expected result. EUs have shown to be site specific, and are likely adaptive proteins. Removing the EU ubiquitous fraction that is found everywhere increases the site specific, adaptive signal even more, thus making sites more dissimilar. The non-ubiquitous EU fraction most likely drives the overall dissimilarity between sites when the entire metagenomic sample is taken into account.

The non-ubiquitous components all follow a similar distance-decay trend to their combined component analysis. This makes sense because non-ubiquitous components have a higher chance of being inherently adaptive because they are not found everywhere. In light of this, the non-ubiquitous fraction is most likely driving the distance-decay trends we observe.

When the ubiquitous fraction is removed from the Unknowns and Knowns their beta-diversity dissimilarities converge, and in the case of the Pacific and Indian region, overlap. This overlap in similarity is more than likely due to the similarity of the Kwp and GU component categories. GUs are taxonomically characterized in sequencing databases similar to Kwp but lack a functional annotation. On the scale of functional characterization, GUs and Kwp are the most similar by definition in the Vanni et. al. workflow. Due to their similarity within the design of the

functional categories, they may have similar biogeographical patterns in the oceans. Another explanation is that the non-ubiquitous Knowns provide a more site specific, adaptive signal and drive dissimilarity to be more similar to the non-ubiquitous GUs. The niche breadth analysis showed that the Knowns have a similar distribution of narrow to wide B components. By removing the ubiquitous fraction, the narrow B components increase in relative abundance and thus drive the increase in dissimilarity.

The Unknowns may represent a hybrid of both “core" and environmentally adaptive proteins. The unknown ubiquitous fraction was making the unknown defined sites more dissimilar. There is a possibility that ubiquitous clusters could be involved in niche adaptation. It has been shown that core genes can be adaptive for the sugar metabolism in the metapangenome of *Prochlorococcus* (Delmont & Eren, 2018). Additionally, Delmont & Eren (2018) demonstrated that the accessory pangenome can be made up of environmental core and environmental accessory genes. This was determined by mapping metagenomic reads to the *Prochlorococcus* and seeing in how many genomes the reads occurred. If an ORF is part of the core genome, then there may be a higher possibility that the protein is ubiquitous.

The ubiquitous fraction of every component category all followed a similar trend of defining sites as extremely similar with no significant distance decay. In other words, this group of components are found in every sample of the global regions analyzed and have a similar dissimilarity signal regardless of distance. These results were expected for the Knowns and possibly the GUs, but were surprising to see for the EUs. Due to the biases of protein characterization of essential genes in traditional microbial experiments, Known components contain a lot of housekeeping and essential functions (i.e. shared between all prokaryotes). Thus a strong and similar ubiquitous signal is to be expected. On the other hand, EUs have been shown to have a small niche breadth and be adaptive responses to the environment (as seen in the niche breadth analysis). The signal of ubiquitous EU components, which have no functional annotation or are found in sequenced or draft genomes, is not in line with the previous results in this thesis. This may demonstrate that the EUs share core microbial functions with the other component categories in the world’s oceans. This is a surprising results because it could infer that a set of proteins from a ubiquitous domain of life and or function in the ocean have been left uncharacterized by metagenomics. In total, 6,587 EUs are in this category and were subjected to further investigation.

3.8 Investigating the Ubiquitous EU fraction

To demonstrate that the ubiquitous EU fraction is a legitimate functional signal, we took steps to remove spurious proteins and to detect distant homology and taxonomy (Table 3.6). First we ran antiFAM HMMERs against the potential EU sequences. This removed 250 spurious proteins from the component set. Next, HHblits detected distant homology in 4,823 of the potential EUs with 3,811 having their best hit being a Hypothetical or uncharacterized protein. Finally, for the EUs that were not hit by HHblits we assigned taxonomy using Kaiju, 81 EUs were classified (Table 3.7). This left us with 1,514 potential EUs with no distant homology or taxonomy.

HHblits iteratively compares one HMM against another HMM to detect distant homologies. HMM vs HMM is the most effective technique for detecting distant homologs because it can detect the similarity of the structural templates that usually diverge slower than the sequences themselves. Proteins may remain structurally very similar long after their sequence similarity has been lost. The fact that HHblits was able to detect distant homology in 4,823 EUs indicates that some of the Vanni et. al. EU components may in fact be GUs. This caveat has been solved in the upcoming version of the workflow (Vanni et. al. in prep). HHblits did however fail to detect distant homology in the 1,514 ubiquitous EUs. This was either because they are truly novel, or simply an artifact of sequencing or assembly.

We used Kaiju with the objective to gather taxonomic information of the ubiquitous EUs using a k-mer based approach that in theory is less database biased. The tool takes advantage of the fact that different taxa have different amino acid compositions signatures.

Table 3.6: Exploration of ubiquitous EUs (6,587) taxonomy and homology to databases

Ubiquitous EU classification step	**Number of ubiquitous EUs**
Removed with antiFAM	250
HHblits distant homology detected	4,823
Kaiju taxonomy classified	81
Potential EU	**1,514**

This strategy is unique compared to other more traditional nucleotide methods like Kraken (Wood & Salzberg, 2014). Thus, the Kaiju hits only imply there is a similar k-mer frequency between the 81 Kaiju hits and any of the taxa on NCBI nr database.

Table 3.7: Kaiju taxonomic annotation of ubiquitous potential EUs with no distant homology

Kaiju taxonomic assignment	**Number of ubiquitous EUs**
Bacteria	64
Eukaryota	3
Environmental	14
Total	**81**

Kaiju did not detect taxonomy in all of the ubiquitous EUs. This may have occurred for a few reasons. First, some ubiquitous EUs may be a new expansion of the tree of life, similar to the Candidate Phyla Radiation (Danczak et al., 2017), and thus have a unique k-mer frequency. Second, they are sequencing artifacts and errors, thus have a random, non-related k-mer frequency to known taxa. A future direction for the Vanni et. al. workflow could be to add tool Spurio to the cluster validation steps to identify and remove spurious genes (Höps et al., 2018).

In light of ubiquitous vs non-ubiquitous genes, the discussion of “core" or “essential" microbial functions can be brought up. The current understanding of “core" genomic functions and metabolic pathways is heavily biased by the low number of deeply characterized microbes grown in the lab (Prosser et al., 2014). The core functions found in common in laboratory isolate may not be representative of the total environmental “core". The Ubiquitous EUs may be a new “core" set of genes that have only been detected by means of environmental gene content comparison.

3.9 Mapping EUs to TARA Ocean MAGs

The onset of techniques to extract metagenomic assembled genomes (MAGs) from metagenomic samples has allowed for great insight into microbial populations. Deeper sequencing technology, allowing for 50x of less abundant microbial populations, and efficient differential coverage binning algorithms has led to an era of Metagenomics 2.0 (McMahon, 2015). Additionally, MAGs can now be manually curated in an efficient

manner leading to higher quality (Eren et al., 2015). Although MAGs represent populations of similar microbes and are not complete genomes, the binned contigs are a step closer to the genetic units they came from, genomes. EUs that successfully mapped to the Delmont et al. (2017) MAGs were then upgraded to a new category of the unknown, Population Unknowns.

Population Unknowns (PU)

ORF clusters with an unknown function but map to contigs in MAGs.

To investigate further if the ubiquitous EU components are indeed real proteins found in the environment, we mapped the clusters belonging to the selected components against a set of high quality metagenomic assembled genomes binned from the TARA Ocean prokaryotic dataset (Delmont et al., 2017). Finding ubiquitous EUs in MAGs from the environment they originated from increases the chance of the component being indeed a real environmental protein sequence because it is occurring in the context of a population of similar genomes in the environment.

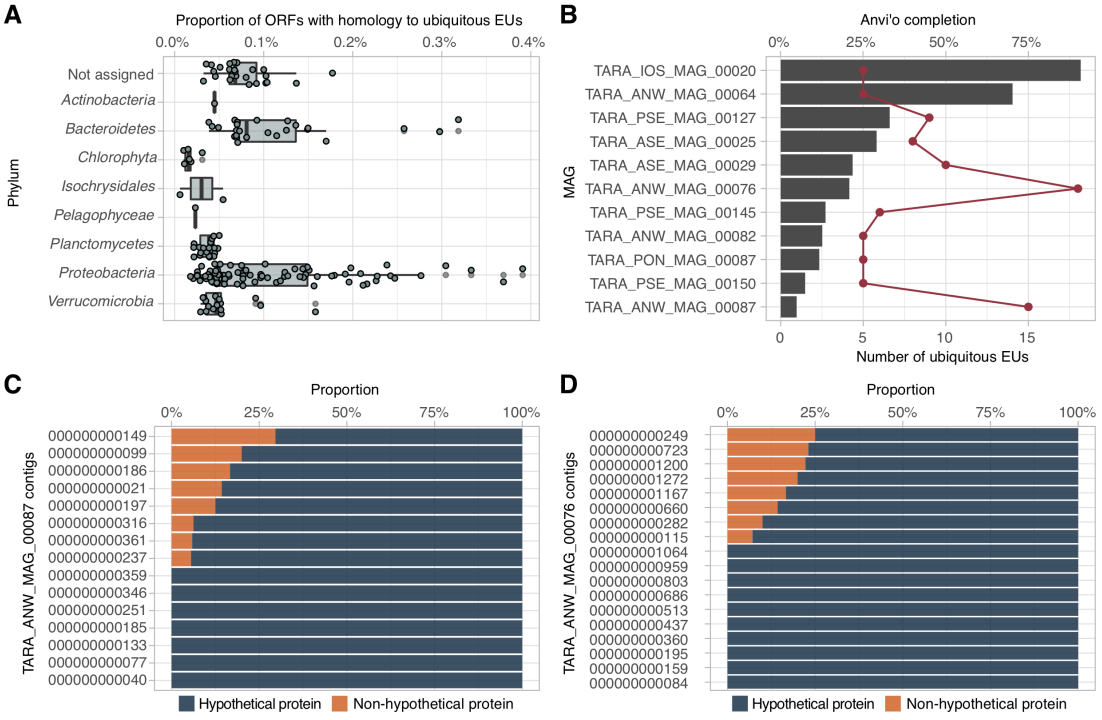

Figure 3.9: **EU mapping results** A) Proportions of ORFS with homology to potential EU within TARA MAGs. X axis is the op hit phylum of potential EU mapping to TARA MAGs. B) Histogram of TARA MAG percent completeness (checkM). Red line represents the number of Potential EUs found in the MAGs. C-D) Contigs from TARA MAGs TARA_ANW_MAG_00076 and TARA_ANW_MAG_00087 in descending order of highest proportion of non-hypothetical ORF content.

Out of the 957 non-redundant MAGs, the ubiquitous EUs mapped to 178 of them. This is a significant amount of the MAGs. If the potential origin of the ubiquitous EUs is from organisms that are part of the rare biosphere, this means that the TARA Ocean project sampled at a critical moment when the rare microbes “bloomed" due to conducive environmental conditions (Gobet et al., 2011). In order to be binned into MAGs, there has to be enough sequencing depth and coverage of the population. Continual deep sequencing of marine waters is needed to catch more of the rare biosphere for metagenomic assembly and expand our knowledge of the world oceans. Our next goal was to examine the genomic neighborhood of the ubiquitous EUs on the contig of which they mapped to. Investigating the genomic neighborhood can lead to the inference of a possible function of the EU. Furthermore, if the EU is surrounded by genes of known function, this adds clarity that the EU is a part of a real contig and possibly an operon.

To select which MAG contig to visualize, we picked the MAGs with the least completeness and highest number of ubiquitous EUs mapped to it (Figure 3.9 B). Next we selected contigs to visualize by choosing the ones with the highest proportion of non-hypothetical proteins (Figure 3.9 C-D). The more annotated proteins on the contig, the more genomic context we can apply to the ubiquitous EU.

TARA ocean MAGS TARA_ANW_MAG_00076 and TARA_ANW_MAG_00087 were selected for contig analysis because they were the most enriched with ubiquitous EUs (Figure 3.9 B). These mags originated from the South West Atlantic Ocean TARA sampling sites. We then picked contigs where the ubiquitous EUs mapped to and arranged them based on content of non-hypothetical genes (Figure 3.9 C-D). This was in order to have a higher chance of having the ubiquitous EU surrounded by genes of known function. If the ubiquitous EU is surrounded by genes of known function, it adds another layer of evidence that the EU is indeed a real sequence. If the EU was spurious, multiple levels of the metagenomic analysis to make the MAG would have had gone wrong including gene prediction, assembly and binning.

In Figure 3.10 A, contig TARA_ANW_MAG_00087_000000000149 is shown. Highlighted in red, is the predicted ORF with significant homology Environmental Unknown clusters (17210924 and 17482518) members of the ubiquitous EU component eu_comp_1 (The EU was blasted against NCBI nt and NCBI genomes to see if there was any significant match to nucleotide sequences in databases). Within its genomic neighborhood are genes relating to DNA and plasmid replication and repair including

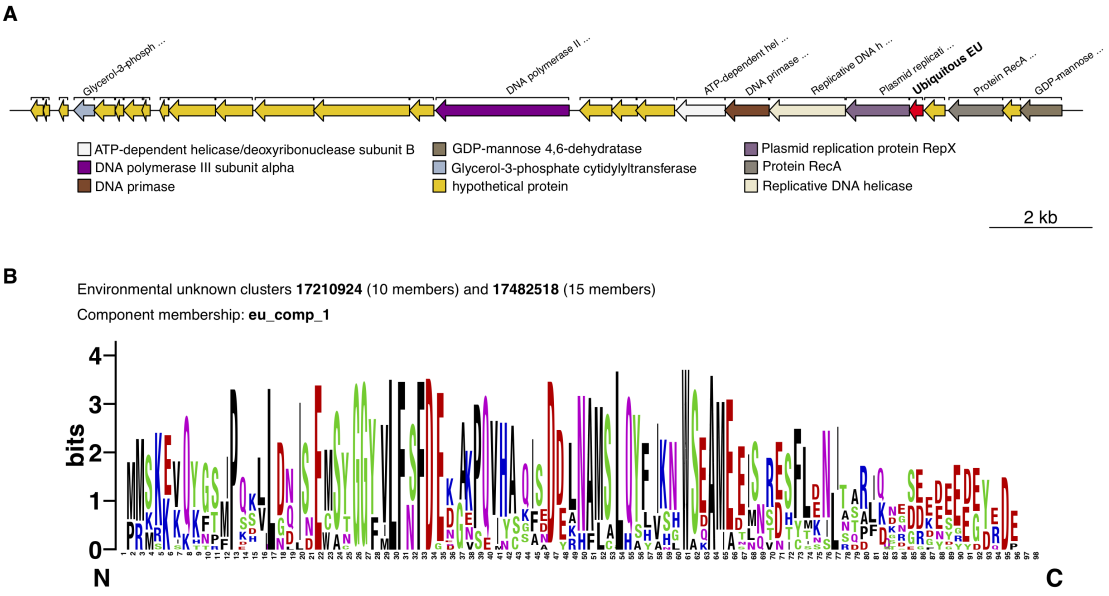

Figure 3.10: **EU gene visualization.** A) Contig genomic neighborhood around a potential EU. B) Conserved consensus sequence logo of of eu_com_1.

DNA polymerase II, DNA primase, and protein RecA. Gene placement in prokaryotic genomes is not random. Genes are grouped together to increase transcriptional efficiency to respond to stimuli in the environment. The lac operon is an example of a group of genes that are transcribed together to metabolize lactose. Since this ubiquitous EU is associated with these DNA and plasmid repair genes, we can hypothesize that it has a related function. Another aspect to consider is that the surrounding genes are essential/core functions for microbial life. Without DNA repair mechanisms, genetic integrity of species would not be able to be maintained. Considering that the ubiquitous EU is surrounded by essential “ubiquitous" biological functions and is found in every sample in the TARA Ocean dataset, there is a higher chance this is a real protein. Furthermore, the consensus amino acid residues after aligning all sequences from both clusters belonging to the component eu_comp_1 (Figure 3.10B) contain two active residues (E, D), one of the conditions defined to identify non-spurious ORFs by Li et al. (2008), supporting one more time that the ORF might be real.

The ubiquitous EUs that were not found in the MAG data set or detected for distant homology have the potential of being part of the genomic repertoire of members present in the rare biosphere. There is mounting evidence that the theory in microbial ecology that “everything is everywhere" but is selected for by the environment is indeed true (Gibbons et al., 2013). All ocean samples may contain a “microbial seed bank" where many rare taxa exist in stationary phase or grow extremely slowly (Pedrós-Alió, 2006). This theory has mainly been investigated via 16S rDNA amplicon sequencing because

the scientific question was focused on phylogenetic diversity. With deeper sequencing, immense diversity of low abundance microbes have been recovered. New protein families have been detected before via mass sequencing efforts and protein clustering (Yooseph et al., 2007). The difference made in this analysis was the Unknown fraction was put in a functional biogeographical context. Due to the deep metagenomic sequencing of the TARA ocean dataset, this maybe the first time that “core" functions from the rare biosphere have been revealed. Due to the Vanni et. al. Unknown cluster categories, the unannotated protein diversity of the TARA ocean data set was able to be accounted for.

Conclusion and Outlook

In this thesis, we have shown that including the entire metagenomic sample in functional biogeography is key to resolve difference between sampling sites. Regardless if a protein cluster has an annotation, it can be accounted for in samples and help define minute differences between geographical sites. By defining a site by their complete functional repertoire, unique insights were generated about the TARA Ocean Southern Ocean sampling sites not observed in the original TARA analysis (Sunagawa et al., 2015). The Vanni et. al. 2018 clusters lay the groundwork for accounting for the entire metagenomic sample and has many future applications in functional microbiology research.

Unknowns as indicators

One direction for immediate application could be to include the Unknowns in objectives for genomic observatories. Since EUs and GUs were associated with small niche occupancy and are selected due to specific environmental factors, their abundances may be good indicators for environmental pollution. Microbial populations respond fast to environmental changes and thus are good indicators for ecosystem health and stress (Buttigieg et al., 2018). Genomic observatories monitor this phenomena to determine ecosystem health (Davies et al., 2012). With the immense influx of high-throughput next-generation sequencing (NGS) data, genomic observatories can effectively monitor ecosystem changes using predictive strategies. Buttigieg et al. (2018) makes the case that a deeper understanding of ocean microbial interactions and functions will improve biomonitoring. It was recently shown that machine learning approaches can effectively analyze differences in ocean microbial taxa can effectively predict anthropogenic impacts (Cordier et al., 2017). A future direction for the Unknowns could be to train Random Forest (RF) algorithms to predict the anthropogenic impacts on the environment and find reliable indicators for environmental monitoring. Including all predicted ORFs in a

sample not only adds information to the site, but also is a better return on investment of the sequencing effort itself. Millions taxpayer dollars and euros have been spent on large environmental sequencing efforts. It is a waste of resources that in most cases not all information for the DNA samples is gleaned and the unannotated fraction is discarded.

A general theme in this thesis was the application of traditional ecological methods to analyze the distribution of the Unknown fraction. Through Levin's Niche Breadth and biogeographical analysis we were able to show unique insights about unknown ubiquitous functions in the ocean. Multidisciplinary approaches may be the key to uncovering more insights about functional biogeography in the world's oceans. There is already dialogue discussing how to apply traditional population genetics methods to the core and accessory genes of the pangenome. Another future direction for the research into the Unknown fraction is to examine their distribution in pangenomes. Research questions can be addressed such as are the Unknowns enriched in the core genome, are Unknowns associated with different phylotypes?

Revealing the function of unknowns

If 99% of microbes are currently uncultivable in the laboratory (Barer, 2015), it can be assumed that many functions will remain unknown until more are isolated and cultured. Computational approaches to marine microbiology have their limitations. Regardless if sequencing technology continues to increase in sequencing depth and quality, gene prediction, assembly, and binning algorithms will never be perfect. NGS data will always be subjected to sequencing and post-processing artifacts. Bioinformatics provides a great platform for hypothesis generation to select and search for important Unknowns, but to truly characterize the Unknown, laboratory experiments on cultured isolates are an absolute necessity. If the Ubiquitous EU component signal is indeed coming from the rare biosphere many innovations will be needed to create cutting edge culturing methods to grow these bugs in the laboratory. The potential of new microbial functions in the rare biosphere is immense and worth investigating due to the untapped resources for biotechnology, environmental applications, and fundamental knowledge.

Another step to investigate the ubiquitous EU would be to express them in microbial vectors and test their function. Many methods are developing to efficiently screen metagenomic libraries, select contigs, and express them in vectors (Leis et al., 2013).

Once expressed, novel genes can be quickly tested for their functions in assays.

Similar to how Wyman et al. (2017) created a list of most wanted FUnkFams, this ecological analysis of the Unknown component fraction is a starting point for targeted environmental protein analysis. Analyzing the patterns and distribution of the Unknown from an ecological perspective can lead to more insights about microbial functions. Additionally, the more functions that are uncovered, the more predictive power genomic observatories may have to detect minute changes in the environment.

It has been suggested that genes of hypothetical function that are conserved throughout phyla should be prioritized for characterization(Galperin, 2004, Hanson et al., 2010). The “core" set of genes have the potential to uncover new phylogenetic markers and essential functions for the definition of life on this planet. This philosophy should be extended to environmentally conserved protein clusters, in particular the Environmental Unknown Ubiquitous clusters.

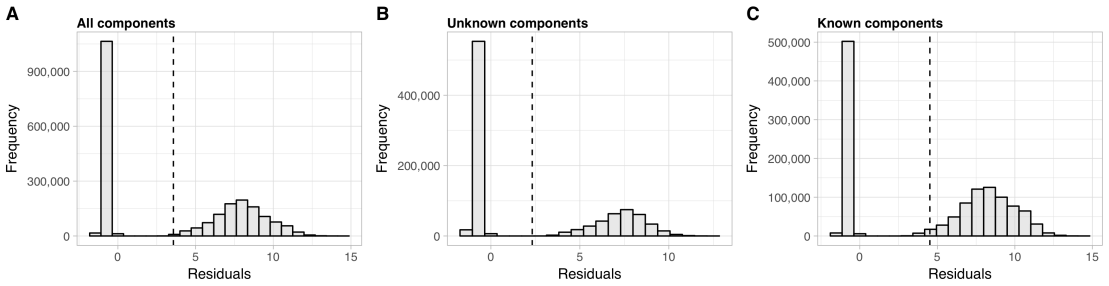

Figure A.1: **TARA Oceans PCA residuals histogram.** A) Residuals of PCA with sites defined by all components. B) Residuals of PCA with sites defined by at the unknown components (EUs + GUs). C) Residuals of PCA with sites defined by the Known components (K + Kwp).

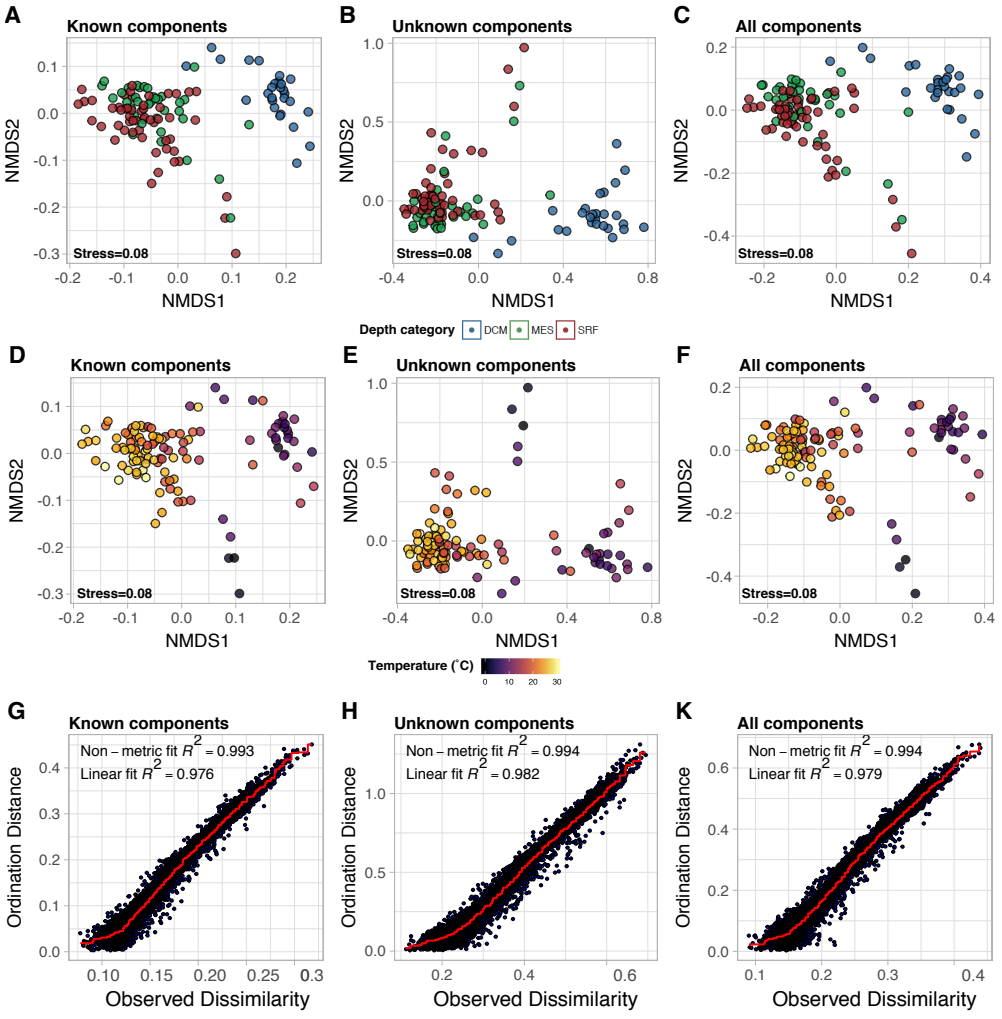

Figure A.2: **TARA Oceans NMDS and Shepard Plots.** A-C) TARA ocean metagenomes defined by (A) Knowns (K + Kwp), (B) Unknowns (EUs + GUs), and (C) ALL components. Sites are colored by depth of origin (Red = surface, Green = deep chlorophyll maximum, and Blue = mesopelagic. D-E) Same plots as A-C but colored by temperature. The darker the color, the colder the sea water temperature. G-I) Shepard plots for NMDS in A-F

Bibliography

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403–410. URL https://doi.org/10.1016/s0022-2836(05)80360-2.

Azam, F. & Malfatti, F. (2007). Microbial structuring of marine ecosystems. *Nature Reviews Microbiology*, **5**(10), 782–791. URL https://doi.org/10.1038/nrmicro1747.

Barer, M. R. (2015). Bacterial Growth, Culturability and Viability. In *Molecular Medical Microbiology*, 181–199. Elsevier. URL https://doi.org/10.1016%2Fb978-0-12-397169-2.00010-x.

Barrientos-Somarribas, M., Messina, D. N., Pou, C., Lysholm, F., Bjerkner, A., Allander, T., Andersson, B. & Sonnhammer, E. L. L. (2018). Discovering viral genomes in human metagenomic data by predicting unknown protein families. *Scientific Reports*, **8**(1). URL https://doi.org/10.1038%2Fs41598-017-18341-7.

Bernard, G., Pathmanathan, J. S., Lannes, R., Lopez, P. & Bapteste, E. (2018). Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. *Genome Biology and Evolution*, **10**(3), 707–715. URL https://doi.org/10.1093%2Fgbe%2Fevy031.

Buchfink, B., Xie, C. & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, **12**(1), 59–60. URL https://doi.org/10.1038%2Fnmeth.3176.

Buttigieg, P. L., Fadeev, E., Bienhold, C., Hehemann, L., Offre, P. & Boetius, A. (2018). Marine microbes in 4D — using time series observation to assess the dynamics of the ocean microbiome and its links to ocean health. *Current Opinion in Microbiology*, **43**, 169–185. URL https://doi.org/10.1016%2Fj.mib.2018.01.015.

Buttigieg, P. L., Hankeln, W., Kostadinov, I., Kottmann, R., Yilmaz, P., Duhaime, M. B. & Glöckner, F. O. (2013). Ecogenomic Perspectives on Domains

of Unknown Function: Correlation-Based Exploration of Marine Metagenomes. *PLoS ONE*, **8**(3), e50869. URL https://doi.org/10.1371%2Fjournal.pone.0050869.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. (2009). BLAST: architecture and applications. *BMC Bioinformatics*, **10**(1), 421. URL https://doi.org/10.1186%2F1471-2105-10-421.

Carradec, Q., , Pelletier, E., Silva, C. D., Alberti, A. et al. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, **9**(1). URL https://doi.org/10.1038%2Fs41467-017-02342-1.

Cavicchioli, R. (2015). Microbial ecology of Antarctic aquatic systems. *Nature Reviews Microbiology*, **13**(11), 691–706. URL https://doi.org/10.1038%2Fnrmicro3549.

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T. & Pawlowski, J. (2017). Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. *Environmental Science & Technology*, **51**(16), 9118–9126. URL https://doi.org/10.1021%2Facs.est.7b01518.

Crooks, G. E. (2004). WebLogo: A Sequence Logo Generator. *Genome Research*, **14**(6), 1188–1190. URL https://doi.org/10.1101%2Fgr.849004.

Danczak, R. E., Johnston, M. D., Kenah, C., Slattery, M., Wrighton, K. C. & Wilkins, M. J. (2017). Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. *Microbiome*, **5**(1). URL https://doi.org/10.1186%2Fs40168-017-0331-1.

Davies, N., Field, D. & Network, T. G. O. (2012). A genomic network to monitor Earth. *Nature*, **481**(7380), 145–145. URL https://doi.org/10.1038%2F481145a.

Delmont, T. O. & Eren, A. M. (2018). Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ*, **6**, e4320. URL https://doi.org/10.7717%2Fpeerj.4320.

Delmont, T. O., Quince, C., Shaiber, A., Esen, O. C., Lee, S. T. M., Lucker, S. & Eren, A. M. (2017). Nitrogen-Fixing Populations Of Planctomycetes And Proteobacteria Are Abundant In The Surface Ocean. *PeerJ*. URL https://doi.org/10.1101%2F129791.

Deorowicz, S., Debudaj-Grabysz, A. & Gudyś, A. (2016). FAMSA: Fast and accurate multiple sequence alignment of huge protein families. *Scientific Reports*, **6**(1). URL https://doi.org/10.1038%2Fsrep33964.

Duarte, C. M. (2015). Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation Expedition. *Limnology and Oceanography Bulletin*, **24**(1), 11–14. URL https://doi.org/10.1002%2Flob.10008.

Eberhardt, R. Y., Haft, D. H., Punta, M., Martin, M., O'Donovan, C. & Bateman, A. (2012). AntiFam: a tool to help identify spurious ORFs in protein annotation. *Database*, **2012**(0), bas003–bas003. URL https://doi.org/10.1093%2Fdatabase%2Fbas003.

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**(9), 755–763. URL https://doi.org/10.1093%2Fbioinformatics%2F14.9.755.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**(5), 1792–1797. URL https://doi.org/10.1093%2Fnar%2Fgkh340.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**(19), 2460–2461. URL https://doi.org/10.1093%2Fbioinformatics%2Fbtq461.

Ellens, K. W., Christian, N., Singh, C., Satagopam, V. P., May, P. & Linster, C. L. (2017). Confronting the catalytic dark matter encoded by sequenced genomes. *Nucleic Acids Research*, **45**(20), 11495–11514. URL https://doi.org/10.1093%2Fnar%2Fgkx937.

Enright, A. J., Van Dongen, S. & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**(7), 1575–1584.

Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L. & Delmont, T. O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, **3**, e1319. URL https://doi.org/10.7717%2Fpeerj.1319.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J. & Bateman, A. (2015). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, **44**(D1), D279–D285. URL https://doi.org/10.1093%2Fnar%2Fgkv1344.

Fischer, D. & Eisenberg, D. (1999). Finding families for genomic ORFans. *Bioinformatics*, **15**(9), 759–762. URL https://doi.org/10.1093%2Fbioinformatics%2F15.9.759.

Galperin, M. Y. (2004). 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Research*, **32**(18), 5452–5463. URL https://doi.org/10.1093%2Fnar%2Fgkh885.

Gibbons, S. M., Caporaso, J. G., Pirrung, M., Field, D., Knight, R. & Gilbert, J. A. (2013). Evidence for a persistent microbial seed bank throughout the global ocean. *Proceedings of the National Academy of Sciences*, **110**(12), 4651–4655. URL https://doi.org/10.1073%2Fpnas.1217767110.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, **8**. URL https://doi.org/10.3389%2Ffmicb.2017.02224.

Gobet, A., Böer, S. I., Huse, S. M., van Beusekom, J. E. E., Quince, C., Sogin, M. L., Boetius, A. & Ramette, A. (2011). Diversity and dynamics of rare and of resident bacterial populations in coastal sands. *The ISME Journal*, **6**(3), 542–553. URL https://doi.org/10.1038%2Fismej.2011.132.

Guy, L., Kultima, J. R. & Andersson, S. G. E. (2010). genoPlotR: comparative gene and genome visualization in R. *Bioinformatics*, **26**(18), 2334–2335. URL https://doi.org/10.1093%2Fbioinformatics%2Fbtq413.

Hanson, A. D., Pribat, A., Waller, J. C. & de Crécy-Lagard, V. (2010). 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list – and how to find it. *Biochemical Journal*, **425**(1), 1–11. URL https://doi.org/10.1042%2Fbj20091328.

Hijmans, R. J. (2017). *geosphere: Spherical Trigonometry*. URL https://CRAN.R-project.org/package=geosphere. R package version 1.5-7.

Höps, W., Jeffryes, M. & Bateman, A. (2018). Gene Unprediction with Spurio: A tool to identify spurious protein sequences. *F1000Research*, **7**, 261. URL https://doi.org/10.12688%2Ff1000research.14050.1.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C. & Bork, P. (2015). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, **44**(D1), D286–D293. URL https://doi.org/10.1093%2Fnar%2Fgkv1248.

Jaroszewski, L., Li, Z., Krishna, S. S., Bakolitsa, C., Wooley, J., Deacon, A. M., Wilson, I. A. & Godzik, A. (2009). Exploration of Uncharted Regions of the Protein Universe. *PLoS Biology*, **7**(9), e1000205. URL https://doi.org/10.1371%2Fjournal.pbio.1000205.

Jordan, I. K., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. (2002). Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. *Genome Research*, **12**(6), 962–968. URL https://doi.org/10.1101%2Fgr.87702.

Kassambara, A. (2017). *ggpubr: 'ggplot2' Based Publication Ready Plots*. URL https://CRAN.R-project.org/package=ggpubr. R package version 0.1.6.

Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I. et al. (2015). The ocean sampling day consortium. *GigaScience*, **4**(1). URL https://doi.org/10.1186%2Fs13742-015-0066-5.

Lahti, L., Shetty, S., Blake, T. & Salojarvi, J. (2012-2017). microbiome R package.

Leis, B., Angelov, A. & Liebl, W. (2013). Screening and Expression of Genes from Metagenomes. In *Advances in Applied Microbiology*, 1–68. Elsevier. URL https://doi.org/10.1016%2Fb978-0-12-407678-5.00001-5.

Levins, R. (1966). The strategy of model building in population ecology. *American Scientist*, **54**(4), 421–431. URL http://www.jstor.org/stable/27836590?seq=1#page_scan_tab_contents.

Li, W. & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**(13), 1658–1659. URL https://doi.org/10.1093%2Fbioinformatics%2Fbtl158.

Li, W., Wooley, J. C. & Godzik, A. (2008). Probing Metagenomics by Rapid Cluster Analysis of Very Large Datasets. *PLoS ONE*, **3**(10), e3375. URL https://doi.org/10.1371%2Fjournal.pone.0003375.

Logares, R., Lindström, E. S., Langenheder, S., Logue, J. B., Paterson, H., Laybourn-Parry, J., Rengefors, K., Tranvik, L. & Bertilsson, S. (2012). Biogeography of bacterial communities exposed to progressive long-term environmental change. *The ISME Journal*, **7**(5), 937–948. URL https://doi.org/10.1038%2Fismej.2012.168.

Louca, S., Parfrey, L. W. & Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science*, **353**(6305), 1272–1277. URL https://doi.org/10.1126%2Fscience.aaf4507.

Marcy, Y., Ouverney, C., Bik, E. M., Losekann, T., Ivanova, N., Martin, H. G., Szeto, E., Platt, D., Hugenholtz, P., Relman, D. A. & Quake, S. R. (2007). Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proceedings of the*

National Academy of Sciences, **104**(29), 11889–11894. URL https://doi.org/10.1073/pnas.0704662104.

Martin, J. H., Gordon, R. M. & Fitzwater, S. E. (1990). Iron in Antarctic waters. *Nature*, **345**(6271), 156–158. URL https://doi.org/10.1038%2F345156a0.

McInerney, J. O., McNally, A. & O'Connell, M. J. (2017). Why prokaryotes have pangenomes. *Nature Microbiology*, **2**(4), 17040. URL https://doi.org/10.1038/nmicrobiol.2017.40.

McMahon, K. (2015). 'Metagenomics 2.0'. *Environmental Microbiology Reports*, **7**(1), 38–39. URL https://doi.org/10.1111%2F1758-2229.12253.

McMurdie, P. J. & Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, **8**(4), e61217. URL http://dx.plos.org/10.1371/journal.pone.0061217.

Menzel, P., Ng, K. L. & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, **7**, 11257. URL https://doi.org/10.1038%2Fncomms11257.

Miklós, I. & Podani, J. (2004). Randomization of presence–absence matrices: comments and new algorithms. *Ecology*, **85**(1), 86–92. URL https://doi.org/10.1890%2F03-0101.

Milici, M., Vital, M., Tomasch, J., Badewien, T. H., Giebel, H.-A., Plumeier, I., Wang, H., Pieper, D. H., Wagner-Döbler, I. & Simon, M. (2017). Diversity and community composition of particle-associated and free-living bacteria in mesopelagic and bathypelagic Southern Ocean water masses: Evidence of dispersal limitation in the Bransfield Strait. *Limnology and Oceanography*, **62**(3), 1080–1095. URL https://doi.org/10.1002%2Flno.10487.

Mirdita, M., von den Driesch, L., Galiez, C., Martin, M. J., Söding, J. & Steinegger, M. (2016). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research*, **45**(D1), D170–D176. URL https://doi.org/10.1093%2Fnar%2Fgkw1081.

Mudgal, R., Sandhya, S., Chandra, N. & Srinivasan, N. (2015). De-DUFing the DUFs: Deciphering distant evolutionary relationships of Domains of Unknown Function using sensitive homology detection methods. *Biology Direct*, **10**(1). URL https://doi.org/10.1186%2Fs13062-015-0069-2.

Nealson, K. H. & Venter, J. C. (2007). Metagenomics and the global ocean survey: what's in it for us and why should we care? *The ISME Journal*, **1**(3), 185–187. URL https://doi.org/10.1038%2Fismej.2007.43.

Nekrutenko, A. (2001). The KA/KS Ratio Test for Assessing the Protein-Coding Potential of Genomic Regions: An Empirical and Simulation Study. *Genome Research*, **12**(1), 198–202. URL https://doi.org/10.1101%2Fgr.200901.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E. & Wagner, H. (2017). *vegan: Community Ecology Package*. URL https://CRAN.R-project.org/package=vegan. R package version 2.4-5.

Parks, D. H., MacDonald, N. J. & Beiko, R. G. (2011). Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinformatics*, **12**(1), 328. URL https://doi.org/10.1186%2F1471-2105-12-328.

Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, **10**(12), 1200–1202. URL https://doi.org/10.1038%2Fnmeth.2658.

Pavlopoulos, G. A. (2017). How to Cluster Protein Sequences: Tools, Tips and Commands. *MOJ Proteomics & Bioinformatics*, **5**(5). URL https://doi.org/10.15406%2Fmojpb.2017.05.00174.

Pedrós-Alió, C. (2006). Marine microbial diversity: can it be determined? *Trends in Microbiology*, **14**(6), 257–263. URL https://doi.org/10.1016%2Fj.tim.2006.04.007.

Perdigão, N., Rosa, A. C. & O'Donoghue, S. I. (2017). The Dark Proteome Database. *BioData Mining*, **10**(1). URL https://doi.org/10.1186%2Fs13040-017-0144-6.

Piepel, G. F. & Aitchison, J. (1988). The Statistical Analysis of Compositional Data. *Technometrics*, **30**(1), 120. URL https://doi.org/10.2307%2F1270335.

Prosser, G. A., Larrouy-Maumus, G. & de Carvalho, L. P. S. (2014). Metabolomic strategies for the identification of new enzyme functions and metabolic pathways. *EMBO reports*. URL https://doi.org/10.15252%2Fembr.201338283.

R Development Core Team. (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org. ISBN 3-900051-07-0.

Remmert, M., Biegert, A., Hauser, A. & Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, **9**(2), 173–175. URL https://doi.org/10.1038%2Fnmeth.1818.

Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. (2015). Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife*, **4**. URL https://doi.org/10.7554%2Felife.08490.

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S. et al. (2007). The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology*, **5**(3), e77. URL https://doi.org/10.1371%2Fjournal.pbio.0050077.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**(14), 2068–2069. URL https://doi.org/10.1093%2Fbioinformatics%2Fbtu153.

Shapiro, B. J. (2017). The population genetics of pangenomes. *Nature Microbiology*, **2**(12), 1574–1574. URL https://doi.org/10.1038%2Fs41564-017-0066-6.

Sharpton, T. J., Jospin, G., Wu, D., Langille, M. G., Pollard, K. S. & Eisen, J. A. (2012). Sifting through genomes with iterative-sequence clustering produces a large, phylogenetically diverse protein-family resource. *BMC Bioinformatics*, **13**(1), 264. URL https://doi.org/10.1186%2F1471-2105-13-264.

States, D. J. & Boguski, M. S. (1991). Similarity and Homology. In *Sequence Analysis Primer*, 89–157. Palgrave Macmillan UK. URL https://doi.org/10.1007%2F978-1-349-21355-9_3.

Steinegger, M. & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*. URL https://doi.org/10.1038%2Fnbt.3988.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K. et al. (2015). Structure and function of the global ocean microbiome. *Science*, **348**(6237), 1261359–1261359. URL https://doi.org/10.1126%2Fscience.1261359.

Tagliabue, A., Sallée, J.-B., Bowie, A. R., Lévy, M., Swart, S. & Boyd, P. W. (2014). Surface-water iron supplies in the Southern Ocean sustained by deep winter mixing. *Nature Geoscience*, **7**(4), 314–320. URL https://doi.org/10.1038%2Fngeo2101.

Tatusov, R. L. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, **28**(1), 33–36. URL https://doi.org/10.1093%2Fnar%2F28.1.33.

Venter, J. C. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, **304**(5667), 66–74. URL https://doi.org/10.1126%2Fscience.1093857.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4. URL http://ggplot2.org.

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. URL https://CRAN.R-project.org/package=tidyverse. R package version 1.2.1.

Wilkins, D., Yau, S., Williams, T. J., Allen, M. A., Brown, M. V., DeMaere, M. Z., Lauro, F. M. & Cavicchioli, R. (2013). Key microbial drivers in Antarctic aquatic environments. *FEMS Microbiology Reviews*, **37**(3), 303–335. URL https://doi.org/10.1111%2F1574-6976.12007.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**, 160018. URL https://doi.org/10.1038%2Fsdata.2016.18.

Wood, D. E. & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, **15**(3), R46. URL https://doi.org/10.1186%2Fgb-2014-15-3-r46.

Wyman, S. K., Avila-Herrera, A., Nayfach, S. & Pollard, K. S. (2017). A most wanted list of conserved protein families with no known domains. URL https://doi.org/10.1101%2F207985.

Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J. et al. (2007). The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families. *PLoS Biology*, **5**(3), e16. URL https://doi.org/10.1371%2Fjournal.pbio.0050016.

Acknowledgements

To begin, I would like to thank my supervisor Antonio who laid the ground work for me to become a successful PhD student. He has instilled critical thinking and methodical philosophy into my workflow. Where ever I end up I will truly benefit from these last 6 months with him as my supervisor.

A HUGE thanks to Chiara for all the support and not getting tired of me asking questions and barging into her office.

Next, I would like to thank my two reviewers Frank Oliver and Pier. Thank you for your time and effort for providing me with constructive criticism on this thesis.

Another thank you to Frank Oliver for allowing me to join the MGG group. I truly enjoyed working with this group and thank you to the whole MGG group for the support!

A separate thank you to Pier who provided with amazing explanations and guidance on my multivariate statistics.

Thank you Henny for being my MSc partner in crime in MGG.

Thank you to my office mate Marcus Klimmek for all the laughs and drastically improving my German vocabulary in the most important ways.

Thank you to Gerard for working next to me during the beginning. It was fun learning Tidyverse with you!

Thank you MarMic class of 2021 for being an inspirational group of young scientists. I enjoyed learning and growing with you guys and cannot wait see how your careers excel!

Thank you to my coffee amigos Cora, David, and Andrea for all the great conversations and laughs!

Thank you to Christiane and Karl-Heinz for all the MarMic support.

Last but not least, many thanks to my Mother, Father, and Sister who encouraged me all along the way from the other side of the world.

Erklärung

Statement

Hiermit versichere ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

I herewith confirm that I have written this thesis unaided and that I used no other resources than those mentioned.

_____ _____

(Ort und Datum / Place and Date) (Unterschrift / Signature)