UNRAVELING THE ECOLOGICAL RELEVANCE OF MICROBIAL DARK MATTER

By

TATYANA ZAMKOVAYA

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2020

To my loved ones

ACKNOWLEDGEMENTS

I'd first like to thank my committee chair Dr. Ana Conesa for her unwavering support, even across continents and time zones, and my committee members Dr. Jamie Foster, Dr. Bryan Kolaczkowski, Dr. George Michailidis and Dr. Valerie de Crecy-Lagard for their encouragement and mentorship throughout my time here at the University of Florida. I'd also like to thank Alex Moskalenko and the HiPerGator team for their invaluable advice and Guillem and Juan Cruz for their bioinformatics expertise and support. I'm most grateful for the support system I've had throughout my graduate studies, particularly the support and constant inspiration I experienced from my incredibly talented, kind, and fun lab group. I thank my labmates Leandro and Raymond for teaching me patience and Venezuelan slang, Rocio for her wonderful advice, and my Spanish labmates Salva, Paco, Pedro, Angeles, Manu, Fran, and Teresa for impromptu Spanish lessons, cross-continental support, and great adventures together. I consider you all not just my labmates and coworkers but great friends and my extended Spanish family. I'd also like to thank all the friends I've made throughout my studies at the University of Florida, particularly Jessi,Addi,and Natalie for believing in me and supporting me at every step with lots of coffee, boba tea, and moral encouragement. Surviving the required microbiology coursework, the PhD, and life in general wouldn't have been possible without you all. Last but not least, I thank my human and poodle family for their unconditional love. To my chosen sisters, Tijana and Alina, and my krishna family, Guillem and Ola, thank you for believing in me more than I've believed in myself, listening to countless presentations, and for being such amazing friends. To my beloved grandma, Nina, thank you for your positive words and videos of encouragement. Your pride and endless love for me has helped me get through even the most trying of days. Mom and Dad, this entire journey would not have been possible without you. Thank you for encouraging me to challenge myself and for plying me with delicious dessert at difficult moments. To everyone mentioned here, and to the countless friends and family I didn't mention by name, know that I'm grateful for each and every one of you for your love and support. Thank you all for everything.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figures page

LIST OF OBJECTS

LIST OF ABBREVIATIONS

16S	16S rRNA Amplicon Sequencing
BLAST	Basic Local Alignment Search Tool
COG	Cluster of Orthologous Groups
DS	Deep Sea
EFI-EST	Enzyme Function Initiative-Enzyme Similarity Tool
eggNOG	Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups)
GO	Gene Ontology
HS	Hot Springs
HY	Hypersaline
JGI Gold	Joint Genome Institute Genomes Online Database
KEGG	Kyoto Encyclopedia of Genes and Genomes
OTU	Operational Taxonomic Unit of Diversity
MAG	Metagenome-Assembled Genome
MDM	Microbial Dark Matter
NCBI	National Center for Biotechnology Information
PO	Polar
Prodigal	Prokaryotic Dynamic Programming Genefinding Algorithm
QIIME	Quantitative Insights into Microbial Ecology
SAG	Single-cell Amplified Genome
SpiecEasi	Sparse InversE Covariance Estimation for Ecological Association and Statistical Inference
StARs	Stability Approach to Regularization Section
WGS	Whole Genome Sequencing

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

UNRAVELING THE ECOLOGICAL RELEVANCE OF MICROBIAL DARK MATTER

By

Tatyana Zamkovaya

December 2020

Chair: Ana Conesa
Major: Microbiology and Cell Science

Complete knowledge of the evolution and adaptation of life is greatly hindered by the unknown properties of the vast majority of the earliest known life forms, prokaryotes. Although next-generation sequencing technologies have made detection of these unknown, uncultured species called Microbial Dark Matter (MDM) possible, many obstacles must still be overcome before complete knowledge of these prokaryotes is reached. The majority of microorganisms and their gene products on Earth remain unknown and it is unclear what magnitude of contributions each unknown microbe provides to its environment. The dominance of novel phyla across Earth biomes, especially in extremophilic conditions, suggests that unraveling the role of unknowns is key to uncover the ecological relevance of these organisms to their communities and subsequently unlock the adaptation strategies needed to survive harsh conditions within and even beyond Earth. Accordingly, completion of the puzzle on the origins and evolution of life requires improved understanding of the ecological niches and subsequent functional properties of relevant unknown organisms. We hypothesized that, by their abundance patterns and possession of a particular set of genes and evolutionary traits, MDM are potentially uniquely adapted to extreme environments and possibly highly relevant to community structure and stability of these environments. We developed a computationally intensive network-based approach to unravel the ecological relevance of unknowns and applied this methodology to a large dataset of diverse extreme aquatic environments. The removal of unknown taxa significantly decreased the overall degree and

betweenness values of different extreme aquatic microbial association networks, indicating that unknowns significantly shape the structure of their communities. Unknown organisms frequently appeared as top hubs, suggesting that these microbes are particularly relevant and adapted to their extreme habitats. Novel gene functions associated to these particularly relevant unknown components were identified using a metagenomics and comparative genomics approach. Candidate conserved uncharacterized genes, including domains of unknown function (DUFs), were predicted to have functional properties related to stress and adaptation. The results of this work provide a subset a subset of species and genes for characterization to further clarify the impact of MDM in the formation, function, and evolution of distinct ecosystems.

CHAPTER 1
PROGRESS IN MICROBIAL DARK MATTER DIVERSITY, INTERACTION, AND PHYLOGENY INVESTIGATIONS

1.1 Introduction

Before sequencing of uncultivated microbes was possible, estimates of the total number of species and biodiversity present on Earth were grossly inaccurate [1, 2]. Extreme environments were considered barren of life and many microbial niches were poorly understood. For these reasons, initial environmental studies aimed to simply understand which species were present.

The near universality and stability of the 16S rRNA gene across prokaryotes made 16S rRNA amplicon sequencing the most popular approach early on to detect novel bacterial and archaeal Candidate phyla and divisions across a wide range of conditions, including oceans [3], arid deserts [4] and hypersaline lakes [5, 6]. However, these analyses suffered from amplification bias (depending on the primer set used) and could only provide information on phylogenetic diversity [7], inciting researchers to use other methods to uncover the functional and metabolic properties of these microorganisms. In response to these limitations, approaches based on shotgun metagenomics (also known as whole genome sequencing, WGS), metagenome-assembled genomes (MAG), and/or single-cell amplified genomes (SAG) were developed to unravel both taxonomic and metabolic diversity. A comprehensive whole genome sequencing approach was first applied to the Sargasso Sea [8],resulting in the identification of 148 previously unknown bacterial phylotypes and over 1.2 million previously unknown genes. With this strategy, genomes were able to be reconstructed for many candidate phyla, including Melainabacteria, Microgenomates (OP11), NC10, SR1, and WWE3 [9, 10], providing new metabolic insight into these unknown bacteria. Significantly, through this approach, the Candidate Phyla Radiation (CPR), a superphylum of Bacteria, was unveiled and was found to account for 15 percent of all bacterial biodiversity [11]. The requirement of a well-annotated reference database and the inability to explain the inherent heterogeneity of cells made assembling complete genomes difficult by WGS alone, inciting researchers to use a combined shotgun sequencing and single-cell sequencing approach. As described in more comprehensive reviews [10, 12], using this integrated strategy within extreme environments yielded novel genomic information for

polyextremophilic candidate phyla like Parvarchaeota, Nanohaloarchaeaota, Acetothermia, and Gracilibacteria [12]. Additionally, SAGs of candidate phyla like TM7 (Saccharimonadia) , TM6, and OP9 were found to contain transporters for multi-drug resistance and glycohydrolases [13], highlighting the clinical significance of studying these unknown organisms. To uncover functional and metabolic activity of a given species, metatranscriptomics, metaproteomics, and metametabolomics, which respectively reveal gene expression, protein expression, and metabolite expression, were implemented.

As technologies improved and became less expensive, the exploration of a given microbial community evolved as well. The evaluation of the composition of an individual environment grew to investigation of multiple environments and even global investigations. Yet, composition analyses alone could not explain the existence and wide distribution of not-yet-characterized archaea and bacteria across a diverse range of ecosystems, nor what role these taxa play in relation to well-characterized members of their environmental communities. Truly understanding any bacterial species, particularly those only detectable through high-throughput sequencing, requires consideration of their ecological context [14], by including all other members and factors affecting a community. Consequently, researchers began to evaluate change in abundance upon environmental and abiotic perturbation of a given ecosystem and to investigate the relationship of candidate phyla to other organisms, either through experimental means, co-occurrence relationships, or presence/absence of core survival genes and pathways. Modeling each community as a network enabled researchers to detect and predict relationships of all members and, through network measures, to quantitatively define the most important “keystone” [15, 16] or hub [17, 18] species. Furthermore, longitudinal, inter-domain and integrated omics-studies enabled enriched understanding of how each microbial member of an ecosystem reacts in abundance, gene, protein, and metabolite expression in response to changes in season [19, 20], salinity [21, 22], time [6, 23, 24], or temperature [25, 26, 27, 28].

Early metagenomics analyses revealed that several novel phyla lack key genes previously believed to be ubiquitous among the Bacterial or Archaeal domains whereas others possessed

archaeal or eukaryotic-like metabolites and varied widely in G+C content [29, 30] or size [31] compared to their well-characterized relatives. To better understand these intriguing differences, further phylogenetic evaluation was required. Consequently, phylogenetic analyses of candidate phyla similarly evolved from evaluating the sequence similarity and position of microbes through phylogenetic trees, to draft genome reconstruction of uncharacterized organisms, to metabolic reconstruction and pathway modeling. At each level of evolutionary analysis, shared traits across domains, symbiotic relationships, and horizontal and lateral gene transfer between organisms became evident. Classic theories on the evolution of life have been challenged with these new discoveries and, a plethora of novel lineages [32, 33, 34, 35, 36, 37, 38], functions [39, 40, 41, 42, 43], and adaptation mechanisms [44, 45, 46] have been revealed.

Through the progression of composition, network interaction, and phylogenetic studies (Fig. 1-1), we have come to know more about how unusual, novel microorganisms commonly referred to as Microbial Dark Matter (MDM) [37] exist, thrive, and adapt to a diverse range of conditions. Here, we present a summary of the insights and limitations of current microbial diversity, interaction, and evolutionary trait studies of unknown organisms. We hope that this review provides a comprehensive overview of the progression and challenges of MDM investigations.

1.2 Microbial Diversity and Composition Analysis

To explore microbial diversity, researchers evaluate community composition of 16S, WGS, MAG, SAG, and RNA-Seq data to determine which species are present, in which proportions, and in what patterns, in a given ecosystem. Microbial community composition of raw amplicon, WGS, single-cell genome sequencing data, or metatranscriptomics data is easier than ever to analyze through the help of various bioinformatics tools and pipelines (Fig. 1-2), which provide researchers with a snapshot of the microbes, genes, and active cells present within a microbial community in an efficient and reproducible manner. Including unclassified organisms or those with no cultured representatives in these analyses has greatly expanded our understanding of life on Earth.

1.2.1 Early Diversity Analyses of Individual Environments

Initial biodiversity studies of a single microbial ecosystem demonstrated that all environments, like soils [47], oceans [3, 48], coral reefs [49], and even more extreme, seemingly uninhabitable areas [34, 4, 50, 21, 51], are teeming with novel diversity. In fact, in hypersaline sapropels, environments most closely resembling the ancient, atmospheric conditions of Mars, 32 of the 59 diverse phyla [52] uncovered were MDM. Similarly, the majority of the prokaryotes in methanogenic bioreactors were found to belong to 15 recently discovered bacterial phyla and 3 novel archaeal phyla [53]. Even characteristically understudied environments, like the Antarctic, have recently been shown to contain organisms belonging to novel lineages of bacteria [54, 55], like Dormibacteraeota (AD3), Eremobacteraeota (WPS-2), Patescibacteria, and Abditibacteriota, which may occupy important roles in carbon fixation and primary production in desert soils.

1.2.2 Investigating Microbial Diversity in Multiple Geographic Regions or Environment Types

The inclusion of more and distinct types of environments has helped unveil where, how, and why novel organisms proliferate and has shed light on the distribution and potential ecological niches of these unknown, uncharacterized phyla. For instance, the widespread detection of Aminicenantes [56] and Candidate Phylum Rokubacteria [57] across diverse biomes suggests that unknown phyla have high intraphylum metabolic diversity and versatile adaptation mechanisms for survival. Similarly, a comparison of Sakinaw Lake (Canada) and Etoliko Lagoon (Greece) metabolic enzymes and genes present within SAGs [58] revealed that candidate phylum Latescibacteria (WS3) likely has an important widespread ecological role related to algal biomass degradation. Taken together, these findings demonstrate that unknown organisms are important drivers of geochemical cycles and suggest that the mechanisms these organisms utilize must be investigated further to enrich understanding of Earth itself.

1.2.3 Global Diversity Investigations

While including diverse landscapes has helped increase knowledge of MDM, significant challenges remain that can only be filled with a global approach. Though early efforts in

investigating MDM helped produce more than 8000 reconstructed genomes [59] and helped reveal more than 15 percent of all biodiversity [60, 61, 11], many microbes have yet to be characterized [62] and the gap between sequenced and cultivated organisms continues to grow, requiring studies to be more systematic in nature.

Results from the relatively few global biodiversity explorations conducted on MDM highlight the potential significance of uncultured taxa in ecosystem function and microbial evolution. One major global abundance metagenome and metatranscriptome-based investigation [63] uncovered that phylogenetically novel uncultured genera and phyla dominated within all nonhuman biomes and were also highly active, with uncultured phyla possibly making up one-quarter of the population of all microbial cells on Earth. Other global studies, focusing on genomic and phylogenetic features of MDM taxa, have expanded our definition and understanding of microbial evolution. The three-domain system no longer seems accurate, with the surprising findings of shared inter-domain characteristics, genes, and mechanisms present within these novel organisms. The massive endeavor to explore 200 single-cell genomes from 29 unexplored branches [37] of the tree of life from candidate phyla representatives, through the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project, uncovered archaeal-like metabolic pathways in bacterial species and complete bacterial-like sigma factors within uncultured archaeal taxa, refuting the long-established notion that molecular features are distinct and domain-specific. Another meta-analysis of over 300 bacterial and archaeal genomes assembled from groundwater, deep subsurface, hydrocarbon-impacted, and oceanic environmental metagenomes revealed, by the presence of a novel archaeal-like RuBisCo enzyme in Candidate Phyla Radiation (CPR) bacteria [64], that lateral gene transfer [65] may have occurred between CPR and DPANN (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, and Nanohaloarchaeota) archaea, again suggesting shared phylogenetic and metabolic diversity across the domains of life.

Findings from global studies have also revealed that certain mechanisms and ecological roles, are more widely distributed than first proposed. For example, a more recent global ocean

study on novel marine ultrasmall prokaryote distribution [66] suggests that these organisms may participate in carbon fixation in addition to carbon degradation, suggesting a novel and broader role in carbon cycling for nano-organisms. Meanwhile, an evaluation of the antibiotic resistance (AR) gene composition of over 6000 sequenced microbial genomes [67] and functional metagenomic data revealed significant AR mechanism enrichment by bacterial phyla (Actinobacteria) and habitat (soils), that seems to be driven largely by ecology. Taken together, these results highlight how inclusion of a broader ecological context, among different ecosystem types, can lead to enriched understanding of prokaryotic survival mechanisms and microbial evolution.

1.3 Microbial Interactions

The presence of MDM across a wide range of environments has led to a better understanding of the diverse composition of soils, marine ecosystems, and the human body while the unique characteristics of MDM have revolutionized ideas concerning microbial evolution. Yet, true understanding depends upon consideration of the interactions and consequent effects of other community members and environmental factors. For these reasons, analyses have progressed from simple taxonomic composition to evaluating composition and interaction upon perturbation of biotic and abiotic factors through genomic and physiological information and, more recently, statistical association networks (Fig. 1-3). By this manner, microbe-microbe, microbe-environment, and multi-omics (i.e. microbe-metabolite) relationships can be depicted, either for a single environment, condition, or timepoint, or for a range of environments, conditions, or times.

1.3.1 Environmental Impact

New ecological roles and niches have been revealed by analyzing the effect of environmental factors upon MDM and by discovering with whom these microbes interact. Evaluating changes in microbial composition over time, or upon addition or perturbation of the system, has led to increased understanding of the survival and distribution of candidate phyla. For example, within the euxonic conditions of Mahoney Lake [22], sulfur-reducing microbes

Deltaproteobacteria and Epsilonproteobacteria were abundant at shallow depths whereas less characterized phyla like Crenarcheaota, Natronoanaerobium, and Verrucomicrobia, were abundant at deeper sediment depths, suggesting that sulfur cycling niches of MDM are stratified by lake water sediment depth. Similarly, a single-cell sequencing comparative analysis discovered that the abundance of acI-A and acI-B clades of Actinobacteria are positively correlated with solar radiation [68], with clade distribution dependent on the pH, carbon, or chlorophyll concentration levels present within the freshwater ecosystem. Both depth and saline content of an Arctic meromictic lake affected the abundance of candidate cyanobacteria phyla, with SAR406 dominating saline-rich waters and Chloroflexi dominating at deepest depths [21]. Taken together, these findings highlight the significant effect diverse environmental features exert over unknown taxa.

The stratification of MDM based on environmental factors can also shed light on the possible functions these organisms may have. For instance, the concentrated abundance of novel bacterial candidate phyla, like GN04 or Hyd24-12 within deep, anoxic zones of Guerrero Negro hypersaline mats suggests that these MDM members use anaerobic metabolism for energy [50, 69]. Another study demonstrated that Marinimicrobia, a species known to act as a greenhouse gas sink for nitrous oxide due to its nitrc oxide reductase expression, could, by its expression of polysulfide reductase gene clusters and its abundance in oxygen minimum zones, also play a role in sulfur and oxygen cycling [6, 70].Via a combined metagenomics, metatranscriptomics, and single-cell sequencing approach, researchers predicted that Thermotogae may be unique syntrophic acetate degraders, Chloroflexi may be homoacetogens, and Marinimicrobia may be proteolytic amino-acid degraders based on their interactions with other microbes within a methanogenic bioreactor [53]. Together, these microbes were shown to form a syntrophy-based food web with Pelotomaculum [71] and other methanogens in order to degrade catabolic by-products. Based on these studies, MDM warrant further consideration as they play roles across all biogeochemical cycles.

1.3.2 Microbial Neighbor Impact

In addition to environmental factors, other microbes or organisms may also significantly affect the presence and distribution of MDM. Upon investigating the relationships of MDM, researchers have uncovered that these unknown, not-yet-characterized microbes often have unusual lifestyles that are predominantly symbiotic in nature. Some CPR bacteria and DPANN (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota) archaea have eukaryotic hosts [61, 72, 65]. Others, like ARMAN (Micrarchaetoa and Parvarchaeota) archaea, were experimentally shown to be episymbionts [61, 73, 74] (symbionts that attach to a host's surface but are not contained within the host itself) of other archaea, like Thermoplasmatales or Cuniculiplasma [75]. The endosymbiotic relationship of a TM6 bacterial strain with free-living amoeba [76], the epibiotic and parasitic relationship of a TM7 phylotype with an Actinomyces odontolyticus strain (XH001) [77], and the symbiotic relationship of "Candidatus Sonnebornia yantaiensis", a member of candidate division OD1, with the ciliated protist Paramecium bursaria [78] further demonstrate the prevalence of symbiotic relationships among MDM and strongly suggest co-evolution between Archaea, Bacteria, and Eukarya.

Increasingly, researchers are turning to network theory to pinpoint key members of an environment and detect otherwise missed relationships [79]. Via network analyses, where a microbial community is mathematically represented as a network made up of nodes (representing different species) and edges (representing the relationships between species), researchers have studied interactions within a variety of environments, including the human gut [16, 80, 81, 82], soils [83, 47, 84, 85], and marine [86, 87, 88] environments. By building networks, researchers can uncover significant factors while simultaneously ignoring massive noise or redundant information. Additionally, network metrics [89, 90, 84, 91], such as hub score, betweenness centrality, degree centrality, and closeness centrality, enable researchers to quantitatively determine the importance of individual elements or of the entire community itself. Each network measure illuminates different properties, such as which node interacts with the most members (degree centrality), which nodes makes up the central core of a network (closeness centrality),

and which node is most important in relaying information as a messenger among other nodes (betweenness centrality). Members of a network that have both high degree and high betweenness centrality are the most highly connected to all other members and are considered "hubs" [18, 92]. The removal of these hubs can be detrimental to the overall network, as many node connections (for instance, species interactions) become lost, leaving the network extremely fragmented. Nodes with low degree but high betweenness centrality are defined as keystone species [18] and represent rare, lowly abundant members of the overall community that play a disproportionately large role due to their large numbers of relationships. Overall, each of these measurements can give importance to different nodes. Consequently, conclusions about the most important member to the network or the topology of the network itself may vary depending on the chosen centrality measure. The overall network structure can also be compared across different networks by its network topology (the arrangement of nodes and edges) and network distribution pattern (whether the ratio of edges to nodes follows a binomial or power-law distribution) to further clarify biological relationships.

Up until recently, an overwhelming proportion of network studies focused on human gut and soils-based ecosystems over more extreme environments and, oftentimes, discarded unclassified bacteria in early steps of network construction. However, in the few network studies conducted on unclassified phyla, researchers have uncovered that MDM play important roles in community stability maintenance. For instance, members of the Sphingomonodaceae and Saprospriaceae families, both of which were first characterized through 16S rRNA gene phylogeny, were found to act as "gatekeepers", interacting across large, small, and midsize streams and wherever water, sediment, and biodiversity from these streams mixed [93]. The presence of these species proved essential for interconnecting the stream community. When these gatekeeper nodes were removed from the network, the network became greatly fragmented, indicating that these taxa are significant for the persistence of stream ecosystem topology and biodiversity. Another co-occurrence network analysis showed that Acidobacteria, Frateuria, and Candidate Phylum members act as "keystone taxa" in soil communities during organic matter

decomposition [94], suggesting that these species play an essential, integral role in the soil community in driving carbon turnover. In these examples, unknown phyla are suggested to be necessary for the protection and upkeep of soil and aquatic environments, implying that unclassified organisms should be included in network analysis to enable full recovery of all ecosystem interactions and detect all key organisms.

1.3.3 Combined Environmental and Biotic Impact

Advances in network analysis have enabled evaluation of changes in an environment over time and, via bipartite networks, investigation of relationships that go beyond pairwise associations, such as cross-domain interactions, host-species interactions, and species' metabolic interactions. In these analyses, nodes of different shapes represent different species (within and across Kingdoms), metabolites, or environmental factors while edges represent relationships between these actors that connect together different node types. These networks enable researchers to more accurately evaluate an ecosystem, by deciphering and considering all interactions, even those of viruses or fungi, within an ecosystem. Results from these analyses, such as the high correlation between unknown eukaryotes and dinoflagellates [95, 88, 96] and time-dependent correlated associations among Sar11 members, stramenophiles, alveolates, cyanobacteria, and ammonia-oxiding archaea [97], suggest novel symbiotic and parasitic associations across domains. Additionally, a decade-long longitudinal study on monthly free-living microbial community interactions across different depths of the San Pedro Ocean station [98] demonstrated that species' abundance and co-occurrence relationships, such as those of Marine Group A and Nitrospina, are potentially shaped by temporal and environmental changes as well as the migration patterns of nearby microbes, again highlighting the importance of studying all underlying interactions within an ecosystem and the strong interdependence between environment, time, and taxa.

Further longitudinal and integrated omics investigations of unclassified phyla are therefore key to better understand ecosystem diversity and interactions and may improve biotechnological research efforts and enable characterization of previously non-cultivable organisms. For instance,

the presence of dense network clusters of non-peptide ribosomal synthetase (NPRS) genes within Rokubacteria genomes [99] suggests that this candidate phyla may be involved in natural product and secondary metabolite biosynthesis. Similarly, the links between viruses and bacterial methanotrophs and detection of virus-encoded glycoside hydrolases in permafrost soil [96] suggest that viruses may potentially act as predictors of methane dynamics, indirectly influencing carbon cycling and degradation. Lastly, the connected nodes of Prevotella oris OT311 with the uncultivable organism Tannerella sp. from a modeled network analysis [100] acted as the groundwork to isolate colonies of Tannerella sp. in the presence of the helper P. oris OT311, resulting in successful enrichment of a previously uncultured microbe, thus demonstrating the power of network predictions to advance MDM research.

1.4 Phylogenetic and Phylogenomic Functional Analysis

The discovery of novel MDM and the interactions or metabolic roles these Candidate Phyla play in unusual environments has led to new theories on the origin and evolution of life. Carl Woese's phylogenetic studies of the 16S rRNA gene and subsequent argument for dividing diversity on Earth into three domains, including a new domain called Archaea, first revolutionized the theory of the evolution of life [101]. Since that time, advancements in sequencing strategies of uncultivated microbes have led to the discovery of new phyla, genes, functions, and pathways that continue to challenge our understanding of evolution.

For instance, shared eukaryotic-like cytoskeleton genes and pathways of Lokiarcheaota and Asgard archaea with Eukarya [102, 103, 104, 38] have convinced some researchers that Eukarya has branched off of Archaea and consequently, all life stems either from Bacteria or Archaea. On the other hand, the existence of novel organisms like virophages [105, 106, 107] and novel genes, like recA and rpoB homologs, which are believed to belong to uncharacterized viruses or ancient paralogs unrelated to any known organisms, have caused researchers to consider adding a fourth domain to differentiate all life on Earth [108].

In any case, the use of phylogenetic marker gene analyses to group MDM sequences to either of the possible domains (Archaea, Bacteria, Eukarya, and viruses) through a process called

phylotyping [108] has led to the discovery of new lineages, more phylogenetic diversity, and many new functions across the domains of life. Through the use of phylogenetic, phylogenomic, and constraint-based modeling tools, researchers are now able to classify and infer functional ecological traits for both existing and novel lineages, evaluate genomic, metabolic, and functional conservation, and even study the cellular metabolism and dynamics of individual species or entire communities (Fig. 1-4), greatly advancing knowledge of both microbial ecology and evolution. Though these computational approaches have mostly been applied to small or simulated datasets, to well-studied organisms with high-quality, well-annotated and nearly complete genomes, and to well-studied environments like the human gut (using tools like MICOM [109]), continued advances in phylogenetic and phylogenomic modeling of the microbiome will illuminate the key metabolic mechanisms and geochemical processes necessary for survival and adaptation of all microorganisms, including those currently lacking genomic representation.

1.4.1 Phylogenetic Analysis

Initial studies of the clustering of MDM within branches of phylogenetic trees acted as catalysts in improving our understanding of the evolution of life. Phylogenetic tree analyses first revealed shared characteristics of Eukarya and Bacteria in Candidate phyla, such as the presence of eukaryote-like proteins, cytosol compartmentalization, and a mitochondrion-like anammoxosome for energy production in Planctomycetes species [110]. On the other hand, a maximum likelihood 16S rRNA phylogenetic tree showed that many CPR genes encode eukaryotic-like self-splicing introns and proteins, suggesting shared traits between Bacteria and Eukarya [11]. Additionally, though methanogenesis was long believed to originate in Euryarchaeota, recent findings of putative methane-metabolizing genes in Bathyarchaeota and methyl co-enzyme reductases in the new archaeal phylum Verstraetearchaeota suggest that methanogenesis is much more phylogenetically widespread [38, 111]. Sequencing of genomic DNA from an acetate-amended aquifer resulted in the discovery of an archaeal-like hybrid type II/III ribulose-1,5-biphosphatecarboxylase-oxygenase (RuBisCO) enzyme in the bacterial phylum OD1 and a previously unknown phylum-branch of Bacteria named PER (Peregrines) [64]. These

bacteria do not have a TCA (tricarboxylic acid) cycle, use anaerobic fermentation, and, like Archaea, are believed to rely upon RuBisCO for energy and metabolism, again suggesting that Bacteria and Archaea may be more similar than believed. Finally, bacterial OD1 members only use acetyl-CoA synthetase for ATP generation, just like sulfur-reducing Archaea, implying that certain enzymes or functions may be shared across domains. Taken together, phylogenetic analyses of the genetic and metabolic properties of MDM have revealed new similarities across the domains of life, broadening our understanding of evolution itself.

1.4.2 Genomic Reconstruction

16S rRNA marker genes alone are insufficient to inform researchers of all the genomic and metabolic diversity of a community. Instead, to recover complete genomes of all microorganisms within a community while circumventing the primer biases and imprecise taxonomic limitations associated with 16S rRNA amplicon sequencing, researchers now rely on analyzing high-quality metagenome-assembled genomes (MAGs), which enable species and strain-level resolution. By this manner, the viral, eukaryotic, and plasmid gene content present within a community can be detected, in addition to the archaeal and bacterial content traditionally found from amplicon analyses, and the phylogenetic position of each organism can be determined based on concatenated sequences of marker genes. These recent advances in genomic reconstruction have enabled functional examination of organisms without cultured representatives, in the process, revealing unique characteristics of novel prokaryotes and viruses that contradict existing criteria used to differentiate within and between domains. For example, the existence of the giant virus Acanthamoeba polyphaga mimivirus (APMV) [112, 31] and ultra-small ultramicrobacteria [113, 114] suggest that the size characteristics used to distinguish between domains need to be re-evaluated. Furthermore, the discovery of a unique 3 aa long insert in RpoB protein in Chlamydiae, Verrucomicrobia, and Lentisphaerae, all part of the PVC (Planctomycetes, Verrucomicrobia, Chlamydiae) superphylum suggests that species from these phyla share “a common ancestor exclusive from all other bacteria” [110], implying deeper evolutionary branching of Bacteria. Genomic reconstruction, in addition to the prevalence of unusual lifestyles

and presence of eukaryote-like protein domains among MDM, suggests that certain clades, like the Candidate Phylum Poribacteria, may belong to separate, evolutionarily distant, bacterial phyla, that bridge the very branches separating Archaea, Bacteria, and Eukarya on the microbial tree of life.

Functional annotation of high-quality MAGs has helped reveal unusual metabolic and physical capabilities within and across domains,demonstrated the prevalence of lateral and horizontal gene transfer among organisms [65, 115, 116, 117],and has illuminated extremophilic microbial dark matter, like halophiles [30]. Comparative genomics analyses have revealed occupation of distinct niches among species or clades of the same class or genus, through key differences within metabolic pathways at species and strain level for members of the genus Acidithiobacillus [44] in response to extremely acidic or metal-enriched conditions and differences in abundance and polysacchardize utilization and degradation genes among six distinct Polaribacter clades [118] during algal blooms. Large-scale reconstruction efforts, like the successful reconstruction of 2,631 draft genomes from global oceans [119], the discovery of over 4000 conserved small proteins with no known functions in the human microbiome [120], the phylogenomic analysis of 10,575 genomes which revealed closer evolutionary proximity than previously believed between Archaea and Bacteria due to the use of 381 marker genes [121], or the reconstruction of over 150,000 genomes from human-related metagenomes to produce thousands of novel species-level genome bins [122], stand to be the key to detect the yet-undiscovered novel organisms and gene products present on Earth. These studies were not able to map all reads [122], suggesting that current large-scale endeavors still fail to capture all possible diversity present. Consequently, capturing all species and strain-level differences within a community seems to require thorough evaluation and de novo discovery of genomes related to non-prokaryotic components as well. A key step towards capturing all species and strain-level differences relies on recovery of non-prokaryotic organisms, such as the recovery of eukaryotic genomes from metagenomes, which has recently been shown to be possible, with the near complete genome reconstruction of three fungi and an arthropod by a k-mer based strategy [123].

Future endeavors must therefore integrate existing and novel strategies to recover both prokaryotic and non-prokaryotic elements of natural communities.

1.4.3 Genome-scale Metabolic Reconstruction

Today, efforts focus on metabolic reconstruction of these high-quality draft assembled genomes, to better understand the complete metabolic processes of each individual organism and community-wide interactions in general. These metabolic reconstruction studies have enriched understanding of adaptation mechanisms and diversity and shed light on the lifestyles and potential roles unknown organisms play within geochemical cycles. For instance, a combined metagenomics and proteomics-driven analysis revealed fermentation-based metabolic properties in candidate phyla including Melainibacteria, OD1, SR1, and TM6, suggesting that these unknown members, by their ability to produce hydrogen, acetate, formate, and lactate, may be important to a wide range of geochemical cycles[124]. Other studies have proposed a fermentative saccharolytic lifestyle for an Aminicenantes member based on the presence of glycosyl hydrolases, secreted peptidases, and complete Embden-Meyerhof glycolytic pathway genes and suggested roles in iron, sulfur, and nitrogen cycling for potential novel Candidatus Acidulodesulfobacterale members based on their expression of nitrogenase-encoding, iron metabolism, and sulfur oxidation genes [125]. Lastly, functions in in carbon acquisition and breakdown were revealed for 21 different phyla, including Aerophobetes, Aminicenantes, TA06, and Bathyarchaeota within deep-sea sediment petroleum seeps after pathway reconstruction from 82 MAGs [126] showed that these species' genomes encode enzymes for anaerobic oxidation and hydrogen and acetate metabolism. Further metabolic reconstruction studies stand to improve understanding of both microbial adaptation and geochemical processes.

Both genome-scale modeling and integrated omics studies have recently been applied to unravel microbial community interactions and adaptations. Genome-scale metabolic modeling in over 800 microbial communities [127] predicted, by prevalent and widespread multispecies interactions, high resource competition, and high metabolic interaction potential within co-occurring subcommunities, that metabolic dependency drives co-occurrence patterns.

Continued development in the field of genome-scale metabolic modeling, particularly with respect to organisms lacking cultured representatives, stands to enhance current understanding of species co-occurrence, adaptation, and evolution. Additionally, the integration of single-cell genomic and transcriptomic sequencing data to existing phylogenetic and phylogeneomic studies is of great interest to illuminate metabolic flux dynamics not just at the organismal level but at the level of individual cells. Recently, a rare subpopulation of *Staphylococcus aureus* cells undergoing prophage induction were successfully distinguished through the novel pipeline PETRI-Seq [128], suggesting a promising start to evaluate single-cell states of microbial communities. Meanwhile, integrated omics studies, such as a combined metagenomics, proteomics, and metabolomics study on soil samples within Northern California grassland [43], have revealed novel contributions to geochemical processes, like carbon turnover by candidate phyla Gemmatimonadetes and CPR and aromatic amino acid degradation by Bathyarchaeaota and Thermoplasmataes respectively. These insights could not have been achieved with one sequencing approach alone, suggesting that complete understanding of microbes, life's cycles and ecosystems requires an interdisciplinary multi-omics approach to better understand all components (species, genes, proteins, and metabolites) of a given community.

1.5 Discussion of Limitations

1.5.1 Current Diversity Analysis Limitations

Continued efforts in characterizing the diversity of Antarctic. hypersaline, non-soil-based (aquatic), and other rare environments are necessary to further improve understanding of biodiversity on Earth. Additionally, primers specifically designed for archaea and newly discovered phyla with alternative ribosomal gene structures or innovative methods that avoid primer biases [129] altogether while enabling maximum recovery of novel diversity should be implemented to capture the most accurate representation of the taxonomic diversity of an ecosystem. Lastly, a global analysis of microbial diversity and dynamics is currently incomplete. To achieve the most comprehensive understanding of microbial diversity, a standardized methodology and benchmarking for collecting, evaluating, and integrating individual datasets

must be created and additional omics data must be considered to unravel the diversity of species' genes, metabolites and pathways.

1.5.2 Current Interaction Analysis Limitations

Though network analyses have proven powerful in generating hypotheses, the prediction of species' metabolic interactions, host-species interactions, and cross-domain interactions from network analyses is still relatively new and challenging, for studying either well-characterized or novel organisms. Technical and biological obstacles, such as resolution, sequencing depth variation, inclusion of other biotic and abiotic factors, and the inherent sparse, compositional characteristics of metagenomics data, hinder differentiation between strains and species, may introduce spurious edges, and may contribute to inaccurate or exaggerated accounts of the true biodiversity of an ecosystem. Additionally, rare taxa are often discarded from network analysis, particularly if too high of a prevalence threshold is used [130]. Biological interpretation of results remains the most challenging, as even directed networks cannot always differentiate relationships [130] like competition from amensalism, occluding full understanding of community interactions. Consideration of these limitations is necessary in future interaction studies to retrieve accurate results.

1.5.3 Current Phylogenetic Analysis Limitations

Despite rapid advancements in phylogenetic and phylogenomic analyses, certain obstacles continue to make interpretation of results difficult. First, no genome assembly method is universally applicable [131, 30], with certain methods better suited for human gut microbiomes and others for halophilic environments. Consequently, though steps have been made to automate processes within genome reconstruction, manual curation remains a necessary part of genome assembly and validation [132]. Genomic variants, particularly strain-specific variants, remain challenging to identify and most assembled genomes are incomplete or may be of poor quality. Subsequent validation of results is imprecise and highly memory intensive [131] due to reliance on closely related phylum representatives for candidate phyla with incomplete or absent genome information. Metabolic reconstruction efforts are also hindered by the lack of functional

annotation or gene-reaction associations for most genes [133]. Tools for predicting metabolic flux dynamics, like SteadyCom [134], are currently only suitable for small datasets of well-characterized organisms. Lastly, integration of different omics types remains daunting. Future studies are therefore required to maximize genome recovery and quality while minimizing challenges associated with data preprocessing, assembly, and interpretation.

1.6 Conclusions and Future Developments

Advancements in sequencing technologies and methodologies for analyzing microbial diversity and function have enabled detection of novel phyla and gene products that drive major geochemical cycles. Though certain challenges remain to best evaluate diversity, interaction, and phylogeny, significant strides in research have improved current understanding of biodiversity and microbial evolution and have made it possible to isolate previously uncultured organisms. Future work centered on long-read sequencing, single-cell genomics, multi-omics analyses, and more network modeling, such as of holo-biont associations, community-wide metabolic associations or inter-domain sequence similarity networks of evolutionary diversification events, will further illuminate understanding of unknown prokaryotes, ecosystems, and the origins and adaptations of life.

One particular gap in knowledge that requires immediate attention is the impact of novel, poorly-characterized phyla (referred to as Microbial Dark Matter (MDM) upon ecosystem function. MDM, like any other form of life on Earth, does not exist in isolation and its role is therefore affected by both the organisms with which it interacts and its surrounding environment.Considering these interactions is vital to understand why and how MDM subsist across environments.While some studies have attempted to make sense of MDM via its ecological context, few have captured a global role of MDM, attempted to model these interactions via networks, or identified key genes responsible for the relative role of MDM within environments.Consequently, there is a crucial need to ascertain the adaptive properties and roles MDM possess to adapt to extreme environmental conditions, without which we will never be able to complete the puzzle on the origins and evolution of life.

We hypothesize that, by their abundance patterns and possession of a particular set of genes and evolutionary traits, MDM are potentially uniquely adapted to extreme environments and possibly highly relevant to community structure and stability of these environments. Accordingly, to enhance understanding of the ecological and functional role of MDM, we propose to 1) Develop a computationally scalable approach for studying the ecological relevance of MDM using network theory (particularly using the SPIECEASI approach to estimate microbial associations and various network metrics to quantify the importance of unknown taxa); 2) Apply this methodology to study the local and global relevance of MDM in extreme environments; 3) Identify novel gene functions associated to key MDM components of extreme environmental networks via a metagenomics and comparative genomics approach (including the use of tools like prodigal, eggnog-mapper, EFI-EST, and Blast2GO to identify and functionally annotate uncharacterized genes). The key microorganisms and functions encoded by these organisms' novel genes produced as results of this computationally intensive approach are expected to improve understanding of how unknown microbes shape other microbial members, their respective environments, and even Earth as a whole, helping to clarify pressing questions of microbial ecology and evolution.

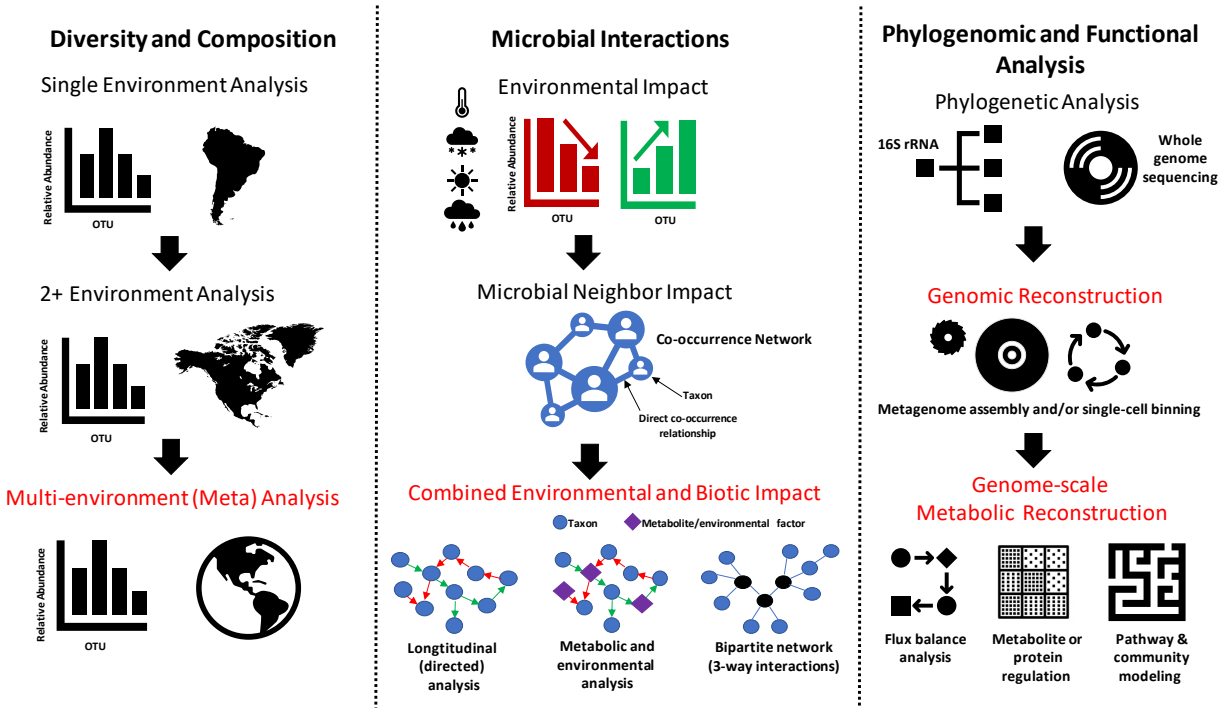


Figure 1-1. Progression of diversity and composition, microbial interaction, and phylogenomic and functional analyses for studying Microbial Dark Matter.Each column panel shows, from left to right, depictions of general strategies for analyzing diversity and composition, microbial interactions, and phylogenomic and functional analysis. Within each panel, downward arrows indicate the chronological progression of each strategy. Marked in red are on-going investigations that remain difficult due to unresolved challenges.

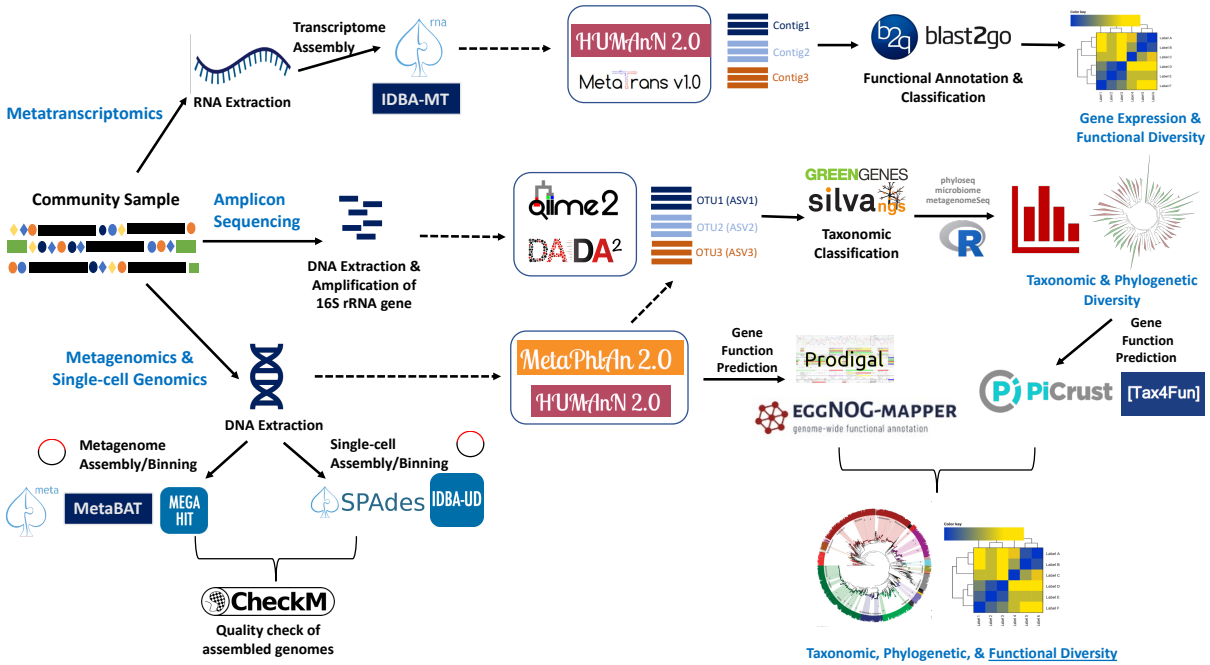


Figure 1-2. Methodologies for Diversity Analyses of NGS Data. Starting with a community sample, gene expression (through analysis of metatransriptomics data), taxonomic and phylogenetic diversity (through analysis of amplicon sequencing data), and functional diversity (through analysis of metatranscriptomics, amplicon sequencing, metagenomics, and single-cell genomics data) can be obtained, using the various tools and pipelines indicated for each omics data type. Arrows show the workflow of analyzing each omics data type. Dotted arrows precede pipeline tools and indicate the shared final output of OTUs (operational taxonomic units of diversity) or ASVs (amplicon sequencing variants) obtained following metagenomics, single-cell genomics, and amplicon sequencing.

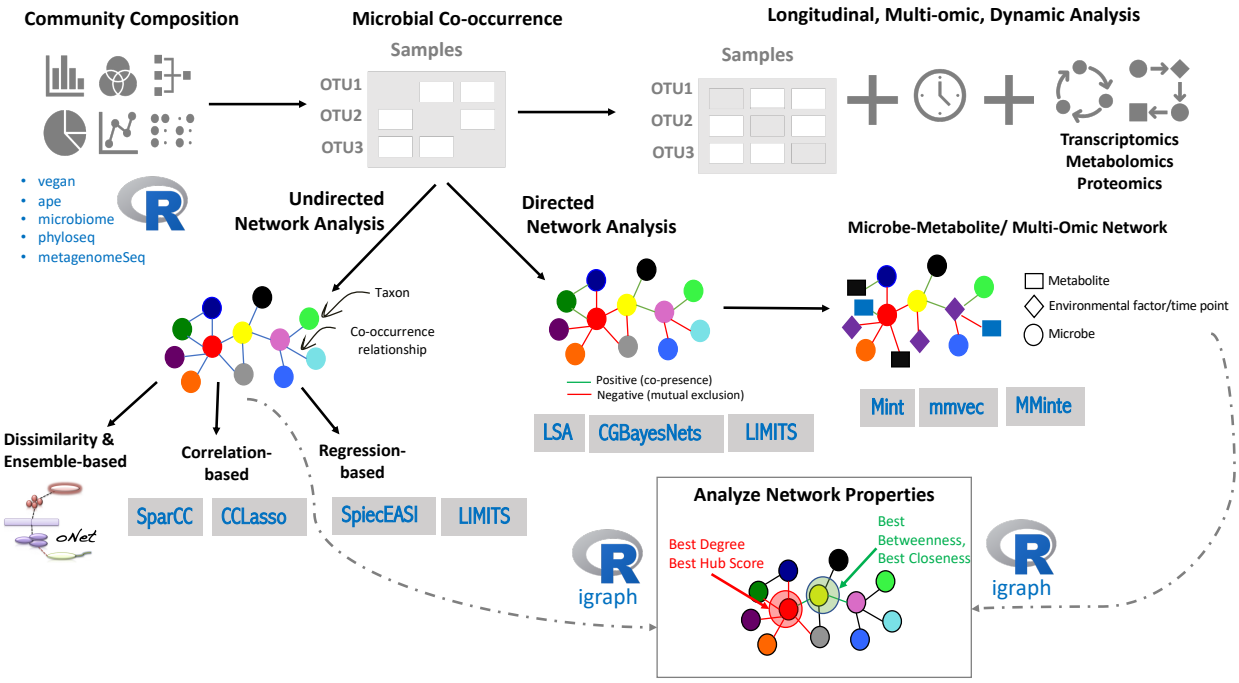


Figure 1-3. Methodologies for Interaction Analyses. Current state-of-the-art tools and R packages for analyzing community composition, microbial co-occurrence, and more complex interaction studies (including longitudinal, multi-omic, or dynamic network-based analyses). Undirected network, directed network, and multi-omic networks are depicted as models, with tools labeled in blue and outlined in gray boxes. All interaction networks can then be analyzed using network properties and metrics, through the R package igraph.

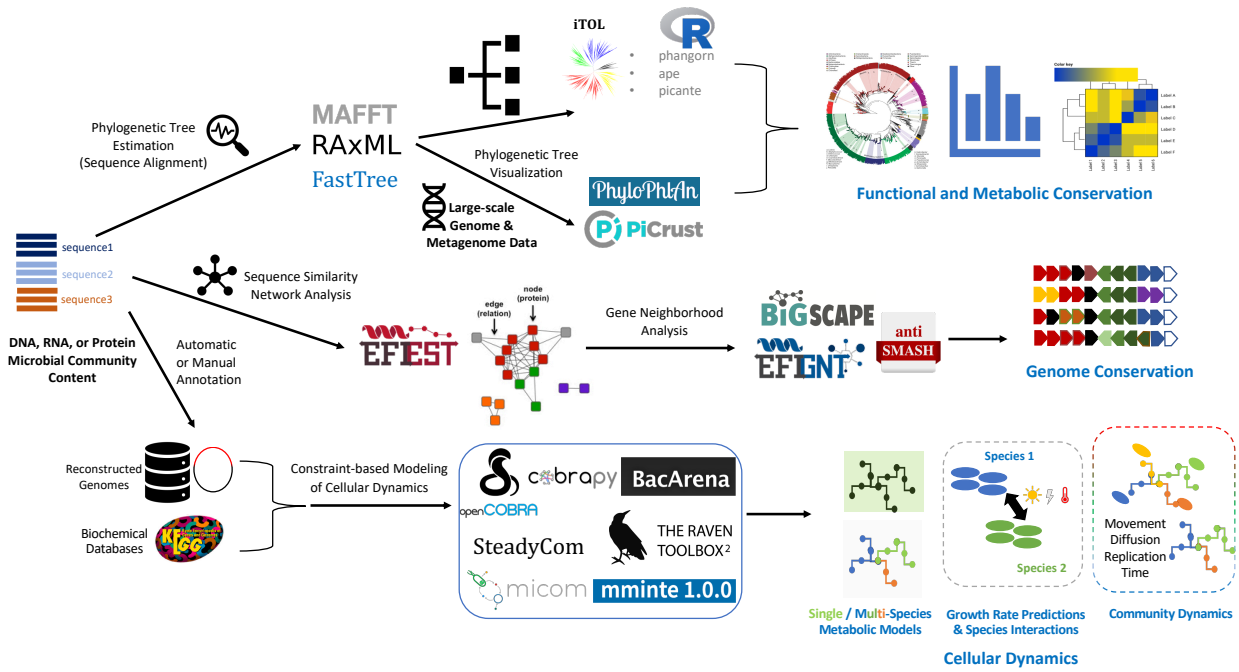


Figure 1-4. Methodologies for Phylogenetic and Phylogenomic Analyses. Current state-of-the-art methodologies for phylogenetic and phylogenomic analysis of a microbial community. The DNA, RNA, or protein content of a given microbial community can be studied to shed light on functional and metabolic conservation (through phylogenetic analysis tools), genome conservation (through sequence similarity network analysis and gene neighborhood analysis tools), and species-specific or community-wide cellular dynamics and growth behavior(through constraint-based modeling tools).

CHAPTER 2
PRELIMINARY NETWORK ANALYSIS ON TOY DATASET

2.1 Introduction

In recent years, unknown, unclassified microorganisms have been found to constitute the majority of active cells and to proliferate in most major, non-human-associated biomes [63]. Metagenomics and 16S rRNA amplicon sequencing-based analyses have revealed a plethora of novel 'candidate' bacterial and archaeal phyla [9, 56, 135], whose inclusion into the microbial tree of life has produced novel branches, discovery of domain-intersecting essential genes, and an altogether new view of evolution itself [61, 38, 136]. These findings have demonstrated the significance of unknown microbes across multiple taxonomic classification ranks, particularly from phylum to genus [63], and argue for further exploration into the realm of unknown taxa across both diverse taxonomic and environmental landscapes. What constitutes as unknown (and how significant this element may be) can thus be evaluated per environmental sample, per taxonomic level, per functional role, and per gene, with each evaluation enriching understanding of microbial ecology and the role of the 'unseen microbial majority'.

The discovery of novel phyla which are evolutionarily divergent [137, 138] in terms of physical characteristics and protein content from most known microbes, the disparity in the number of microorganisms discovered through sequencing methods versus those able to be cultured [62], and the overwhelming proportion of unknown genes within most microbial genomes [139, 140, 141] are evidence that advances in next-generation sequencing have only illuminated how little we know and how much we still have to learn about unknown microorganisms. Clearly, there are multiple levels of unknowns [141], from unknown microorganisms to their gene content, that have yet to be fully examined.

One major question that stems from the recent findings of unknown taxa as dominant, active cells across the globe is whether or not unknown taxa are also ecologically relevant, and if so, in similar or different ways to their known microbial neighbors. While the complete ecological relevance of unknowns will likely take many years to solve, one method that may help address this challenge is to use network theory to identify and quantify the contributions of unknowns. The mathematical modeling of ecological and biological systems as networks [142, 91, 143] has

already enabled researchers to accurately visualize and predict the dynamics of well-characterized genes, taxa, metabolites, and abiotic factors within various ecosystems [144, 145, 146, 147, 50, 20]. Meanwhile, network measures like degree, betweenness and hub score centrality metrics [143, 148, 149] have enabled quantification of the contribution of each element, revealing the most significant taxa [17, 150, 151], genes [152, 153, 154, 155], or pathways [156] responsible for driving a particular system. We hypothesize that application of network theory can also be useful for studying the unknown components of microbial environments, particularly for clarifying the relative position and ecological relevance of unknown taxa.

Here, we aim to develop a complete pipeline to study the relevance of unknown microorganisms in environmental communities through network theory. A reduced 16S rRNA amplicon dataset was used to test and design this network-based approach. We evaluated if networks including unknowns are possible, if network measures are useful for analyzing unknown taxa, and which other variables of a typical metagenomics analysis must be controlled to address and avoid potential confounding factors along each step of the pipeline. These validation checks enabled the completion of a methodology that would be robust and applicable for large-scale multi-environmental 16S rRNA amplicon data.

2.2 Results

2.2.1 Analysis of the Effect of Data Source Variability

Presently, thousands of 16S rRNA amplicon-based studies are accessible from public repositories, like the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) or the Joint Genome Institute (JGI Gold), for large-scale metagenomic analysis. Though utilization of these publicly available data allows for an affordable and time-independent analysis of microbiomes compared to manual data collection and sample retrieval from various ecosystems, it also introduces new challenges in the integration and comparison of diverse datasets. Effective comparison of data requires evaluation of elements that could introduce biases in data and potentially confound results, including but not limited to differences in sequencing

platform, primer set, sequence quality, sequencing depth, sample size, and sampling location across project studies [157, 158, 159, 160]. The effect of each of these variables must be evaluated to better understand the data sources themselves and determine to what extent differences in data quality or sequencing technology may change the results of a microbial composition analysis of one or more environments. Other choices, like the selected 16S rRNA pipeline approach [161, 162], quality filtering method [159, 163], taxonomic assignment (clustering) method [164], and reference taxonomic database [165, 163] may also affect downstream analysis results, potentially causing significant differences in both taxa prevalence and abundance. Researchers must therefore carefully consider the parameters chosen at each step of amplicon-based metagenomics analysis to ensure that results are driven by biology and not by technical variability.

Consequently, our first step in developing a network-based pipeline was to evaluate the impact of a number of parameters (such as sequence quality, sequencing depth, sampling location, sample size, and reference taxonomic database) on a 26-sample toy dataset (PRJNA401502 -Biogeography of cyanobacteria project) that was produced using a consistent sequencing platform (Illumina MiSeq) and standardized targeted primer region (V4). The latter elements (sequencing platform and primer set) needed to be standardized for this toy dataset, and any multi-sample study in future, as literature has demonstrated that use of different sequencing platforms and amplicon hypervariable regions is most error-prone, leading to the greatest number of inconsistencies and technical variability [166, 167] in subsequent microbial composition and phylogenetic resolution [168]. A dataset based on Illumina sequencing and V4 primers was chosen due to consistent agreement among literature of the higher quality [159] and improved phylogenetic resolution [168] associated with this particular sequencing platform and primer set.

2.2.1.1 Evaluation of read quality

The chosen toy dataset consisted of 5 different ecosystems (freshwater, hot springs, marine, plant, and terrestrial) and 26 globally distributed samples (5, 10,5,1,5 for each ecosystem respectively), (Fig. 2-1A) ensuring that a sufficient number and diversity of samples (excluding

plant) was present for an accurate comparison of results within and between environments. To develop a suitable way to measure the impact and potential bias introduced by differing sample quality in a large-scale study, we analyzed the total read counts, total taxa identified, and the proportion of reads passing quality filtering and taxonomic assignment among environments within the toy dataset. Though the number of total reads and total operational taxonomic units (OTUs) was directly proportional to the sample size of a given environment, a similar proportion of reads passed quality filtering and taxonomic assignments across all environments studied. As shown in Figure 2-1C, the number of total reads, total reads passing quality filtering, and total number of OTUs recovered was highest for hot springs (the environment with the most samples) and lowest for plant (the environment with the lowest samples),yet, on average, a similar and high percentage (95-98 percent) of reads passed quality filtering for all environments, indicating that similar read quality levels can be expected from different environments. Similarly, though a wider range was exhibited in the percentage of mapped reads across environments, a consistently high proportion of reads (ranging from a mean of 92 to 97 percent) was mapped to OTUs, indicating that similar read mapping rates are possible across environments, and therefore should be required in a larger scale analysis. These preliminary results also suggest that mapping rate is not likely to act as a bias in a multi-environmental study. These results were promising and suggested that integration of samples from diverse environments is possible.

2.2.1.2 Identification and breakdown of unknown OTUs

Having established that different environments could be evaluated in a nonbiased manner by read quality, we then designed a strategy to identify the unknown components of each environment, using the QIIME open reference clustering approach and the SILVA database as a reference for taxonomic classification. For a deeper understanding of unknown distribution and contributions, unknowns were defined and evaluated at each taxonomic rank from phylum to genus, where any taxa classified as unknown, ambiguous, or uncultured by SILVA were reassigned to an unknown status for that particular taxonomic rank. By this manner, the relationship between taxonomic rank and unknown contribution could also be explored, enabling

a more comprehensive overview of the ecological impact of unknown taxa. The total proportion of unknowns to knowns of all OTUs and the breakdown of unknowns initially classified as unknown, ambiguous, or uncultured was calculated at genus rank.

The majority of OTUs consisted of knowns for all environments, with marine and hot springs harboring the greatest proportion of unknowns (Fig. 2-1C). These latter environments consisted of samples originating from extreme environmental conditions (Yellowstone National Park and Antarctica respectively), suggesting that targeting extreme aquatic environments may result in greater recovery of novel taxa. For all environments, the greatest proportion of unknowns were initially deemed unassigned by the SILVA database during OTU mapping, with the second largest proportion of unknowns belonging to taxa with ambiguous taxonomic classification for all environments but plant. The methodology used to identify unknowns was shown to produce consistent and similar patterns in the breakdown of unknowns across environments, suggesting that accurate comparisons could be made for diverse data.

2.2.1.3 Evaluation of prevalence

Next, prevalence (the number of samples in which a given OTU is found) was evaluated to test the sparsity of different data sources and to determine if any adjustments are needed to account for differing rates of sparsity among environments. Prevalence curves were used to show the proportion of taxa present in relation to an increasing number of samples. The prevalence of both known and unknown taxa was evaluated for each environment to investigate the effects of taxonomic classification status and environment on prevalence (i.e. if unknown taxa are significantly rarer and less prevalent than known taxa and if the proportions of certain taxa are significantly different for particular environments). Despite the low proportion of unknowns to the overall taxonomic composition (unknowns made up at most 20 percent of total OTU counts per environment), we found that the prevalence of these taxa was comparable and similar to knowns, with most OTUs, regardless of their known or unknown status, only appearing in one or few samples (Fig. 2-1B) for all environments. The results of the toy dataset show that, regardless of environment or taxon type, taxa abundance tends to decrease as the proportion of all samples in

which that taxa must be observed increases. In fact, this observation is even more pronounced for environments with a larger total sample size.

As illustrated in Figure 2-1B, the majority of taxa present in the hot springs community of the toy dataset are found in just 3 out of 10 samples, with the number of OTUs sharply dropping after inclusion of at least 5 samples. The sharp decline in OTU counts in response to an increased sample prevalence threshold suggests that studies with larger sample sizes may require an adjusted, less stringent sample prevalence threshold to feature an adequate number of OTUs. For these reasons, we applied a prevalence percentage filter of 60 percent to marine, terrestrial, and freshwater environments due to their similar total sample size (5 samples) and a slightly less stringent percentage filter of 40 percent sample inclusion to hot springs, enabling a similar number of OTUs (at least 100 OTUs) to be featured in networks across environments.

In future, prevalence curves will be applied to a large study to understand what filters must be set to ensure similar taxa prevalence across different data sources. Prevalence curves were thus helpful in evaluating what percentage of prevalence to filter the dataset and adjusting for differences in sparsity and sample size. The analysis confirmed that 16S rRNA data is both compositional and sparse, regardless of environment studied. Meanwhile, the similar prevalence curves among known and unknown taxa suggest that unknowns may be just as integral to their respective microbial communities as their known microbial counterparts, and therefore merit inclusion in microbial co-occurrence networks. In conclusion, although this prevalence analysis indicates that environment specific prevalence thresholds might be required to ensure comparable data, it also shows that prevalence levels will likely similarly influence the presence of known and unknown OTUs in environmental networks and therefore not significantly affect their comparisons.

2.2.1.4 Evaluation of shared taxa

Another sparsity-related problem that needed to be evaluated was to determine if any OTUs were found in only one specific subset of samples or locations of an environment as these location-specific species may introduce biases in our network analysis by creating strong,

spurious links that are in reality only due to a small number of samples. Although this would need to be evaluated per environment for a large-scale study in future, here, due to the small sample size of the toy dataset, we addressed this sparsity problem by evaluating the abundance of OTUs across all 26 samples. A heatmap of the OTU abundance was created to investigate whether certain samples or locations were enriched in either known or unknown taxa, with counts shown after centered-log-ratio transformation to enhance visualization. The heatmap results were useful in identifying patterns of abundance for each taxon across different locations and samples.

Applied to the toy dataset, the results demonstrated that, regardless of OTU status (known or unknown), the majority of OTUs had low counts within and across environments (Fig. 2-1E) , with only a few samples per location having an abundance of a particular OTU.,These consistent results confirmed that all prevalent and abundant known and novel microbial members should be featured in co-occurrence networks to better represent the inherent biology of each habitat and that the role of unknowns in the subsequent environmental network topology of each of these habitats does merit full investigation. Future large-scale microbiome analyses should therefore incorporate the use of heatmaps to evaluate taxonomic abundance patterns and discard any location-specific OTUs to prevent strong but spurious co-occurrence associations from being introduced into subsequent microbial co-occurrence networks.

2.2.1.5 Distribution of known and unknown OTUs across environments

Next, we investigated how often OTUs were shared between environments to better understand the composition of each environment and determine whether a combined network or distinct environmental networks would best describe overall microbial composition. Both unknown OTUs and the total OTUs present were shown to be environment-specific (Fig. 2-1D), with only a small fraction of OTUs shared across all five environments. The largest proportion of OTUs were shared between freshwater and hot springs and between terrestrial and hot springs, due to the shared geographic location of Yellowstone National Park among samples of these environments. The least OTUs were shared between marine and plant environmental samples, due to their divergent ecosystem characteristics and geography (with samples representing marine

and plant environments originating from Antarctica and Hawaii respectively). Due to the distinct microbial composition present within each community, microbial co-occurrence networks were necessary to construct for each environment type as an integrated multi-environmental network would, at best, represent only 155 of the total 19,445 taxa recovered after OTU mapping. Venn diagrams of the shared OTUs across environments were thus useful for determining the specificity of the microbial content and should be used for future large-scale studies to determine if integrated networks are possible or beneficial.

2.2.2 Creation and Evaluation of Network and Node Properties using SpiecEASI

The previous steps applied to the toy dataset enabled us to identify potential biases and differences among differing sampling data and allowed us to define thresholds to filter out rare taxa to avoid spurious associations from being included in subsequent environmental networks. With these initial problems solved, a correlation method that was particularly suited for sparse, compositional data was required to most accurately visualize each environment as a network of microbial co-occurrence relationships. The correlation method would need to estimate the co-occurrence relationship (edge) between taxa (nodes) without introducing spurious edges between taxa that did not directly co-occur with one another. Furthermore, the sparsity of each environmental matrix of OTU counts would need to be represented as sparse graphs, with lowly prevalent taxa needing to be discarded to reduce subsequent noise in networks. To address these challenges, the sparse inverse covariance estimation for ecological association and statistical inference (SpiecEASI [169] Meinshausen and Bühlmann (MB) neighborhood selection method was chosen to estimate and model direct co-occurrence relationships among taxa of a given environment. Before applying this method, we automatically discarded the plant environment from analysis as co-occurrence (which calls for the presence of a given taxa in at least 2 or more samples) could not be measured for a single sample. As mentioned previously, to discard lowly abundant taxa that would only contribute noise to networks, sample prevalence thresholds were applied to the remaining four environments, using a slightly lower sample prevalence threshold of 40 percent (presence of a taxon in at least 4 of 10 samples) to discard lowly abundant taxa in hot

springs and a sample prevalence of 60 percent (presence of a taxon in at least 3 out of 5 samples) to discard unwanted taxa in freshwater, marine, and terrestrial communities. The SpiecEASI algorithm then estimated the co-occurrence relationships among the remaining taxa, resulting in a sparse graph representation of the frequently co-occurring taxa for each environment. Though a resultant freshwater microbial co-occurrence network could not be attained due to insufficient taxa meeting the prerequisite sample prevalence and co-occurrence thresholds, networks were successfully created for hot springs, marine, and terrestrial microbial communities.

2.2.2.1 Network composition

A similar fraction of data was retained (Table 2-1), with a comparable and suitably large number of nodes (ranging from 143 to 234) and edges (ranging from 208 to 551) for the three environmental networks. Strikingly, marine and terrestrial networks were most and least dense respectively, despite an equivalent sample prevalence threshold. The lack of a freshwater network and difference in density among marine and terrestrial networks despite an equivalent number of samples and similar initial OTU counts for all three environments demonstrates that network interconnectedness and initial data (sample and taxa) size do not necessarily need to be related. Instead, node and edge composition of each network appear to reflect the inherent biology of each environment, irrespective of sampling selection applied. Likewise, a deeper evaluation of edge composition revealed that the proportion of unknown-unknown (U-U) edges was also unrelated to initial proportions of unknowns. Though both hot springs and marine environments had the largest proportions of unknowns initially, the edge composition differed drastically among these environments, with the majority of edges being Known-Known for the former and Unknown-Unknown for the latter respectively. Consequently, for large-scale studies, network, node, and edge composition may vary depending on the intrinsic characteristics of a specific environment, leading to different arrangements and roles of unknowns.

To further understand the presence and potential impact of unknowns on microbial community structure and composition, networks for marine, terrestrial, and hot springs communities were visualized including and excluding unknowns (hereafter referred to as

"Original" and "Without Unknown "networks) at each taxonomic classification level, with different colors signifying the respective taxonomic classification of each node.

Though the proportion of unknowns (colored in gray) and effect of unknown node removal varied per taxonomic level and per environment (with few unknowns present overall), the presence of unknown nodes was generally highest and most detrimental at genus level (Fig. 2-2), with the absence of unknowns leading to the largest fragmentation (removal of the greatest number of edges) of each "Original" environmental network. The range in network position and impact of unknowns justified evaluation of the potential ecological relevance of unknowns across the spectrum of taxonomic classification for future studies. Due to the large number of unknown nodes at family and genus levels and subsequent substantial decrease in nodes and edges upon unknown removal compared to those at higher taxonomic ranks, a method to control for the effect of node size would be necessary to most accurately assess subsequent changes in network appearance upon removal of unknowns. A method to control for node size bias would be particularly necessary for networks like the marine network that consist of a large number of unknowns, as removal of unknowns at genus-level for this toy example led to a 61 percent decrease in nodes (from 218 initially to 85 nodes) and 80 percent decrease in edges (from 551 to 120 edges).

2.2.3 Evaluation of Network Metric Influence on Taxon Importance

Though the network visualization described above helped clarify where unknowns were situated and in what proportions per environment, further understanding and a quantifiable comparison of the relative importance of unknowns within each environmental network required the use of network metrics. We focused on using the network metrics degree, betweenness, closeness, and hub score as these would, by definition, identify the most co-occurrent and prevalent [170] taxa, the taxa most responsible for interconnectedness [171], the most centrally located [172] taxa, and the taxa most essential for community structure maintenance [173, 152, 174]respectively. Since each of these measures may give weight to different taxa based on how they are calculated, we first explored which taxa were found to be most significant

according to each network measure to identify any similarities among the taxa selected and among the network measures themselves.

As illustrated in Figure 2-3A, the range of values and the proportion of the most significant hot springs taxa (colored in dark red) differed for each network metric. The widest range (0-1600) in scores was exhibited for betweenness and the narrowest range (0.0006 to 0.0011) for closeness. Only a few taxa had the highest hub scores while the majority of other taxa had low hub scores, confirming that the hot springs network, like most other environmental biomes, is scale-free and that top-scoring hub taxa are microbes whose individual absence induces the largest damage to network structure. In contrast, a larger proportion of taxa were identified as most important (colored red) by closeness, degree, and betweenness, suggesting that changes to these network measures would be most effectively studied by investigating the effect of overall unknown node removal. Additionally, though certain taxa were identified as recurring top members by all network measures (the top-scoring hub taxon also had a high degree and closeness centrality score), other taxa were uniquely selected as important by betweenness or closeness centrality, reaffirming the use of multiple network measures to capture different network properties.

Further evaluation of the distribution in network centrality scores per genus (Fig. 2-3B) demonstrated that while members of a specific genus often have similar distributions in scores among all four network measures, these patterns do not always hold, especially for selection of the highest scoring taxa by an individual network metric.For example, though members of the genus Sphingopyxis were high-scoring overall in degree and closeness and Acidovorax members were low-scoring in all network measures evaluated, certain unknown microbes at genus level had the highest hub scores and betweenness scores despite the relatively low to average distribution of degree, betweenness, and hub score values of unknowns overall (Fig. 2-3B). Furthermore, comparison of the distribution in degree and closeness scores for members of Marinobacter showed that despite their low degree (low connections to other taxa), these microbes were among the most central (highest-scoring in closeness centrality) to the hot springs network.

Different network measures therefore illuminate different aspects of networks, as the most important nodes were not the same across all metrics. For future work, including large-scale studies, researchers must therefore take into effect that the network measures used may affect or bias the selection of important microbes.

Thus, using a variety of network metrics enables identification of important microbial drivers that may have been missed using one network measure evaluation alone. Outliers in hub score that contrasted with the overall hub score distribution of members of a genus emphasize the need to evaluate this particular measure by individual node. Overall, all network measures evaluated were found to be effective in revealing different properties of OTUs present in microbial communities and were suitable for capturing the diversity in the network position of OTUs. Our results suggest that degree, betweenness and closeness are appropriate metrics to compare networks globally by analyzing changes in distribution and mean values. The more skewed distribution of hub score values,however,suggests that this metric would be more useful for an individual comparison of nodes in the selection of top-scoring hub taxa.

2.2.4 Network Measure Application to Identify Local and Global Impact of Unknowns

Consequently, the exploration and comparison of these network measures led us to evaluate changes in degree, betweenness, and closeness for all nodes present in each environmental network and to evaluate top-scoring hubs individually. We examined the mean degree, betweenness, and closeness values for all networks with and without unknowns at each taxonomic rank and again observed a stark contrast in the distribution of all network measure scores at the genus level, where unknowns made up a majority of nodes.

2.2.4.1 Effect of unknown node removal on network metrics

Using the marine network as an example, evaluation of the mean degree, betweenness and closeness values for networks including and excluding unknowns showed that upon unknown removal, start differences in mean values appeared. In particular, a substantial decrease (from 22.5 percent to 60 percent decrease) in the average degree and closeness value occurred at family and genus levels upon removal of unknown nodes. Though this observation is striking, it is

unclear whether the change in mean network metric scores is due to the role of the unknowns or due to the large number of nodes that are unknown at family and genus level. Both the increase in fragmentation to the network structure (Fig. 2-2 and Table 2-2) and the decrease to the average degree and closeness scores upon unknown removal compared to the original network appearance and properties required validation using a control group.

To ensure that the changes we observed to both the network appearance and overall degree, betweenness, and closeness scores upon the removal of unknowns were significant and truly due to the unknown status and not node number, a methodology to control for any confounding effects due to node size was introduced. At each taxonomic rank, we randomly sampled and removed an equal number of known nodes as unknowns and repeated this process 100 times, generating 100 networks with random knowns removed, and calculated the degree, betweenness, and closeness scores of all nodes in all consequent networks. By this manner, a null distribution of "Bootstrap" networks and network measure distributions was created. We then evaluated the effectiveness of these bootstrap network- driven results as a control for node size effects for future statistical evaluation of changes in network topology and node centrality for each environment. To clarify, we evaluated how the network appearance differed at each taxonomic rank and for each environment when random knowns were removed compared to when we removed unknowns. A comparison was also made in the change to degree, closeness, and betweenness centrality distributions among all nodes present in each environmental network, across taxonomic rank levels to evaluate whether degree, betweenness and closeness scores followed similar patterns (indicating no significance and a confounding effect of node size) or whether the scores produced after unknown removal were truly different from both the random known removal networks and the original network (indicating a significant effect due solely to the unknowns).

2.2.4.2 Analysis of network metric distribution upon removal of unknown versus known nodes

Degree, betweenness, and closeness centrality scores were calculated for all nodes present in the "Original", "Without Unknown", and "Bootstrap" networks and the change in score

distribution for each of these measures was evaluated by the Wilcoxon test, with p-values adjusted using the Holm Method. Boxplots were used to visualize the effect of unknown removal compared to an equivalent removal of random knowns upon degree, betweenness, and closeness scores for the hot springs, marine, and terrestrial community overall. At the same time, to showcase the most important novel or known taxon of each given environment, hub score was calculated for all nodes present in the "Original" networks and networks were visualized once more, with nodes sized as a function of hub score, from phylum to genus.

Comparison of the degree, betweenness, and closeness distributions excluding unknown nodes (Without Unknown, orange boxplots) to those resulting from an equal number of random known nodes (Bootstrap, blue boxplots) at each taxonomic level proved effective in identifying significant differences in network measure changes from the Original network. Using the bootstrap network results, it was clear that the impact of unknown removal to network appearance and to betweenness and closeness values of the marine and terrestrial networks was ultimately insignificant across almost all taxonomic levels, as the same effect and similar network metric values were obtained when a similar number of random knowns were removed. On the other hand, change in betweenness upon a removal of unknowns proved to be significant at genus level for the terrestrial network, as there was a significant decrease to betweenness values overall compared to no significant change in betweenness values upon removal of random knowns from the original network. Consequently, using the bootstrap values allowed us to distinguish significant results caused by unknowns from spurious results obtained from noise and node size. Bootstrap networks are thus necessary to include as a control step in the overall pipeline for investigating the role of unknowns.

When we visualized the differences in degree, betweenness, and closeness distributions among the three network types (Original, Without Unknown, and Bootstrap), the impact of the removal of unknowns varied by environment, taxonomic classification and network measure. Nevertheless, a consistent pattern emerged at deeper classification levels, where removal of unknowns had a large and more significant effect upon all network measure distributions for all

environments. For instance, removal of unknowns consistently significantly decreased overall degree and closeness scores at genus level (Fig. 2-4B) for all three environments compared to a removal of an equal number of random knowns (Bootstrap network values. In this toy example, unknowns consequently appear to be prevalent, connected and essential drivers of overall community structure. These observations suggest that removal of unknowns in a more large-scale study is most likely to produce a more significant impact at deeper taxonomic classification than at ranks like Phylum or Class. Additionally, the findings for the toy dataset justify the use of different network measures and evaluation at different taxonomic classification levels due to the varied significance of the effect of the removal of unknowns.

The evaluation of the frequency of unknowns as hubs across taxonomic levels also led to surprising, similar observations among environments. Although a small number of unknowns made up the hot springs and terrestrial networks compared to the marine network, unknowns were among the top hubs (Fig. 2-4A) at genus level for all environments. A large-scale study should therefore evaluate the frequency of unknowns as top hubs at genus level and should not neglect to evaluate networks with small proportions of unknowns as number alone does not account for whether a taxon has a significant effect on community structure.

Closeness score distribution changes upon removal of unknowns varied most among the network measures evaluated. Closeness scores significantly decreased for hot springs microbes (for all ranks from class to genus), only decreased significantly at genus level for marine microbes, and significantly increased at genus level for terrestrial microbes. In this example, the comparative increase and decrease in closeness centrality distribution is directly related to the appearance of unknowns as more central members in the hot springs community and more peripheral or equally dispersed members of the marine and terrestrial microbial communities. Large-scale studies should therefore consider the relationship between closeness and relative positioning of unknowns, keeping in mind that closeness score changes will greatly vary depending on the network and node composition of a given environment. Regardless of network measure studied or at which taxonomic level results were evaluated, the comparisons in network

measure distributions among networks with and without unknowns and without an equal number of random knowns reaffirm that the global and individual effect of unknowns is possible to be accurately measured using different network metrics and that bootstrap networks are useful and effective controls. Each environment may be impacted in different ways and at different taxonomic levels by unknowns, yet a more significant impact is most probable at deeper taxonomy. The use of hub score demonstrates that even though unknowns may make up a minority of all nodes or shared edges in an environment, certain individual unknown taxa, particularly at genus level, may be potential essential players that highly impact their respective communities and may be particularly useful targets for future characterization studies.

2.2.5 Effect of Correlation Estimation Method and Sample Prevalence Threshold on Integrated 26-Sample Network

To ensure that the results of the global and individual effect of unknowns on network metric and structure and the subsequent network relationships of unknowns would not be biased by initial network construction steps, we compared four co-occurrence estimation approaches (SpiecEASI (used previously), Sparse Correlations for Compositional data (SparCC) [80] Correlation inference for Compositional data through Lasso (CCLASSO) [175], and Pearson correlation). Both SparCC (which infers correlation by taking the linear log of Pearson correlations) and CCLasso (which infers correlation by least squares lasso regression), like SpiecEASI, were designed specifically for compositional data while Pearson correlation is a simple but effective and widely utilized correlation method. As each method estimates correlation in slightly different ways, potentially resulting in differences in correlation values among taxa, it was important to verify to which extent featured nodes, network appearance, and network-metric scores are dependent on the choice of correlation method.

Since a microbial community of 5 to 10 samples is unlikely to effectively portray the biology of a given ecosystem, we applied each of these co-occurrence estimations to the integrated 26-sample dataset at a 40 percent sample prevalence and also created networks including taxa present in all 26 samples at different sample prevalence thresholds (ranging from

20 to 45 percent prevalence, at 5 percent increments). By this manner, the integrated dataset of all five environments acted as a proxy for a larger, more realistic biological 16S rRNA dataset, enabling the most accurate validation of our overall pipeline. Using the resultant networks from the different co-occurrence estimation and sample prevalence strategies, we then re-calculated all network metrics and re-evaluated changes in the subsequent network structure, positioning of unknowns, and effect of unknown removal in hopes of clarifying any potential biases that may have been introduced by these parameters through our methodology.

2.2.5.1 Evaluation of the robustness of network analysis results using different estimation methods

We found that the correlation method used to estimate taxa co-occurrence relationships highly affected resultant network structure and topology, yet still produced similar results regarding the measured impact of unknown taxa by different network metrics. As illustrated in Figure 2-5A, though the number of nodes was constant among the different methods (147), the use of different estimation methods resulted in very different representations of co-occurrence relationships (large differences in edge size). Networks were least dense and least interconnected using the CCLasso method (which estimated only 20 edges (direct co-occurrence relationships among 147 nodes) yet most dense using Pearson and SparCC (estimating up to 1025 direct edges among nodes) (Fig. 2-5A). Such a relatively noisy network was expected for Pearson as this method ignores compositionality and sparsity and is most likely to introduce spurious edges. On the other hand, the marked difference in network topology for SparCC and CCLasso was surprising, as both methods are supposedly robust to the inherent compositionality of 16S rRNA data. However, one explanation for this observation could be that SparCC still relies upon initial Pearson correlation in estimation while CCLasso implements a strong l1 penalty during correlation inference. These results show that methods with strict penalty parameters in place (SpiecEASI and CCLasso) result in the sparsest network representations of a given environment, encouraging use of such models (in place of models based on simple Pearson correlation) in future on more complex metagenomics datasets. In particular, the sufficiently large number of

both nodes and edges found using the SpiecEASI algorithm led us to conclude that this correlation method would be the best model for estimating microbial co-occurrence for large-scale studies.

Though these networks differed in appearance and edge properties, in all four networks, unknowns featured as some of the highest-scoring hub taxa (Fig. 2-5A) at genus level, confirming that the methodology driven by network measure evaluation is robust and consistent, regardless of differences in estimating correlation or resultant edge size. The impact of unknown removal compared to an equal removal of random knowns on overall degree, betweenness, and closeness distribution (Fig. 2-5B) also followed a somewhat similar pattern to our initial results, despite a widened range of median degree, betweenness, and closeness scores among each of the correlation inference metrics. For instance, though SparCC-derived and CCLasso-derived networks had the highest and lowest degree scores overall, for each correlation metric, degree values generally significantly increased upon removal of random knowns (as can be seen by blue, Bootstrap values) and comparatively decreased upon removal of unknowns (in green). Interestingly, both betweenness and closeness centrality values tended to increase upon removal of unknowns compared to Bootstrap values, regardless of correlation metric used to infer co-occurrence, unlike the initial effects of unknowns to betweenness and closeness in hot springs, marine, and terrestrial networks. These results suggest that betweenness and closeness may be least robust to changes in co-occurrence estimation. In contrast, impact to degree distribution appears to be most consistent across different correlation metrics and serves as validation of our initial SpiecEasi-derived network results.

2.2.5.2 Evaluation of the robustness of network analysis results using different sample prevalence thresholds

The examination of the effect of sample prevalence was also useful in validating our methodology and, when applied to the toy dataset, showed, despite differences in network sparsity and prevalence of unknowns, similar patterns in terms of the local and global impact of unknowns. Sparsity increased with increased (more stringent) sample prevalence thresholds, with up to a 32 percent decrease in node size (going from 35 percent to 30 percent sample prevalence) and up to a

67 percent decrease in edge size (going from 25 percent to 20 percent sample prevalence) (Fig. 2-6A). Despite the large difference in node and edge composition among different sample prevalence thresholds, the frequency of unknowns as hubs was consistent with earlier network analysis results. For example, even at the less stringent 30 percent sample prevalence, unknown hub taxa continued to appear as top hubs at genus level (Fig. 2-6A). The most stringent sample prevalence thresholds most clearly illustrated the frequency of unknowns as top hubs and most effectively delineated predicted biological relationships among all taxa. When we examined the changes to degree, betweenness, and closeness distributions, we found that the range in degree, betweenness, and closeness scores varied greatly due to sample prevalence percentage applied, but a clear (and similar) pattern was present in regard to the impact of unknown node removal. As shown in Figure 2-6B, all network measure values increased for networks produced using the lowest (most lenient) sample prevalence thresholds at 20 percent and decreased for networks produced only including taxa present in at least 45 percent of all samples compared to the original 40 percent prevalence filter. Thus, the observations shown here suggests that while the sample prevalence threshold used significantly changes node and edge size, it should not drastically change biological interpretation of unknown ecological relevance. Nevertheless, the noisy networks at sample prevalence thresholds of 25-30 percent sample inclusion demonstrate that a higher (more stringent) sample prevalence threshold is needed to reduce noise and most accurately portray the inherent biology and relationships present within a given community.

2.3 Discussion and Conclusions

The results of each of these analyses and validations helped finalize our overall pipeline (Fig. 2-7) for unraveling the ecological relevance of unknowns. Future multi-environmental datasets of a suitably large size require evaluation of read quality upon quality filtering and OTU mapping to ensure that samples are comparable. After this check is completed and OTUs are recovered by the open-reference clustering approach, unknowns can be identified by finding all OTUs designated as unassigned, ambiguous, or uncultured at any taxonomic level. To ensure that both knowns and unknowns merit inclusion in networks, the prevalence and shared abundance of

taxa across different samples or environments must be evaluated. Prevalence curves should be used to find the most appropriate sample prevalence threshold for a given environment. To ensure that no taxa are over-represented in any given sample or environment type, the OTU abundance across samples should be assessed using heatmaps. After these initial checks of the data, a stringent and suitable sample prevalence threshold should be applied to remove rare, infrequent taxa and a method designed for sparse, compositional data, ideally with strong penalty parameters, like the SpiecEASI neighborhood algorithm, should be used to estimate taxa co-occurrence relationships. After estimation, networks can be constructed including and excluding unknowns and the individual impact of unknowns can be measured by calculating the hub score of each node. To determine the overall impact of unknowns, differences in score distribution for the network measures degree, betweenness, and closeness should be compared among networks with and without unknowns (among "Original" and "Without Unknown" networks) must be conducted, using a removal of randomly selected knowns (a "Bootstrap network") as a control for node size. Lastly, the hub score metric can be used to determine the frequency of unknowns as top-scoring hubs to identify unknown taxa that are particularly relevant for community structure.

Using a 26-sample toy dataset of 5 distinct environments (freshwater, hot springs, marine, plant, and terrestrial), we successfully developed a methodology for recovering and identifying particularly relevant unknowns from diverse 16S rRNA data, evaluating and controlling for parameters that may bias results, like choice of sample location, size, quality filtering threshold, and taxonomic assignment method. Unknown taxa were shown to merit inclusion in networks due to their similar prevalence as known taxa. Degree centrality, betweenness centrality, closeness centrality, and hub score were found to be useful in quantifying the global and individual relevance of unknowns across taxonomic classification ranks and environments.

Unknowns were found to act as top hubs in marine, hot springs, and terrestrial networks, despite low presence of unknowns among the two latter networks, suggesting that the importance of unknowns is not dependent on overall abundance. The application of bootstrap networks as a

control for the effect of node size upon network and network measure properties was evaluated and found to be useful in distinguishing which results were significant and solely due to the removal of unknowns. Lastly, validation checks of the effect of sample prevalence filters and different manners of estimating co-occurrence relationships revealed that the pipeline and subsequent results following network metric evaluation were robust and mostly consistent, regardless of sample prevalence or network estimation parameter changes. Based on our results, future network analyses should incorporate a stringent sample prevalence threshold and a correlation estimation method like SpiecEASI to most accurately model microbial co-occurrence relationships while bypassing noise or spurious associations.

Consistent observations of unknowns as important hubs and drivers of community interconnectedness across different correlation estimation methods, sample prevalence thresholds, and diverse environments suggest that the pipeline is robust and applicable for future, large-scale datasets. Furthermore, the results of the network pipeline on the toy dataset demonstrate not only the utility of network theory for microbiome analysis but also the importance of including and quantifying the contributions of unknown microorganisms in future analyses. Future work will consist of applying this network-driven approach to a multi-environmental dataset of at least 20 publicly available studies.

2.4 Methods

2.4.1 Data Retrieval and Sample Preprocessing

All raw data were retrieved from NCBI and converted to fastq format using NCBI SRA Toolkit (http://ncbi.github.io/sra-tools/). Quality filtering and preprocessing of all sequence data was performed through the split_librairies_fastq.py script from the Quantitative Insights into Microbial Ecology (QIIME) pipeline (Version 1.9.1) [176] using a Phred quality threshold of 19. All sequences passing quality filtering were clustered at 97 percent sequence similarity and classified to OTUs using the SILVA (v128) SSU reference database by the script called pick_open_reference_otus.py. All singletons were discarded. Lastly, the script filter_taxa_from_otu_table.py was used to remove any OTUs related to Archaea, mitochondria, or

chloroplast so that the analysis would be targeted to the Bacteria. All subsequent statistical and network analyses were conducted in R (v 3.5.1)

2.4.2 Identification Strategy for Unknown Taxa

At each taxonomic level, from phylum to genus, unknown taxa were identified as any OTU taxonomically assigned as: “uncultured”, “uncultured bacterium”, “Unknown”, “Unassigned”, “Ambiguous taxa” or “NA” by the open-reference picking strategy. These OTUs were renamed as “Unknown” for all subsequent analyses and comparisons. An OTU was only labeled as “Unknown” at the specific taxonomic classification level at which it could not be taxonomically assigned beyond the higher classification descriptors. For example, if an OTU had a known order but unknown family description assigned by the reference database, it would only be designated as an unknown in the network analysis for family and genus taxonomic classification and, at all higher classification ranks, would be referred to its known classification status.

2.4.3 Network Creation

To create networks of each environment, OTU biom tables and corresponding mapping information (i.e. sample ID, geographic location, longitude, and latitude) were imported into R using the package phyloseq (version 1.28.0) [177]. The filterTaxonMatrix() function from phyloseq was used to apply filtering criteria. For each environment, the filtered OTU phyloseq object was normalized, transformed, and converted into an adjacency matrix based on covariance by the spiec.easi() function from the package SpiecEasi (version 1.0.7) [169], using Meinshausen-Buhlmann’s neighborhood selection (method = ”mb” parameter) to estimate the conditional dependence of each pair of OTUs, particularly designed for sparse, compositional amplicon, and metagenomic data and, least likely to introduce spurious, indirect relationships from being included in the networks [149]. The Stability Approach to Regularization Selection (StARS) method was used to find the optimal sparsity parameter, with the StARS variability (i.e., minimum lambda) threshold set to 0.05 for all networks. Each adjacency matrix was then converted into an igraph object and visualized as a network using the adj2igraph() and plot_network() functions from SpiecEasi. Networks were created and visualized at each

taxonomic classification level, using the function plot_network() of phyloseq, with nodes representing OTUs and edges representing direct co-occurrence relationships between OTUs. Each resulting network contained at least 100 nodes (i.e., OTUs) per environment and edge-to-node ratios that varied from 1.9 and 2.9 (Table 1).

To create reduced networks, either excluding unknown OTUs ('Without Unknown' network) or randomly selected known OTUs ('Bootstrap' network), the delete_vertices() function (package igraph (version 1.2.5)) was used. To create bootstrap networks, known nodes were randomly selected by the sample()function, with the x parameter equal to the total number of OTUs, known and unknown, at that specific taxonomic classification level for the chosen environmental network and size parameter equivalent to the number of unknown OTUs for that network. Networks were created using the same SpiecEasi pipeline as described above. These randomly reduced networks were created 100 times for each taxonomic classification level from phylum to genus for each of the four target environments (i.e., hot springs, hypersaline, deep sea, and polar habitats).

2.4.4 Network Analysis Strategy

For each environment, at each taxonomic level from phylum to genus, network-level and node-level measures of networks including unknown OTUs (i.e., "Original"), excluding them (i.e., "Without Unknown"), and excluding an equal number of randomly selected known OTUs (i.e., "Bootstrap") were evaluated and compared against each other and visualized as boxplots. The network measures evaluated for all nodes present within each network were degree, closeness, betweenness, and hub score and were calculated by using the igraph package functions: degree(), betweenness(), closeness(), and hub_score(). Wilcox test was used to evaluate the statistical significance of changes in degree, betweenness, and closeness between the three network types using the stat_compare_means() and compare_means() functions from the ggpubr (version 0.3.0) package. P-values were adjusted using the Holm method [178] and boxplots were created using the package ggplot2 (version 3.3.0).

2.4.5 Sample Criteria Validation

To account for the possible confounding effect of sample size, networks were reconstructed and reanalyzed using the complete dataset (all 26 samples) and using a range of different percentages (from 20 percent to 45 percent at 5 percent increments) of samples in the initial filtering process of the OTU table, discarding any taxa that did not meet this sample percentage threshold from being included in the networks. Networks excluding unknown and random known nodes were constructed using the same methods as described previously and network measures were recalculated and visualized as boxplots to determine the effect of different sample size criteria and its statistical significance.

2.4.6 Network Tool Validation

To determine whether other measurements of species' co-occurrence relationships changed our final network analysis results, three other methods were used to calculate the relationship between pairs of species in R: SparCC [80], CCLasso [175], and Pearson correlation. The R rcorr()function, with the parameter type set to "pearson" from the package Hmisc (version 4.4-0) was used to calculate Pearson correlation. The R cclasso()function (https://github.com/huayingfang/CCLasso/blob/master/R/cclasso.R) was used to calculate CCLasso correlation, and the adapted sparcc()function in the SpiecEasi package was used to calculate SparCC correlation. Subsequent networks were created using igraph. Boxplots were created to compare the median network measure scores of degree, betweenness, and closeness for the three network types at all classification levels for the four (SpiecEasi, SparCC, CCLasso, and Pearson) correlation networks.

Table 2-1. Evaluation of sample prevalence threshold filtering on node, edge, and network properties

Environment	Sample Prevalence Threshold (%)	Node Size	Fraction of Data	Edge Size	Edge Breakdown (K-K, K-U, U-U Edges)	Density (Edge-to-Node Ratio)
Hot Springs	40	234	0.020	472	304,142,26	2.02
Marine	60	218	0.035	551	120,183,248	2.58
Terrestrial	60	143	0.029	208	157,41,10	1.45

K-K: Known-Known; K-U: Known-Unknown; U-U: Unknown-Unknown

Table 2-2. Evaluation of network property changes upon unknown removal (marine network as example)

Taxonomic Level	Node Size (%)	Mean Degree (%)	Mean Betweenness (%)	Mean Closeness (%)
Phylum	0.0	-20.3	+21.0	-10.0
Class	-0.5	-20.3	+13.8	-60.0
Order	-1.8	-22.5	+13.1	-50.0
Family	-1.8	-22.5	+13.1	-50.0
Genus	-61.0	-36.4	+8.4	-60.0

Plus sign: Increase; Minus sign: Decrease (Upon Removal of Unknown Nodes)

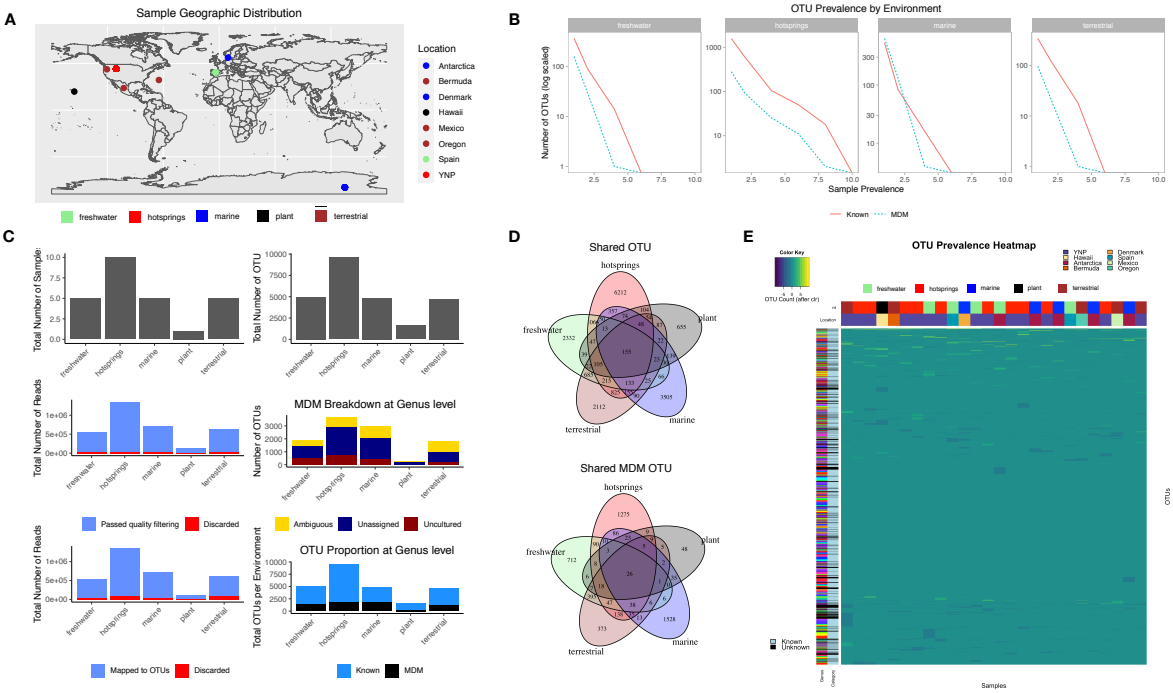


Figure 2-1. Initial data quality evaluation across environments.a. Map of sample geographic distribution. Circles symbolize sample location and are colored by environment. b.Prevalence line curves of OTU (log-scale) abundance resulting from changing sample prevalence. Solid, red lines signify known OTUs and dotted, blue lines signify unknown OTUs. c. Barplots of total samples, reads, OTUs, and MDM OTUs across environments. Left-hand bar plots show, from top to bottom, the total number of samples, total number of reads passing (blue) and failing (orange) quality filtering, and total number of reads able (blue) and unable (orange) to be mapped to OTUs. Right-hand bar plots show,from top to bottom, the total number of OTUs, the proportion of known (blue) to unknown (black) OTUs, and the breakdown of unknown OTUs as ambiguous (yellow), unassigned (blue), and uncultured (in red). d. Venn diagrams of the shared total and unknown OTUs between environments. e. OTU Prevalence Heatmap. OTU counts are shown after centered-log-ratio (clr) transformation, with blue indicating low and yellow high values respectively. Rows are samples, colored by environment and location type. Columns are OTUs, colored by genus and unknown status.

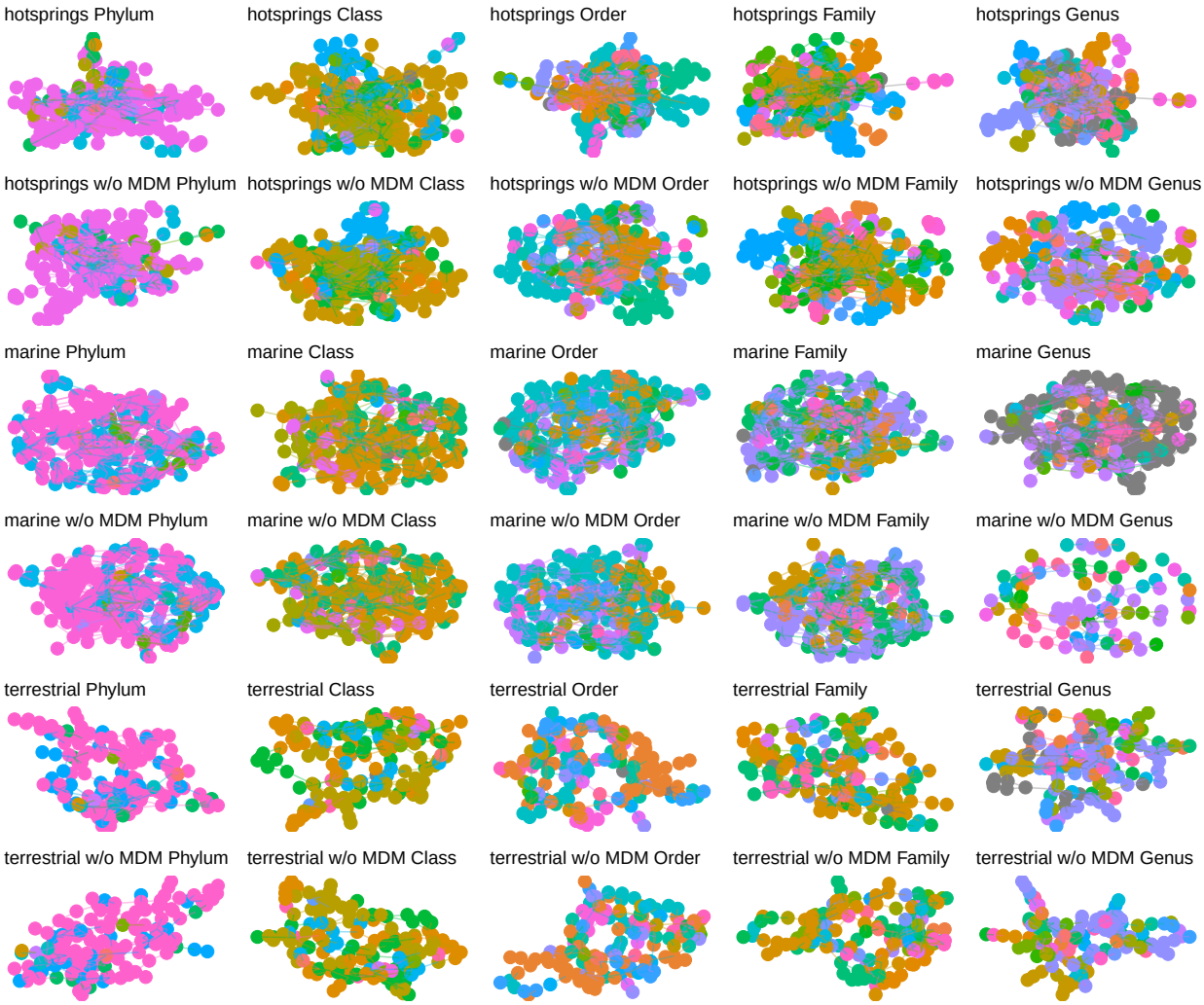


Figure 2-2. Microbial co-occurrence networks visualized across classification levels and environments. Rows 1 and 2 show hot springs networks, rows 3 and 4 marine networks, and rows 5 and 6 terrestrial networks with and without unknowns while columns show each network from phylum to genus, with nodes colored by respective taxonomic classification. Odd rows show networks including both known and unknown (colored in gray) taxa while even rows show networks with unknown taxa removed at each classification level. For each network with and without unknowns, node and edge size are indicated.

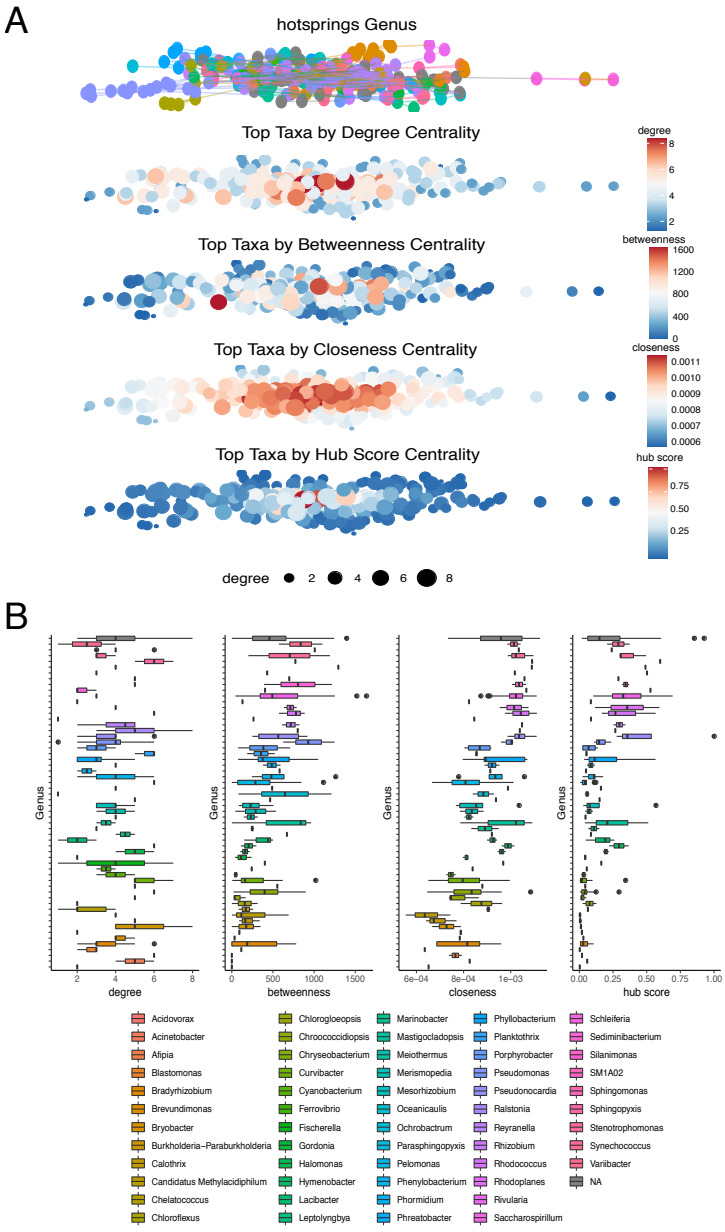


Figure 2-3. Network measure visualization and comparison on the selection of the most important taxa. a. The most important taxa for the hot springs genus network calculated by degree, betweenness, closeness, and hub score metrics. From top to bottom are the original hot springs genus network and networks recolored by degree, betweenness, closeness, and hub score and sized as a function of degree, with higher degree indicating larger node size. For all metric models,nodes are colored using a blue-white-red color scale (blue:low, red:high values). b. Boxplots of the genus score distribution of different network measures. From left to right are shown the distribution of degree scores, betweenness scores, closeness scores, and hub scores for each genus, with unknowns at genus colored in gray.

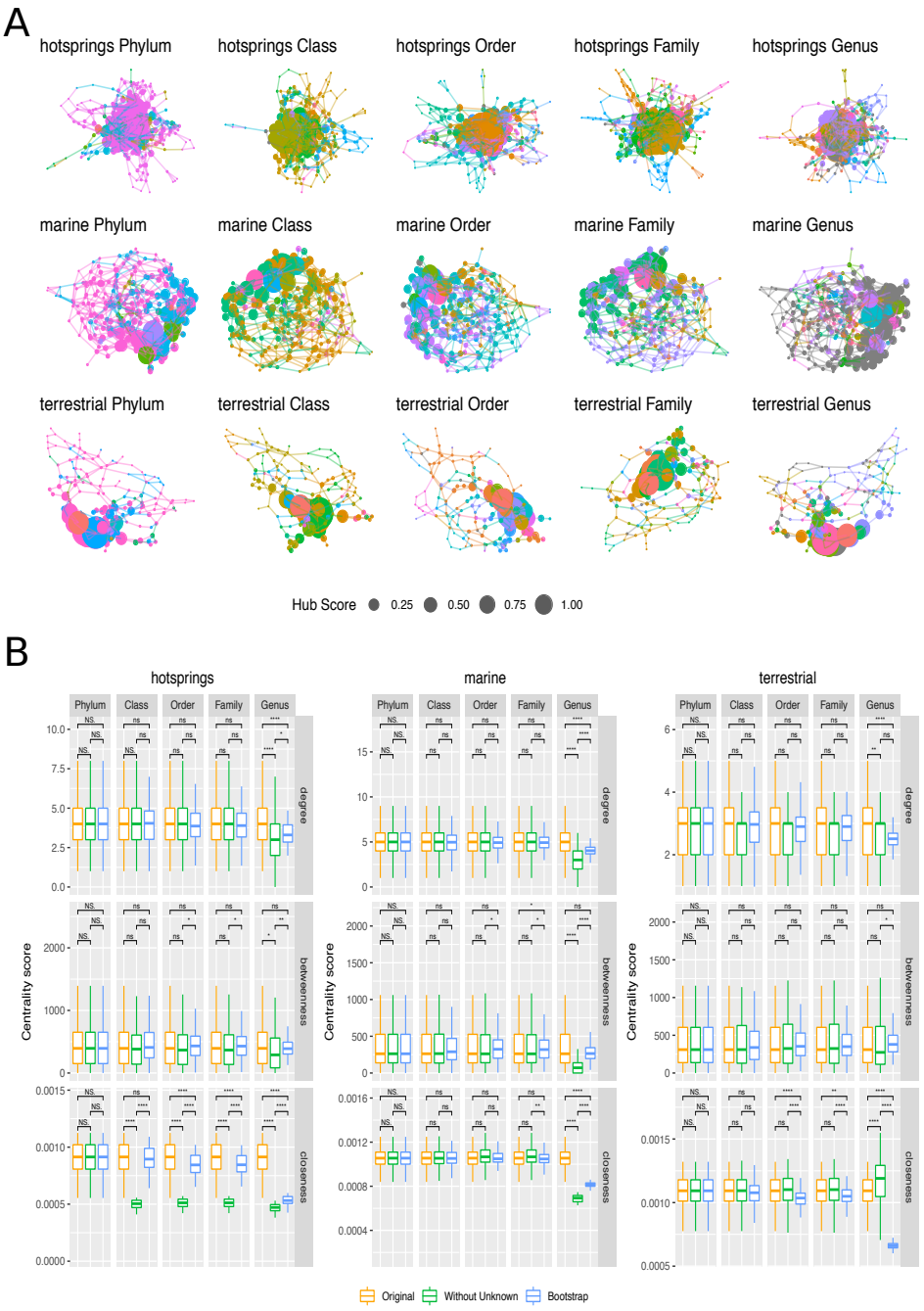


Figure 2-4. Local and global impact of unknowns to network metrics across environments. a. Environmental Networks at genus level, with nodes resized as a function of hub score. Nodes are colored by genus classification, with nodes in gray representing unknowns at genus level. b. Boxplots of degree, betweenness, and closeness centrality values of nodes present in hot springs, marine, and terrestrial networks at different taxonomic levels (left to right- phylum to genus). Wilcoxon pairwise comparisons were used to assess significance between the three network types (Original-Without Unknown, Without Unknown-Bootstrap, Original-Bootstrap) for each taxonomic level. For each comparison, p-values after Holm adjustment are depicted as stars (significant) or ns (not significant).

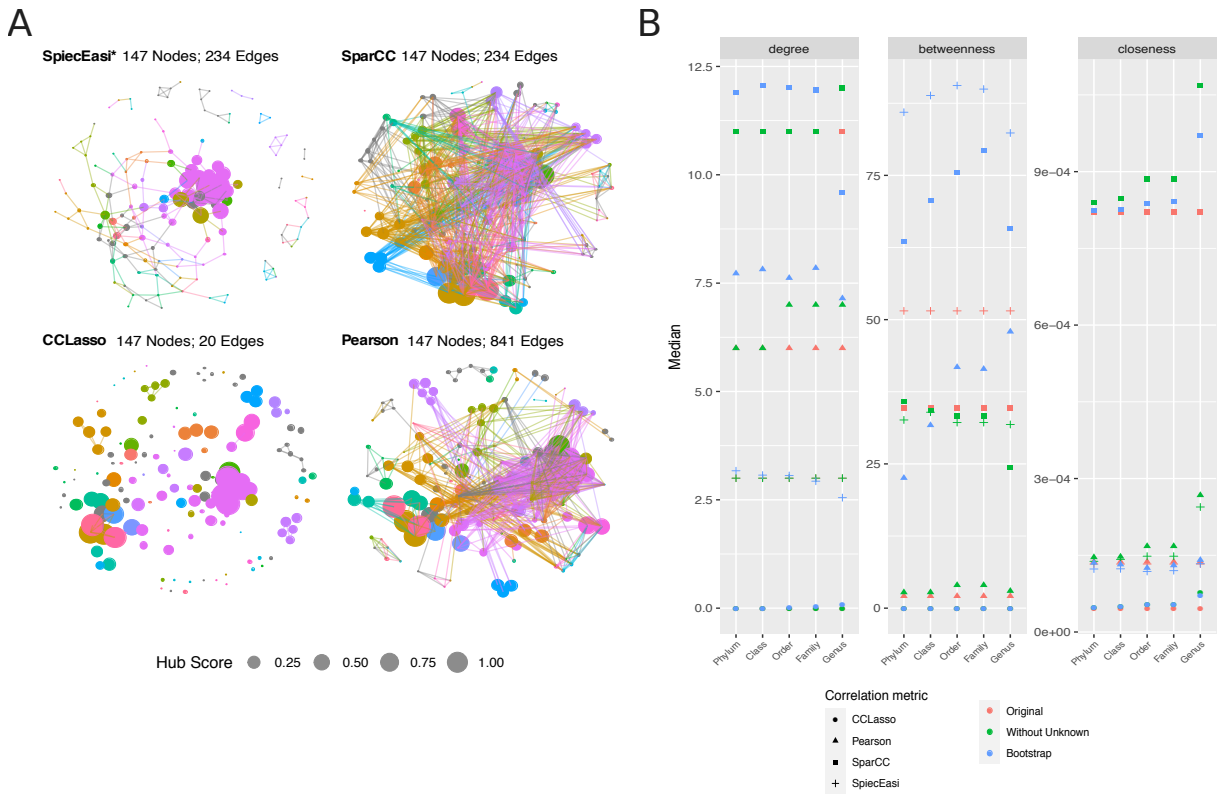


Figure 2-5. Effect of correlation estimation method on network results. a. Hub-score networks for the integrated dataset created by the SpiecEASI MB, SparCC, CCLasso, and Pearson methods. Nodes are colored by genus classification, with unknowns represented in gray, and sized as a function of hub score. Asterisk (*) denotes network used in initial network analysis . b. Boxplots of degree, betweenness, and closeness centrality values of nodes at different taxonomic levels (left to right-phylum to genus) for each of the correlation metrics. Network types (Original, Without Unknown, and Bootstrap) are outlined in red, green, and blue respectively. Shapes correspond to correlation metric applied. Wilcoxon pairwise comparisons were used to assess significance between the three network types (Original-Without Unknown, Without Unknown-Bootstrap, Original-Bootstrap) for each taxonomic level. For each comparison, p-values after Holm adjustment are depicted as stars (significant) or ns (not significant).

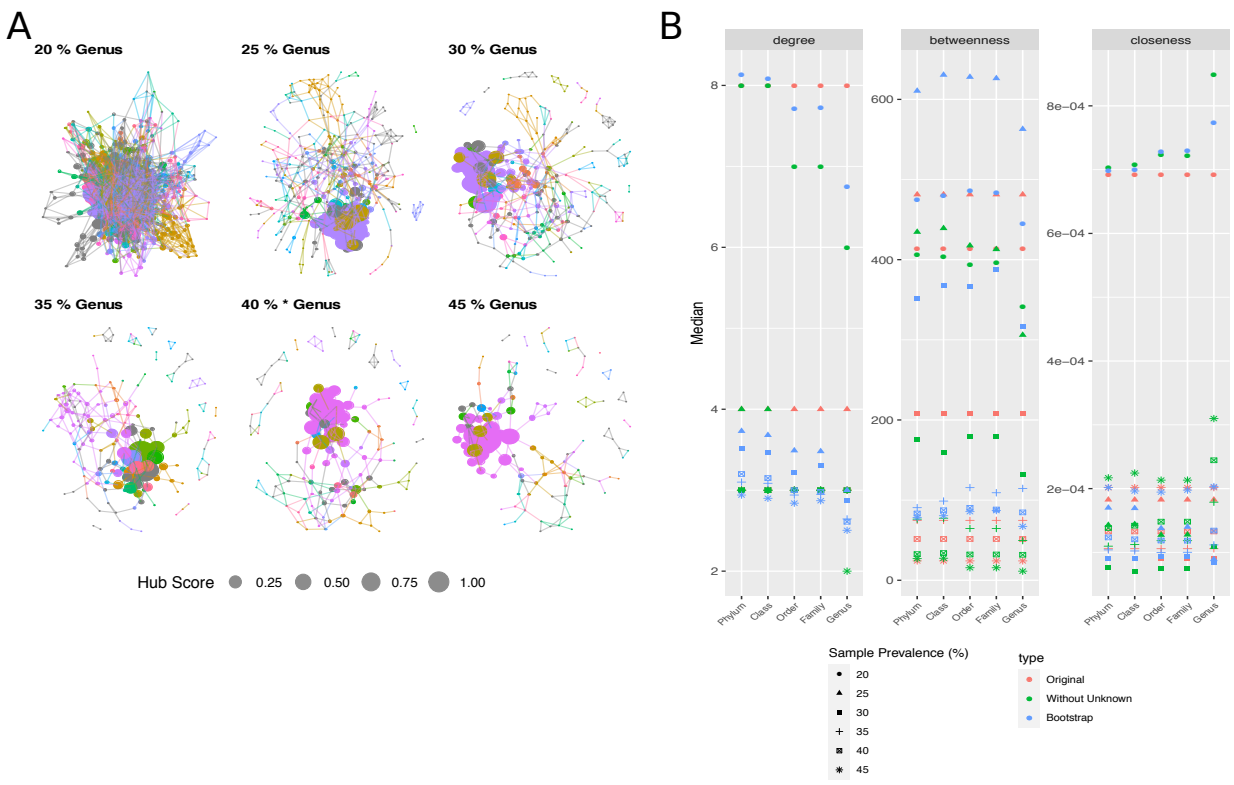


Figure 2-6. Effect of sample prevalence threshold on network results.a. Hub-score networks for the integrated dataset across sample prevalence thresholds, from 20 to 45 % prevalence. Nodes are colored by genus classification, with unknowns represented in gray, and sized as a function of hub score. Asterisk (*) denotes network used in initial analysis. d. Boxplots of degree, betweenness, and closeness centrality values of nodes at different taxonomic levels (left to right-phylum to genus) across sample prevalence thresholds. Network types (Original, Without Unknown, and Bootstrap) are outlined in red, green, and blue respectively. Shapes correspond to sample prevalence percentage applied. Wilcoxon pairwise comparisons were used to assess significance between the three network types (Original-Without Unknown, Without Unknown-Bootstrap, Original-Bootstrap) for each taxonomic level. For each comparison, p-values after Holm adjustment are depicted as stars (significant) or ns (not significant).

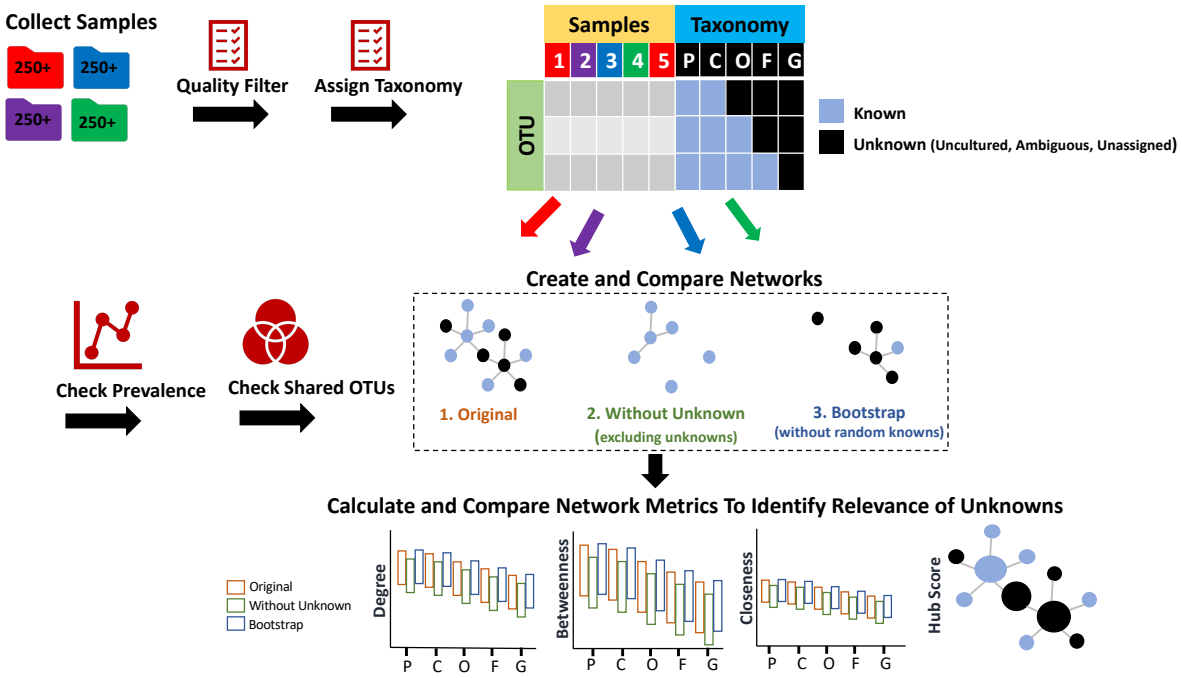


Figure 2-7. Complete pipeline to assess ecological relevance of unknowns. All validation steps (quality filter, assign taxonomy, check prevalence, and check shared OTUs) are highlighted in red graphics.

CHAPTER 3
A NETWORK APPROACH TO ELUCIDATE AND PRIORITIZE MICROBIAL DARK MATTER IN MICROBIAL COMMUNITIES

3.1 Introduction

*For billions of years, microbes and their metabolic activities have been shaping Earth's physical, chemical and mineralogical landscape. Although microbes comprise the majority of the planet's biomass, most microbial species and their genomes remain uncharacterized [37, 63]. These unknown aspects of microbial life, colloquially called 'microbial dark matter' [179, 37], represent a fundamental impediment to microbial ecology, as microbe-dominated ecosystems cannot be reliably characterized without a thorough understanding of the roles microbes and their gene products play in ecosystem processes.

Currently, most of our knowledge of the microbial world is skewed by a few taxa that lend themselves to cultivation and genetic manipulation. Of the cultivated species, 88% are derived from just four phyla (Proteobacteria, Firmicutes, Actinobacteria, and Bacteroidetes) [37]. The uncultured and unsequenced microbial majority on Earth likely represents major evolutionary lines of descent within the tree of life and is expected to have played key roles in ecosystem formation, evolution and function. Without a mechanistic approach to characterize the roles of 'microbial dark matter' in the ecosystem, we will not have a full understanding of how these organisms impact their neighbors, environment, or life as a whole.

Recent efforts to provide insight into the uncharacterized and uncultured majority through next-generation sequencing technologies have significantly expanded the microbial tree of life [59, 61, 180]. Yet, despite the recent explosion of nucleic acid sequencing of microbial environments, much remains to be learned [181, 182]. Truly understanding the ecological role of unknown taxa within communities requires a more comprehensive assessment of why unknown taxa persist, or with whom they interact on a global scale. More importantly, it is unclear whether the presence of these unknown organisms confers a value that is not already provided by other, well-characterized microbes within the same ecosystem.

*Zamkovaya, T., Foster, J.S., de Crécy-Lagard, V., Conesa, A. A network approach to elucidate and prioritize microbial dark matter in microbial communities. ISME J (2020). https://doi.org/10.1038/s41396-020-00777-x

To more fully understand microbial life, particularly the contributions of unknown taxa, it is first critical to understand the connectivity and structure of microbial ecosystems. Networks have long been used as analytical tools to better understand species' roles and interactions [143, 183, 80, 184, 88, 130]. By mathematically modeling a microbial community as a network, where nodes are different species and edges represent their relationships [148, 143], researchers can depict species interactions and study the structure of the environmental system. Network metrics, such as hub score, betweenness centrality, closeness centrality, and degree centrality [142, 149, 143], can be used to quantitatively describe these communities and pinpoint the most important taxa of a given environment, thereby providing essential clues about how specific taxa or gene products may contribute to ecosystem functioning [185]. While many advances have been made in microbial ecology using network-driven approaches [186, 187, 188, 189, 144, 190, 191, 192, 193], few, if any, revolve around unknown taxa. Most network analyses focus on the role of known species, usually excluding any taxonomically unassigned or ambiguous taxa in early filtering steps. If included, any unassigned taxa are only briefly mentioned in connection to their interaction with more characterized, abundant phyla, leaving the role of unknown and uncultured taxa largely unexplored.

Here, we present a network-based approach to determine the ecological relevance of unknown taxa within a targeted community. To provide both a broad and accurate interpretation of results, a comprehensive dataset encompassing four different aquatic environments was compiled. Networks were constructed with and without the unknown taxa and changes in the network metrics degree, betweenness, and closeness were evaluated. In this manner, a systematic identification of previously undetected biological patterns of unknown taxa was possible, and the contribution unknown taxa provide to the overall community structure was evaluated and compared across taxonomic levels. To identify the most important unknown components of each environment, the hub score of all nodes was calculated and the frequency of unknowns among the top hubs was noted. These hub network analyses provide a means to prioritize unknown hubs for future characterization efforts. We demonstrate one of several possible applications of this

approach, using particularly relevant hub unknowns as probes to detect novel adaptation-related genes within metagenome scaffolds. The application of network theory to identify and prioritize key unknown microbial members may thereby help shed light on potential adaptation mechanisms of successful unknown taxa while enabling a more comprehensive understanding of ecosystem structure under a diverse range of environmental conditions.

3.2 Results

3.2.1 Overall Strategy to Detect the Relevance of Unknown Taxa

A pipeline based on network analysis was developed to detect and quantitatively measure the overall and individual impact of unknown taxa on their environmental communities (Fig. 3-1). Briefly, over 250 Illumina 16S rRNA sequencing fastq files belonging to four distinctive aquatic extreme environments (i.e., hot springs; hypersaline; deep sea; and polar habitats that included both Arctic and Antarctic samples) were collected from public repository databases and 45 different BioProjects (Fig. 3-2A, Supplementary Dataset S1). We included different environment types to assess general and environment-specific results and chose to use extreme habitats as they comprise some of the harshest and relatively understudied habitats on Earth, and therefore, are likely to contain uncharacterized organisms.

Object 3-1. Link for Supplementary Datasets S1 and S2

After quality filtering, reads were mapped to operational taxonomic units (OTUs) and annotated against the SILVA (v128) reference database [194] by an open-reference strategy, i.e. allowing the detection of unknown OTUs [176]. Over two million known and novel taxa were observed. Only those taxa associated with the domain Bacteria or unclassified at the domain-level were targeted for downstream network analyses to demonstrate the feasibility of this approach across ecosystems. Unknown taxa, labeled as uncultured, unassigned, or ambiguous by the reference database, were identified and grouped as ‘microbial dark matter’ at each taxonomic classification level (e.g. phylum to genus). For each environment, networks reflecting across-samples co-occurrence relationships between all taxa, known and unknown, were

constructed and referred to as the "Original" networks (Fig. 3-1). To assess the role of the unknown taxa on network structure and properties, the unknown nodes were removed from the 'Original' network and a new network, referred to as 'Without Unknown' was reconstructed. To ensure that changes in network properties were not caused just by the number of nodes, a third network, referred to as the 'Bootstrap' network, was created where a random set of nodes of the same number as the unknown OTUs was removed. The relevance of the unknown taxa was assessed by comparing changes in degree, closeness, and betweenness scores between the three network types and by evaluating the frequency of unknowns as top hubs within each of the 'Original' environmental networks.

3.2.2 A Similar and Significant Fraction of Unknown Taxa Populates Diverse Environments

To assess whether there were distinctive patterns or trends of unknown taxa within the four targeted environments, data from each habitat type were mined from several geographical locations across the globe (Fig. 3-2A). Reads were collected from the online repositories National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) and Joint Genome Institute Genomes Online Database (JGI Gold). For each environment, between 255 to 286 16S rRNA samples and between 51 million and 57 million reads were included in the analysis, resulting in a total of 219,980,340 reads from 1,086 samples (Fig. 3-2B).

After processing, quality filtering, and OTU assignment steps were completed, there were 2,102,595 unique OTUs totaling 164,896,127 bacterial read counts derived from these samples (Fig. 3-2B). The relative proportions of unknown OTUs, which were designated as unassigned, ambiguous, or uncultured by the SILVA database, were compared between environments. Results indicated that all environments showed similar relative contributions of these three unknown types, with unassigned and uncultured OTUs making up the majority of the unknown component composition (Fig. 3-2C). Regardless of environment type, within each of the four habitats, between 25 and 38% of unique OTUs were cataloged as unknown (Fig. 3-2B and 3-2D). Samples collected from polar habitats were significantly enriched (Fishers Exact Test p-value < 0.05) in

unknown OTUs despite having the highest total read counts. The higher proportion of unknown OTUs in the polar samples compared to the other habitats likely reflects the less-well characterized biological diversity of these Arctic and Antarctic ecosystems.

Next, the proportion of shared known and unknown OTUs between environments was evaluated. Most OTUs, regardless of assigned or unassigned taxonomic status, were environment-specific, with only 11,318 out of the 2,102,595 OTUs present in all four of the environments (Fig. 3-2D). The majority of shared OTUs were observed between the hypersaline and polar environments and between hypersaline and hot springs environments, possibly reflecting the widespread distribution of hypersaline habitats across diverse thermal zones. Unsurprisingly, polar and hot springs environmental samples shared the least number of OTUs (Fig. 3-2D). Given this low common OTU pool across environments, network analyses were applied to each environment independently.

Lastly, the prevalence (i.e. the percentage of samples with no-zero counts) of known and unknown OTUs was evaluated at the genus level to assess the consistency of OTU detection within each environment. The OTU matrix was sparse, with the majority of taxa observed in < 50 (i.e., ~ 20 %) samples (Fig. 3-2E). However, prevalence curves were generally very similar for known and unknown OTUs in all four environments, indicating that unknown OTUs are not necessarily rarer than already characterized species (Fig. 3-2E). Moreover, we confirmed that unknown OTUs, like known OTUs, were generally present and consistent across multiple studies within the same environment and did not tend to concentrate in any particular project (Supplementary Fig. S1-4).

Object 3-2. Link for Supplementary Figs. S1-30

Consequently, these results indicated that a network analysis of these data would be a reflection of the co-occurrence structure of the community and not of potential compositional bias.

3.2.3 Network Analysis of OTU Abundance at Different Taxonomic Levels Reveals the Connectivity of Unknown Microbes

Having demonstrated that the unknown taxa comprise a substantial proportion of unique OTUs and have comparable abundances to known taxa within a community, network metrics were used to effectively compare the ecological relevance of both known and unknown taxa in subsequent networks. Microbial association networks were constructed that featured only significant co-occurrence correlation relationships for OTUs with a notable prevalence in each environment, meaning that any OTUs that were not detected in a sufficient number of samples were removed. To select a suitable prevalence threshold, the proportion of known and unknown taxa across a range of sample percentages was evaluated (Supplemental Fig. S5). Across all taxonomic levels, the known and unknown taxa of hot springs and polar habitats were more prevalent than those of hypersaline and deep sea communities; therefore, a slightly more stringent prevalence threshold (40%) was chosen for the former, and a lower threshold value (30%) chosen for the latter environments. These thresholds resulted in the retention of a similar fraction of data from the initial OTU count (Table 3-1), with 102 to 297 nodes present per environment, making the networks both suitably large and comparable.

For each of the environments, microbial co-occurrence networks were constructed using the SpiecEasi Meinshausen-Buhlmann (MB) neighborhood algorithm, which estimates the conditional independence between any given pair of OTUs [169]. This approach is robust and ideal for sparse, compositional amplicon and metagenomic data and, unlike other correlation methods, prevents spurious, indirect relationships from being included in the networks [169, 195]. The Stability Approach to Regularization Selection (StARS) method was used to find the optimal sparsity parameter, with the StARS variability (i.e., minimum) threshold set to 0.05 for all networks [196]. Each resulting network contained at least 100 nodes (i.e., OTUs) per environment and edge-to-node ratios that varied from 1.9 and 2.9 (Table 3-1).

Although most OTUs that passed the prevalence criteria became elements of the networks, no relationship between the initial data size (i.e., number of samples and taxa) and the

interconnectedness (i.e., nodes/edges ratio) of the resulting network was observed (Table 3-1). The two environments with the highest number of initial OTUs (i.e., hypersaline and deep sea) had the lowest number of prevalent members, 193 and 102 respectively, and very different edge/node ratios, 2.7 and 1.9 respectively, indicating that high prevalence does not necessarily correlate with high co-occurrence. Similarly, the polar and hot springs networks retained a high number of prevalent OTUs but differed in edge number, yet again, indicating that the observed network structures were the result of the intrinsic properties of each environment and were not dependent on the sampling procedure. Interestingly, when evaluating the proportion of unknown edges and Unknown-Unknown connections at the genus-level, similar patterns were observed across environments. Between 45 to 62% of all connected nodes were unknown OTUs and a higher proportion of Unknown-Unknown versus Known-Known links was present at the genus-level (Table 3-1).

The results of this global analysis of network construction and composition demonstrate that although the general community connectivity might be environment-specific, the relative contribution of known and unknown taxa to these networks is similar. Once again, network properties were not a direct outcome of sampling biases, but more likely, reflect the biology of their respective ecosystems.

3.2.4 Unknown Taxa Play Important Roles in Interconnectedness and Connectivity of Extreme Environmental Microbial Networks

Next, the position and neighborhood of unknown nodes were examined. At the phylum-level, unknown taxa were present in the hot spring, hypersaline and polar environments, but were not found in the deep sea network (Supplemental Fig. S6). The class-level was the first taxonomic classification rank in which unknown taxa were present in all environmental networks. To accurately assess the role of unknowns, we evaluated class-level networks and observed that the hypersaline and polar unknown OTUs created distinctive clusters, whereas hot springs and deep sea unknown OTUs were intermixed with known taxa (Fig. 3A). Hypersaline and polar unknowns consistently appeared to be isolated and peripherally located compared to the centrally

positioned hot springs and deep sea unknown nodes across almost all classification ranks (Supplementary Fig. S6). Consequently, these results suggest that the clustering pattern is unrelated to a higher abundance of unknowns and is more environment-dependent. For example, the targeted hot spring environments had the highest number of unknown OTUs, yet showed the most dispersed connections between the unknown taxa (Fig. 3A). Thus, the inclusion of the unknown taxa in our environmental networks models was, as anticipated, not solely the consequence of their level of prevalence, but rather a reflection of a particular abundance co-variation pattern.

For all four environments, unknown OTUs had more frequently shared edges among them than with classified taxa (Fig. 3-3B). Unknown OTUs were found to frequently co-occur with each other within each environmental network, although the frequency of within-class interactions for unknowns at the class-level was found to be statistically no greater than all other within-class interactions for each environment (Supplementary Fig. S7). In accordance with other studies, these results demonstrate that OTUs of the same taxonomic classification most frequently co-occur with each other [197, 198]. Furthermore, the high frequency of shared edges between unknown classes suggests that unknown OTUs might be taxonomically related.

To ensure that the observations found were reproducible, robust, and not biased by earlier steps of the analysis, the diversity and position of unknown taxa in the networks were examined for several parameters. Although the number of unknown nodes changed at each level (Supplementary Fig. S6), the environment-specific network patterns observed at the class-level (Fig. 3-3A) were retained at other taxonomic levels. Additionally, to determine whether the topology of the network was a direct consequence of our correlation metric of choice or the prevalence threshold, three other network construction approaches were used: SparCC [80], CClasso [175], and Pearson correlation. Network analyses were performed across a range of prevalence thresholds (15% to 35%, at 5% increments). Again, regardless of the network construction approach or sample percentage applied, network shape and unknown OTU position remained consistent and each environment exhibited a distinctive pattern of unknown taxa

inclusion. For example, unknown nodes continued to occupy peripheral positions for hypersaline and polar networks, whereas nodes in hot springs and deep sea environments were more centrally located when applying different correlation metrics (Supplementary Fig. S8). Although networks appeared "noisy" at more lenient prevalence thresholds (15 - 20%; Supplementary Fig. S9 - S12), the networks and positioning of unknowns at higher percent thresholds were similar in appearance to the "Original" networks for all environments. Based on these results, we found that our general analysis strategy was robust across parameter choices and, therefore, these networks captured critical features of the relationships among taxa within each distinctive environment.

3.2.5 Microbial Dark Matter Acts as Unifiers and Frequent Hubs Within Extreme Environmental Networks

Although these results suggest that unknown taxa were highly interconnected, these observations did not reveal how the presence of unknowns affected the overall community structure. To more fully understand this role, we analyzed how network properties changed in the presence and absence of unknown OTU nodes. We evaluated changes in degree, betweenness, and closeness, as different network metrics reveal different aspects of the relevance of nodes within their networks. This approach has the potential to ascertain whether certain unknown OTUs were more centrally positioned (e.g., due to high closeness scores), more essential for joining other taxa (e.g., high betweenness), or simply more prevalent and likely to co-occur with others (e.g., high degree). To control for the effect of node removal and distinguish effects of unknown taxa from network size, networks were generated that excluded several randomly picked known nodes equal to the number of unknown OTUs. This process was repeated 100 times to create a distribution of "Null" or "Bootstrap" networks for statistical comparisons. Differences in network parameters between networks without unknown OTUs and the "Original" or "Bootstrap" networks were determined by the Wilcoxon test and p-values were adjusted using the Holm method [178]. Strikingly, removal of the unknown taxa caused a statistically significant impact on all measured network metrics in all four studied environments (Table 3-2; Fig. 3-4; Supplementary Fig. S13 – 23). In the polar environments, for example, removal of unknown

OTUs caused a significant decrease in degree (p-value $<$ 1E-5) and betweenness (p-value $<$ 1E-4) scores and a significant increase in closeness (p-value $<$ 2.22E-16) scores across multiple taxonomic levels (Fig. 3-4), suggesting that the unknown taxa in the polar environments are critical to the community structure and preserve local connections. Closeness score was the only parameter that behaved differently between environments (Table 3-2), increasing upon exclusion of the unknowns for hypersaline and polar networks but decreasing at family-level for the deep sea environment, due to the different topology of unknown OTUs across environmental networks. For example, the removal of centrally located unknowns at the family-level in the deep sea network increased node distance and network fragmentation. In contrast, in the hypersaline and polar networks, the removal of the peripherally located unknowns resulted in a more connected network structure and appearance. The impact of unknown node removal on degree, betweenness, and closeness persisted across taxonomical levels for the respective environments (Fig. 3-4; Supplementary Fig. S13 - 15) and followed a similar pattern across tested correlation metrics (Supplementary Fig. S16 - 19) and prevalence thresholds (Supplementary Fig. S20 - 23).

In summary, despite these minor differences in closeness score changes, these results suggest that unknown OTUs had significant and comparable relevance for community structure in all four environments. The exclusion of unknown nodes from their environmental networks led to a drastic change in network structure and interconnectedness, illustrating that the unknown taxa are critical for establishing co-occurrence relationships and for maintaining overall network shape within each distinct ecosystem.

3.2.6 Unknown Taxa Act as Important Hubs Within Extreme Environment Networks

After studying the overall relevance of the ‘microbial dark matter’ in extreme environmental networks, we then concentrated on finding the most important unknown components of each community. To do so, the hub score of each node was calculated to determine which taxa caused the most fragmentation or loss of network structure when removed, and therefore, may be a critical component for the microbial community structure of the target environment (Supplementary Dataset S2).

Hub scores were calculated at the genus-level for each node and the environmental networks were recreated, resizing the nodes as a function of the scores (Fig. 3-5). The overwhelming majority of the top hub scores, as symbolized by the largest-sized nodes in each network, were unknown for the hypersaline and polar environments, even at higher taxonomic ranks (Supplementary Fig. S24). Moreover, at least four of the nodes within the top 20 hub scores were unknown at the genus-level for all environments (Supplementary Dataset S2), providing further evidence that unknown taxa are key components of microbial community structure in all four extreme environments.

Student's t-tests were performed to evaluate the statistical significance of the differences between hub score values of known and unknown genera. Unknown genera in the hypersaline (p-value = 2.01E-13) and polar (p-value =2.41E-9) habitats had significantly higher hub scores than known genera, whereas in the deep sea environments the unknown nodes were marginally significant (p-value = 1.77E-3). No significant differences were observed between the known and unknown nodes (p-value = 0.6) in the studied hot springs environment. Based on these analyses, we concluded that while unknowns OTUs occupied and dominated key positions with the hypersaline and polar networks, all environments harbored relevant unknown hubs within their microbial communities.

3.2.7 Network Analysis of Unknown Hubs as a Tool to Prioritize Taxa for the Search of Novel Genes with Targeted Functions

The high frequency of currently unknown microbes as top hubs within their habitats implies that these organisms have high prevalence and co-occurrence within their microbial communities, indicating that they have successfully adapted to survive in these harsh ecosystems. Therefore, we hypothesize that our network approach, particularly using the hub score to prioritize unknowns, could serve as a critical foundation to study novel pathways and gene functions present in these yet-uncharacterized microbes. This hypothesis, however, poses two problems. First, this network-based approach relies on 16S rRNA sequencing data, which holds no additional functional information, and second, reference genomes for these uncultured organisms are

lacking. To circumvent these challenges, we used the 16S rRNA sequences of the top unknown hubs to probe large metagenomics databases, where extensive and diverse genome information is available on these microbial habitats, thereby facilitating the recovery of gene content associated with these unknown organisms. In parallel, we searched the literature for gene functions described to be associated to adaptation to our four studied extreme environments, resulting in a list of 86 "adaptation" related terms that mostly contained metabolic and stress response processes (Supplementary Table 1). Once high-confidence matching scaffolds were found, putative novel genes with the targeted functions could be computationally identified by searching for genes of unknown function in the neighborhood of genes previously annotated with the list of targeted terms.

Object 3-3. Link for Supplementary Tables S1 and S2

To explore this idea, the 16S rRNA gene sequences of the top five unknown and known hubs at the genus-level for each environment were blasted against 100 draft and permanently assembled metagenomes in the IMG/M database [137]. Hits with high similarity (i.e., $> 95\%$) and equivalent partial taxonomic classification in SILVA to the blasted hub 16S rRNA gene sequences were selected. These scaffolds were then filtered for a high (≥ 50) gene content and searched for operons containing gene descriptions matching any of the terms in our "adaptation" list, as well as genes labeled as "hypothetical proteins" or of "unknown function". Hierarchical clustering of gene distance between pairs of genes was used to identify putative operons, with all results validated by checking gene neighborhood information for each scaffold using the IMG/M genome browsing and annotation platform.

Overall, metagenomic screening with the 16S rRNA gene sequences of top known and unknown hubs returned numerous high-match, gene-rich scaffolds that varied by environment (Supplementary Table 2). The variation between habitats likely reflected the compositional differences of the metagenome database. Nevertheless, similar numbers of genes labeled as "adaptation", "hypothetical genes", and similar "putative adaptation-related" operons (see

Methods for assignment criteria) were identified for all environments, regardless of hub type, with scaffolds consistently containing an average of three to four potential adaptation-related genes to extreme environments. Interestingly, the average number of hypothetical genes found per scaffold differed most between environments. Deep sea hubs consistently had the lowest average number of hypothetical protein genes per scaffold (12.8 for unknown, 20.6 for known), whereas polar, hypersaline, and hot springs scaffolds had similar, larger numbers of hypothetical genes (up to 35 for known hypersaline hubs). These results suggest that a low total scaffold count does not prohibit detection of a high, equally abundant, number of hypothetical genes across environments or hub types. Importantly, across all habitats, an equivalent mean number of putative operons, where hypothetical and environment-relevant genes were found in close genomic proximity, were detected. Altogether, we recovered 6734 un-annotated genes present in 535 putative operon structures potentially involved in the metabolic and stress responses to extreme environmental conditions.

An example of the outcome of this approach for the targeted hot springs habitat is depicted in Fig. 3-6. Blast results showed a 97% sequence similarity match to the fasta sequence of unknown hub AB176701.1.1510 (Supplemental Dataset S2) within the metagenomic scaffold Ga007390_1000203,which was originally generated from Dewer Creek hot spring sediment in British Columbia. The scaffold contained 98 genes in total, of which 19 were annotated as 'Hypothetical Proteins' and distributed across seven operons (Fig. 3-6A). One of the identified operons contained three putative genes with predicted proteins of unknown function (i.e., two were labeled as hypothetical and one was labeled with unknown function). Further, this operon contained a gene annotated as Fe-S oxidoreductase and two well-known oxidative stress genes encoding DNA-binding ferritin-like protein (DPS) and rubrerythrin (RBR) suggesting possible roles in oxidative stress-related functions. An exact match for all six of the protein sequences in the same operonic structure was found in the genome of the bacterium *Blastocatellia* by searching the NCBI non-redundant database (Fig. 3-6B), which was recently sequenced as part of a hydrothermal vent metagenome project [199]. Further bioinformatic analysis of the poorly

characterized open reading frames (ORFs) in this operon led to the identification of Ga0073930_100020343 as a domain of unknown function (DUF) DUF3501 family member and Ga0073930_100020344 as part of the COG0247 family (FeS oxidoreductase). The link with oxidative stress was reinforced by the clustering of these same four genes with the hydrogen peroxide-inducible genes activator OxyR in *Sideroxydans lithotrophicus ES-1* (Fig. 3-6C). The two other hypothetical genes, Ga0073930_100020341 and Ga0073930_100020340, were identified as a putative SH3 domain-containing protein and a putative lysine synthase protein, respectively. Notably, the DUF3501-COG0247-RBR functional association had already been reported in a previous study where the authors proposed that the DUF3501 and COG0247 protein families enabled the functional adaptation of rubrerythrin from a thermophilic/anaerobic to a mesophilic/aerobic environment [200]. The identification of DUF3501 by our OTU prioritization analysis of the hot springs habitat supports this prior work suggesting the gene may be part of the microbial adaptation strategy to high temperatures. These results illustrate how the hub score prioritization can be successfully combined with computer-intensive mining of publicly available metagenomic data to identify novel and potentially ecologically relevant genes and gene products.

3.3 Discussion

Microbial communities are complex and dynamic, however, with the vast majority of Earth's microbes yet to be cultured or characterized, our understanding of these systems is likely limited or skewed by this large gap-in-knowledge. To more fully understand the impact of 'microbial dark matter' on ecosystem structure and function, we have developed a network theory-based approach to assess the relevance of the uncultured and unknown taxa within their microbial communities. Implementation of this bioinformatic pipeline demonstrated that: 1) specific patterns of the microbial network could be identified and compared for targeted ecosystems across taxonomic levels; 2) the comparison of centrality metrics between networks including and excluding the unknown taxa is an effective strategy to reveal the relevance of these organisms within their communities, 3) certain unknown taxa act as key players (hubs) in ecosystem structure due to their high prevalence and strong central connections; and 4) network

metrics can be used to prioritize unknown taxa for downstream analysis by using the 16S rRNA gene sequences of top-ranked hubs as probes to screen publicly available metagenomic datasets for the identification of ecologically relevant gene functions.

3.3.1 Harnessing the Power of Networks to Elucidate 'Microbial Dark Matter'

No matter the environment, previous research has shown that 'microbial dark matter' represents a significant limitation to the exploration of the global microbiome [201, 37, 12, 202, 33, 135, 203, 101, 63]. To address and overcome this challenge, we developed a combined bioinformatics pipeline and network theory approach that was applied to a large, geographically diverse 16S rRNA dataset of four extreme aquatic environments to determine the ecological relevance of 'microbial dark matter' in these communities. Although correlation-based microbial networks cannot infer the nature of ecological relationships, such as syntrophy or competition, they are indicative of social interactions within the community and can serve as important focal points for downstream analyses [204]. Our analysis clearly showed that the unclassified and uncultured taxa were prevalent and represented a highly significant proportion of the microbial diversity in all ecosystems examined. Both known and unknown OTUs were found to be environment-specific, agreeing with previous reports of habitat-specific, niche-partitioning species of hypersaline lakes [52], deep sea vent communities [205, 206], and polar lakes [21, 55]. Furthermore, our work extended beyond simple composition analysis and demonstrated the consistent and significant contribution of unknown OTUs to microbial community structure.

By using network metrics to study these four extreme environment networks, additional insights into the relevance of these unknown organisms could be gained. The unknown OTUs positively contributed to betweenness and degree centralities (i.e., denoting microbial interactions). More importantly, the exclusion of unknown taxa was more detrimental to the overall network than the exclusion of random known components, as more central node connections (e.g. ecotype interactions) were lost, causing a greater network fragmentation. Moreover, unknown taxa established co-occurrence relationships with themselves, suggesting that they might be phylogenetically related, as is the case for known taxa [197, 198]. The results

presented here reveal that many unknown taxa are key elements of microbial ecosystems and strongly advocate for the inclusion of unknown taxa in any metagenomics or amplicon composition and interaction studies, as key biological interactions may remain undiscovered otherwise.

3.3.2 Networks can Prioritize the Most Ecologically Relevant Unknown Taxa in a Community

Unclassified microbial taxa often occurred as top hubs across all examined environments. Since hubs, by definition, significantly contribute to network structure and cohesiveness, these unknown microbes can be considered keystone taxa, most likely playing vital and meaningful roles within key ecosystem processes in these habitats. Moreover, hub positions indicated that these unknown taxa were prevalent in their environments and co-varied with many other taxa, and hence can be considered successful components of their microbial communities. These results support the value of our analysis and suggest that this approach could be used to identify the highly interacting OTU components of any microbial community. The frequent dominance of unknown taxa as top-scoring hubs stresses the need for further exploration and functional characterization of these novel species and also offers new tools for prioritizing novel taxa for follow-up studies.

3.3.3 Filling in Gaps-in-Knowledge of Ecosystem Functioning with Hubs of Unknown Taxa

Understanding the emergent properties of an ecosystem, i.e., those taxa, genes, and functions that are important for a particular niche, can have a big impact on our understanding of that environment. Using amplicon-based approaches to address these questions, however, can be limited. For example, amplicon-focused tools, such as PICRUSt [207] and Tax4Fun [**?**], can retroactively predict gene function from 16S rRNA gene data [208]. However, these tools require reference genomes to be present and well annotated for each ecotype, and therefore, cannot be applied to novel, uncultured organisms.

Here, we envisioned an alternative approach to probe metagenomics databases with 16S rRNA sequences prioritized from the top unknown hubs of a given environment and used those

sequences to investigate the gene content of associated scaffold hits. The approach leverages the wealth of metagenomics data currently available within public databases, which encompasses hundreds of terabytes of data [209, 210] and represents untapped sources of valuable information that can, and should be, exploited both for fundamental science and for potential biotechnological applications.

The successful retrieval of unknown genes, potentially involved in environmental stress responses, from uncultured and unclassified organisms, indicated that this network and hub identification approach is an effective strategy to use prioritized OTUs for direct data-mining efforts. Notably, in this proof-of-concept effort, just a few top hub score OTUs within the hot spring networks were used to screen a fraction of the available metagenomics information, still recovering a substantial number of candidate genes. With one specific example, we illustrated the power of this methodology to unveil interesting gene functions. Supported by sufficient computational resources, up-scaling of this concept holds the potential for the large-scale discovery of novel gene functions and pathways, further unraveling roles that these unknown taxa may play within their respective ecosystems.

In summary, this approach has the potential to be extended to other aspects of the environmental microbiome, including, but not limited to, the archaeal and eukaryotic taxa, as well as other multi-omic platforms (e.g. metaproteomics, metabolomics, and metatranscriptomics). As reference databases continue to grow, taxa and gene co-occurrence network analyses and measurements can also be used to evaluate changes in ecosystem structure over different temporal and spatial scales. The application of this strategy to a variety of microbial ecosystems from both extreme and non-extreme environments could be used to more fully understand those features of the hidden microbial world that are critical for environment-specific or global attributes of microbial ecosystems.

3.4 Methods

3.4.1 Data Retrieval

To mine samples from public databases, search queries of "16S", "V4", "V3", "Illumina", "hot springs", "hypersaline", "Arctic", "Antarctic", "deep sea", and "hydrothermal" were utilized to find suitable studies from NCBI SRA and JGI Gold. All samples were classified to their respective environmental categories (i.e., hot springs, hypersaline, deep sea, and polar) using available information provided by the studies in the public repositories. All raw data were converted to fastq format using NCBI SRA Toolkit (http://ncbi.github.io/sra-tools/). Although the complete dataset included both paired-end and single-end samples, only the forward reads of paired-end samples were used in subsequent steps, due to the consistently noted lower quality of reverse reads of Illumina samples.

3.4.2 Sample Preprocessing, Filtering, and OTU Mapping

Quality filtering and preprocessing of all sequence data was performed through the split_librairies_fastq.py script from the Quantitative Insights into Microbial Ecology (QIIME) pipeline (Version 1.9.1) [176] using a Phred quality threshold of 19. All sequences passing quality filtering were clustered at 97% sequence similarity and classified to OTUs using the SILVA (v128) SSU reference database by the pick_open_reference_otus.py script. All singletons were discarded. The filter_taxa_from_otu_table.py script was used to remove any OTUs related to Archaea, mitochondria, or chloroplast. The collapse_samples.py script was used to compare OTU presence across environments (i.e. hot springs, hypersaline, deep sea, and polar). The filter_samples_from_otu_table.py script was used to separate the global OTU biom table into four environment-specific biom tables. All subsequent statistical and network analyses were conducted in R (v 3.5.1).

3.4.3 Identification Strategy for Unknown Taxa

At each taxonomic level, from phylum to genus, unknown taxa were identified as any OTU taxonomically assigned as: "uncultured", "uncultured bacterium", "Unknown", "Unassigned", "Ambiguous taxa" or "NA" by the open-reference picking strategy. These OTUs were renamed as

"Unknown" for all subsequent analyses and comparisons. An OTU was only labeled as "Unknown" at the specific taxonomic classification level at which it could not be taxonomically assigned beyond the higher classification descriptors. For example, if an OTU had a known order but unknown family description assigned by the reference database, it would only be designated as an unknown in the network analysis for family and genus taxonomic classification and, at all higher classification ranks, would be referred to its known classification status.

3.4.4 Network Creation

To create networks of each environment, OTU biom tables and corresponding mapping information (i.e. sample ID, geographic location, longitude, and latitude) were imported into R using the package phyloseq McMurdie, , phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. OTU tables, now converted into phyloseq objects, were filtered using the filterTaxonMatrix() function from phyloseq, keeping only taxa present in a given percentage (from 30 - 40%) of samples per environment to reduce sparsity and ensure robust results. For each environment, the filtered OTU phyloseq object was normalized, transformed and converted into an adjacency matrix based on covariance by the spiec.easi() function from the package SpiecEasi [169], using Meinshausen- Buhlmann's neighborhood selection (method = "mb" parameter) to estimate the conditional dependence of each pair of OTUs. Each adjacency matrix was then converted into an igraph object and visualized as a network using the adj2igraph() and plot_network() functions from SpiecEasi. Networks were created and visualized at each taxonomic classification level, using the function plot_network() of phyloseq, with nodes representing OTUs and edges representing direct co-occurrence relationships between OTUs.

To create reduced networks, either excluding unknown OTUs ('Without Unknown' network) or randomly selected known OTUs ('Bootstrap' network), the delete_vertices() function (package igraph) was used. To create bootstrap networks, known nodes were randomly selected by the sample()function, with the x parameter equal to the total number of OTUs, known and unknown, at that specific taxonomic classification level for the chosen environmental network and

size parameter equivalent to the number of unknown OTUs for that network. Networks were created using the same SpiecEasi pipeline as described above. These randomly reduced networks were created 100 times for each taxonomic classification level from phylum to genus for each of the four target environments (i.e., hot springs, hypersaline, deep sea, and polar habitats).

3.4.5 Neighborhood Analysis

For each environmental network, the as_edgelist()function from igraph was applied to identify all edges between nodes. Using the taxonomic classification information found in the original OTU table that was retrieved by tax_table() function on the phyloseq object, all nodes were matched up with their taxonomic classification. A data frame containing all taxonomic information of each pair of connected OTUs was then used to identify which classes shared edges.

3.4.6 Network Analysis Strategy

For each environment, at each taxonomic level from phylum to genus, network-level and node-level measures of networks including unknown OTUs (i.e., "Original"), excluding them (i.e., "Without Unknown"), and excluding an equal number of randomly selected known OTUs (i.e., "Bootstrap") were evaluated and compared against each other and visualized as boxplots. The network measures evaluated for all nodes present within each network were degree, closeness, and betweenness and were calculated by using the igraph package functions: degree(), betweenness(), closeness(), and hub_score(). Wilcox test was used to evaluate the statistical significance of changes in degree, betweenness, and closeness between the three network types using the stat_compare_means()and compare_means() functions from the ggpubr package. P-values were adjusted using the Holm method [178] and boxplots were created using the package ggplot2.

3.4.7 Hub Blast Against Metagenomes

For each environment, the fasta sequences of the five known and unknown taxa with the highest-ranked hub scores were retrieved with the subseq function from the SEQTK toolkit (https://github.com/lh3/seqtk), using a text file of the hub sequence names and the fasta file of complete sequences (new_refseqs.fna) produced by the QIIME pick_open_reference_otus.py script

as the input. Next, using the selected genomes blast feature in IMG/M, the five known and unknown hub sequences were blasted (blastn, using the default number of hits (500) and e-value (1E-05) specified) against 100 publicly available finished, draft, and permanent draft metagenomes pertaining to each environment. Blast results were exported and saved in text format and further analyzed in R. Only blast hits meeting $\geq$ 95% match to the query hub sequence were retained. Of these hits, only metagenome scaffold hits where at least 50 genes were present were retained. For each scaffold that met the percent identity for the 16S rRNA gene (i.e., 95%) and gene number (50%) criteria, the gene content information was accessed manually and exported. The gene name, strand position, and start and end coordinate positions given in the file were used to identify the number of functionally annotated and unknown hypothetical protein genes and also used to identify putative operons of genes with similar functions.

3.4.8 Identification of Hypothetical and Putative Adaptation Genes and Operons

The term "hypothetical" was used in the grep()function in base R to identify hypothetical protein genes on each strand for each scaffold. A list of keywords related to metabolic and extreme environmental stress response functions retrieved from literature was used to parse for adaptation gene matches among all genes present on each strand (Supplementary Table 1). To identify closely related genes, hierarchical clustering was performed based on the distance between gene start and end coordinate positions of each gene pair using the packages ape and dendextend, and the function hclust()in base R. Closely clustered genes on a single branch represented putative operons and any ten genes within 5000 bp or less to one another were considered to belong to one operon. If targeted genes were among the ten closest neighbors (i.e., less than 5000 bp away) to a hypothetical gene, we defined this set of hypothetical and potentially extreme stress adaptation-related neighboring genes as a putative adaptation operon. Gene clusters were then analyzed using the PubSeed database [211] and visualized with the Gene Graphics tool [212].

3.4.9 Sample Criteria Validation

To account for the possible confounding effect of sample size, all networks were reconstructed and reanalyzed, using a range of different percentages (from 15% to 40% at 5% increments) of samples in the initial filtering process of the OTU table, discarding any taxa that did not meet this sample percentage threshold from being included in the networks. Networks excluding unknown and random known nodes were constructed using the same methods as described previously and network measures were recalculated and visualized as boxplots to determine the effect of different sample size criteria and its statistical significance.

3.4.10 Network Tool Validation

To determine whether other measurements of species' co-occurrence relationships changed our final network analysis results, three other methods were used to calculate the relationship between pairs of species in R: SparCC [80], CCLasso [175], and Pearson correlation. The R rcorr()function, with the parameter type set to "pearson" from the package Hmisc was used to calculate Pearson correlation. The R cclasso()function (https://github.com/huayingfang/CCLasso/blob/master/R/cclasso.R) was used to calculate CCLasso correlation, and the adapted sparcc()function in the SpiecEasi package was used to calculate SparCC correlation. Subsequent networks were created using igraph. Boxplots were created to compare the median network measure scores of degree, betweenness, and closeness for the three network types at all classification levels for the four (SpiecEasi, SparCC, CCLasso, and Pearson) correlation networks.

3.4.11 Scripts and Documentation

To encourage a deeper investigation into the role of microbial dark matter, ready-to-use scripts and documentation to apply this methodology to other ecosystems are available. All scripts used in this analysis, along with a complete documentation of the bioinformatics pipeline, are available at http://github.com/Conesalab/MDM.

Table 3-1. Breakdown of node and edge attributes across extreme environmental networks.

Environment	Prevalence Threshold (Number of Samples)	Number of Selected Nodes [fraction from initial data]	Number of Connected Nodes [Number of Unknown at genus]	Number of edges in network [K-K, K-U, U-U at genus]	Ratio of Edge vs. Node Size
Hot Springs	104	291[4.4E-4]	290[133]	552 [167,230,155]	1.9
Hypersaline	85	193[3.7E-4]	191[118]	516 [89,125, 302]	2.7
Deep Sea	75	102[1.6E-4]	97[51]	194 [59,65,70]	1.9
Polar	113	279[9.6E-4]	274[171]	797 [112,244,441]	2.9

K-K: Known-Known. Edge between two nodes, both of which are known at genus level; K-U: Known-Unknown, Edge between two nodes, one of which is known and one unknown at genus level; U-U: Unknown-Unknown. Edge between two nodes, both of which are unknown at genus level

Table 3-2. Comparison of unknown impact on network measures.

Environment	Degree	Betweenness	Closeness	Proportion of unknown OTUs out of 20 top hubs
Hot Springs	$\downarrow (F,G)$	$\downarrow (G)$	*NS*	40 %
Hypersaline	*NS*	$\downarrow (G)$	$\uparrow (P-F)$	100 %
Deep Sea	$\downarrow$	$\downarrow (F,G)$	$\downarrow (F)$	20 %
Polar	$\downarrow$	$\downarrow (P-G)$	$\uparrow (P-G)$	90 %

P: Phylum; *F*: Family; *G*:Genus; *NS*: Not Significant; Downward pointing arrow: Significant decrease; Upward pointing arrow: Significant increase

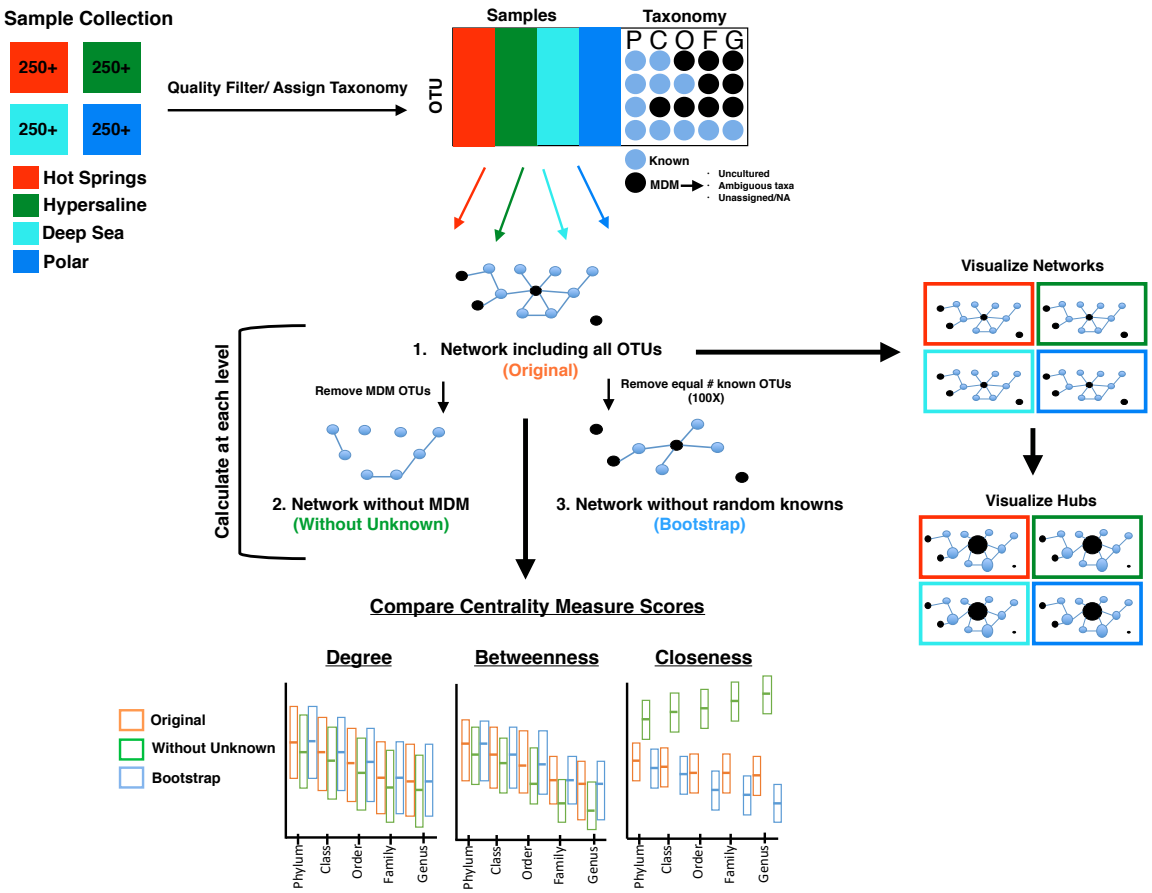
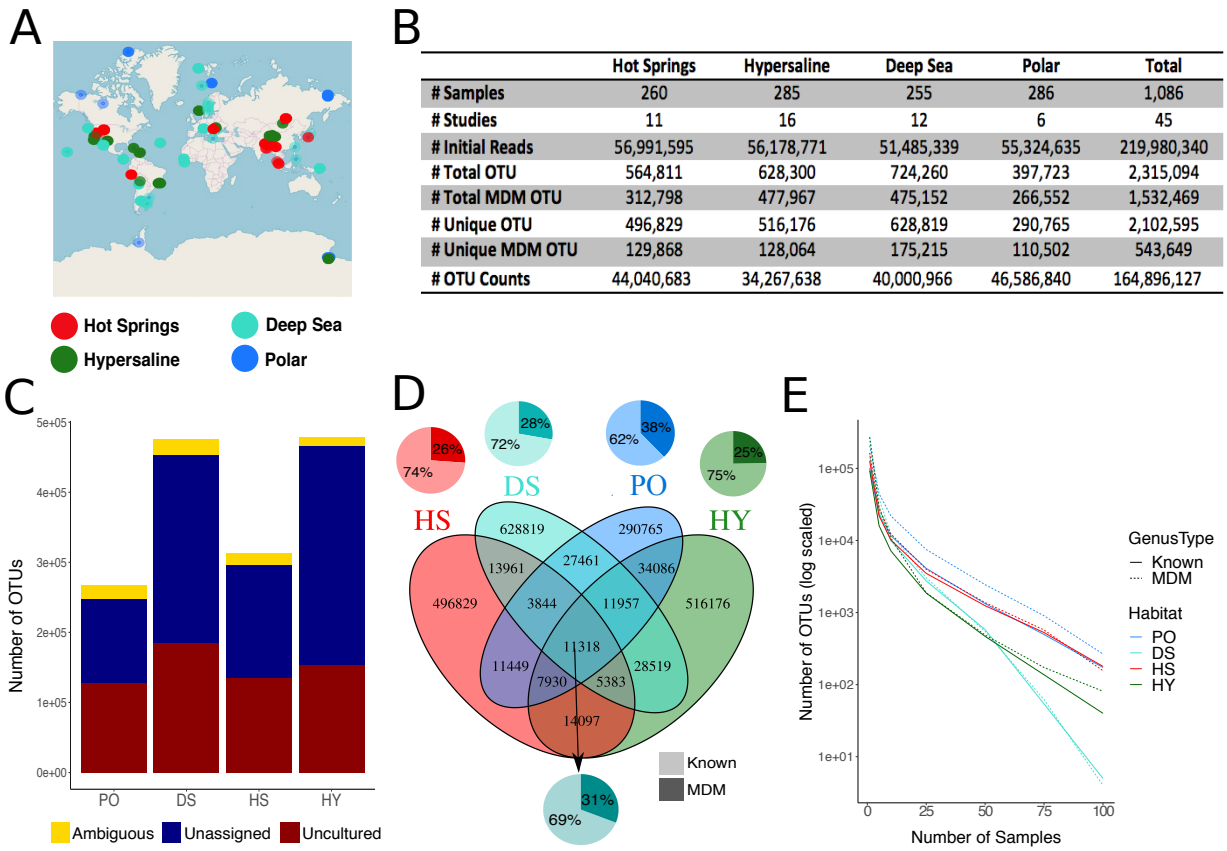


Figure 3-1. Overview of the analysis pipeline. A minimum of 250 samples was retrieved for each of the four different extreme environments- hot springs (red), hypersaline (dark green), deep sea (turquoise), and polar (blue). Sequence reads were quality filtered, assigned to a taxonomy, and clustered to OTUs. At each classification level, any unassigned, ambiguous, or uncultured OTUs were designated as unknowns, or 'microbial dark matter' (MDM). For each environment, at each classification level, the direct co-occurrence relationship between all OTUs was mathematically modeled as a network. Networks were created for each environment, across all taxonomic classification levels, including all OTUs (Original, orange), excluding MDM (Without Unknown, light green), and excluding an equal number of random knowns (Bootstrap, blue). Network centrality metrics (i.e. degree, betweenness, and closeness) were calculated for each node, compared, and visualized as boxplots between these network types. Hub scores were calculated for each node in the Original network and networks were recreated, resizing by hub score, where the largest size node indicates a top hub species.



	Hot Springs	Hypersaline	Deep Sea	Polar	Total
# Samples	260	285	255	286	1,086
# Studies	11	16	12	6	45
# Initial Reads	56,991,595	56,178,771	51,485,339	55,324,635	219,980,340
# Total OTU	564,811	628,300	724,260	397,723	2,315,094
# Total MDM OTU	312,798	477,967	475,152	266,552	1,532,469
# Unique OTU	496,829	516,176	628,819	290,765	2,102,595
# Unique MDM OTU	129,868	128,064	175,215	110,502	543,649
# OTU Counts	44,040,683	34,267,638	40,000,966	46,586,840	164,896,127

Figure 3-2. Summary of environmental 16S rRNA gene data. A. Map of the sample sites used in this study. Circles symbolize sample locations and are color-coded by environment: hot springs (HS, red); hypersaline (HY, dark green); deep sea (DS, turquoise); and polar (PO, blue). B. Summary of data used in this study. OTUs counts are provided at the genus-level. C. Proportion of 'microbial dark matter' (MDM) OTUs for each environment labeled as unassigned (dark blue), uncultured (dark red), and ambiguous (yellow) after SILVA and UCLUST-based taxonomic assignment to OTU. D. Venn diagram of shared OTUs in four extreme environments. Each pie chart depicts the proportion of unique OTUs that were known (lighter shade) and unknown (darker shade) for each environment, with the bottom-most pie chart showing combined data for all environments. E. Prevalence curves indicate the number of unique OTUs consistently present at an increasing number of samples. Dotted lines signify the prevalence of MDM OTUs and solid lines signify the prevalence of known OTUs.

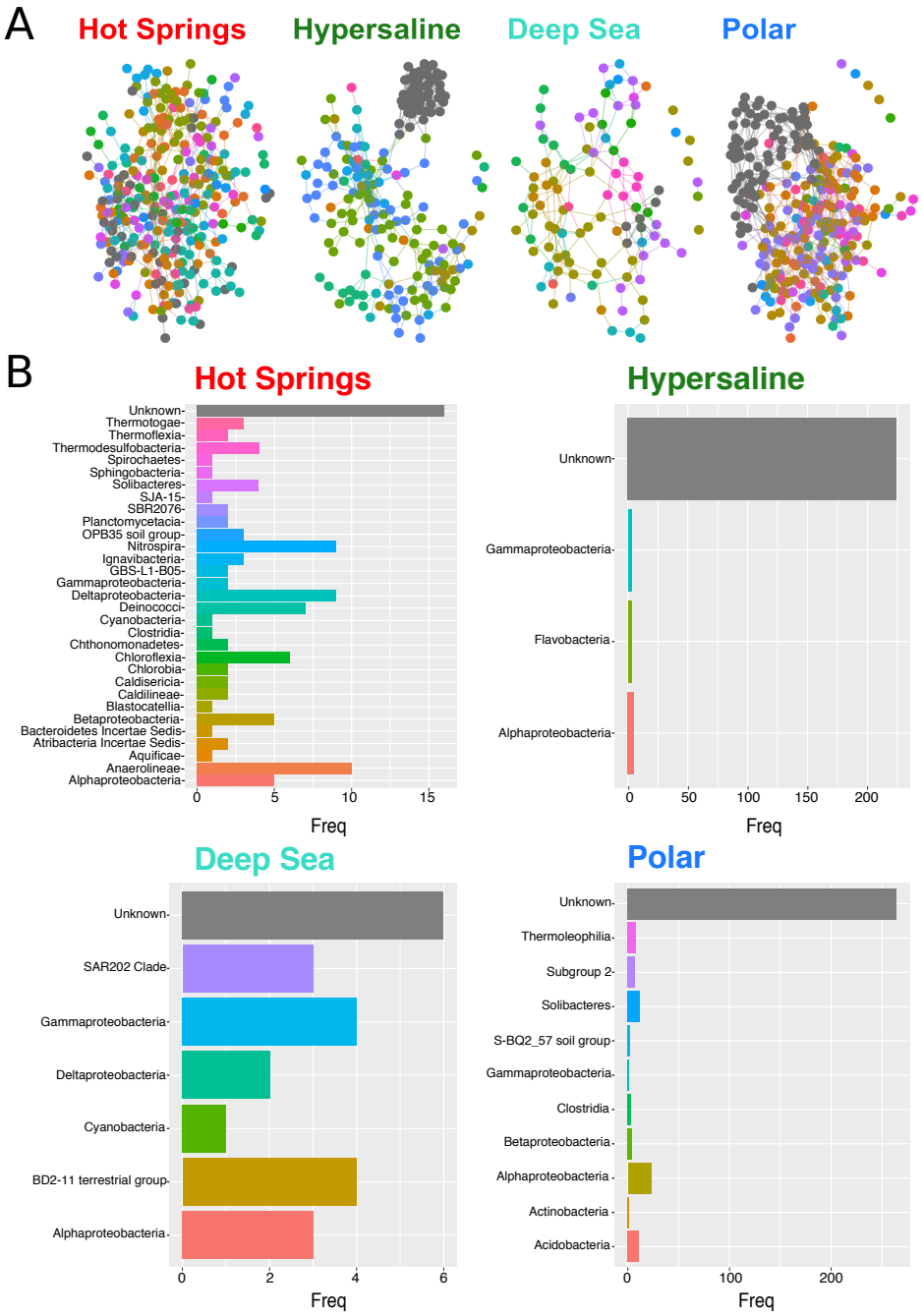


Figure 3-3. Analysis of environmental network taxa interconnectedness. A. Microbial networks at class taxonomic classification level. Nodes are colored by class assignment, with gray nodes representing unknown taxa at the class-level. B. Bar graphs of the co-occurrence relationships (i.e., edges) of unknown OTUs with other taxa at the class-level within each environmental network. Y-axis labels and colors signify the different classes with which unknowns were found to co-occur. Unknown-unknown relationships are represented in gray.

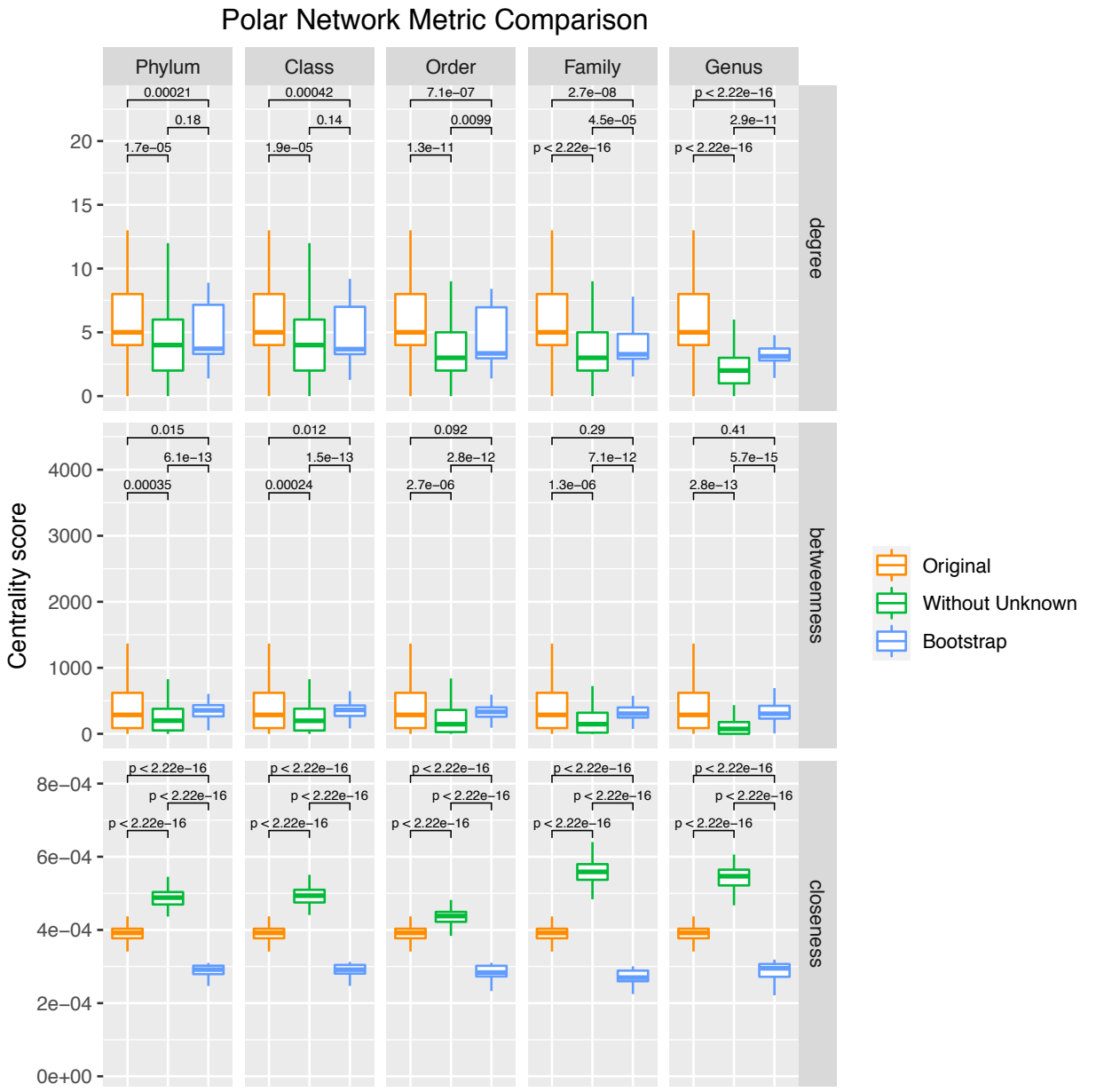


Figure 3-4. Impact of unknown taxa on polar network metrics at different taxonomic levels. Boxplots of degree, betweenness, and closeness centrality values of nodes present in different network types at different taxonomic levels. Wilcoxon pairwise comparisons were used to assess significance between the three network types (Original-Without Unknown, Without Unknown-Bootstrap, Original-Bootstrap) for each taxonomic level. For each comparison, p-values after Holm adjustment are shown.

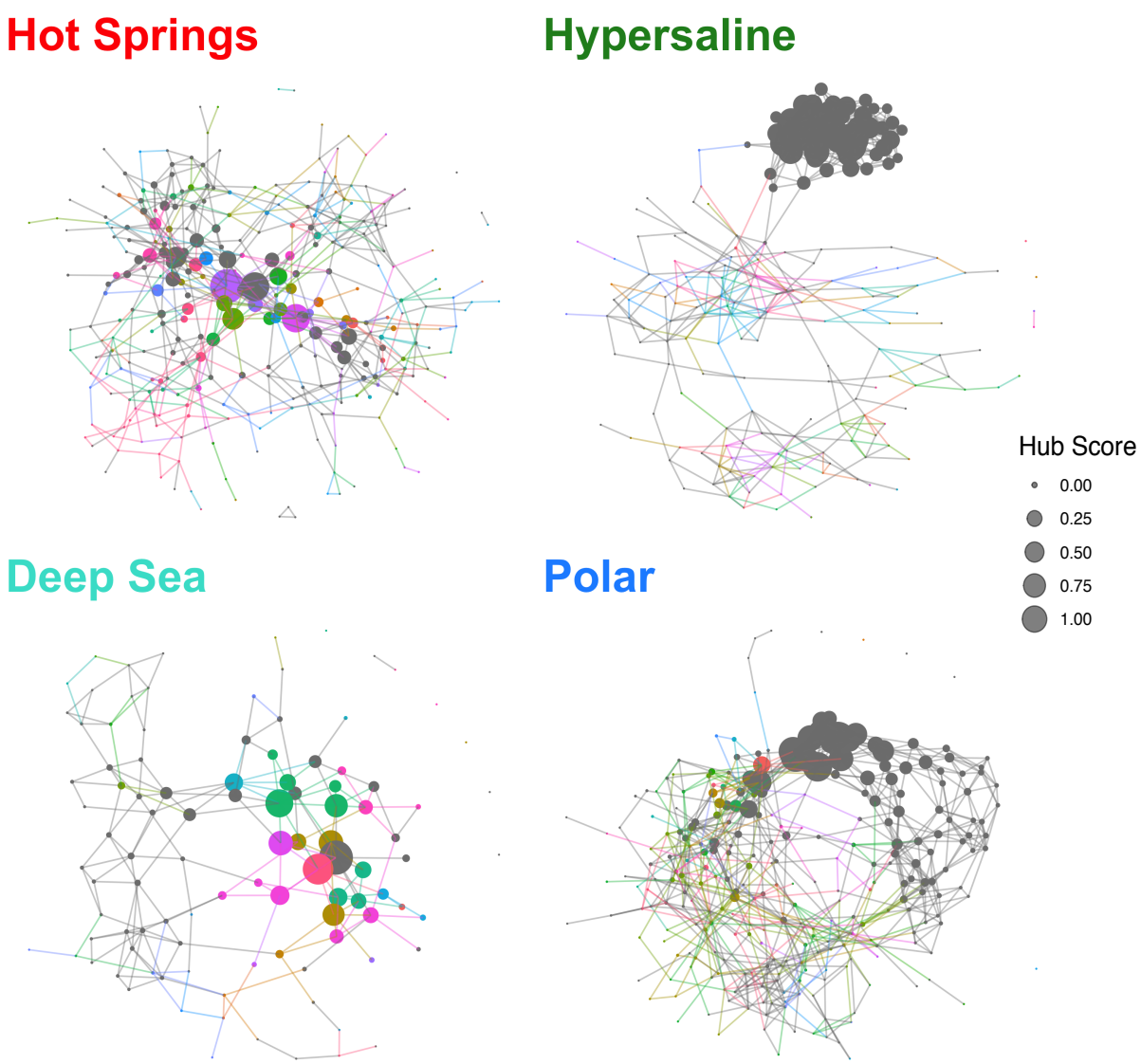


Figure 3-5. Hub analysis of extreme environmental networks. Environmental networks at the genus-level with nodes sized as a function of hub score. Nodes are colored by genus classification with ambiguous, unassigned, or uncultured taxa depicted in dark gray.

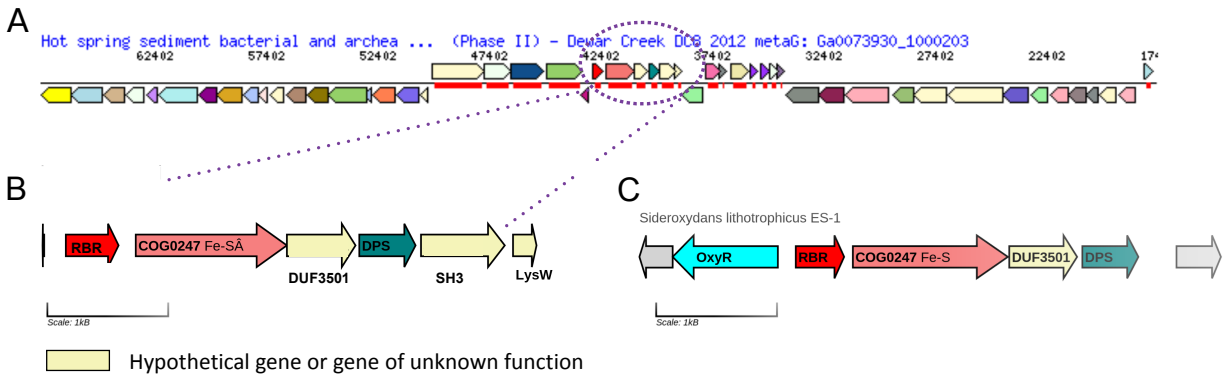


Figure 3-6. Metagenomics analysis of top unknown OTU AB176701.1.1510.A. Overview of operon annotation in scaffold hit Ga0073930_1000203 obtained from the JGI metagenomics browsing platform. Genes in light yellow color represent hypothetical genes or genes of unknown function. B. Zoom-in of the selected operon with adaptation-related gene annotations. Bioinformatic function predictions for genes of unknown function are indicated under their gene boxes. C. DUF3501-RBR containing operon for Sideroxydans lithotrophicus, including the OxyR gene. RBR: Rubrerythrin gene; COG0247 Fe-SA: Fe-S oxidoreductase gene; DUF3501: Domain of unknown function 3501; DPS: DNA-binding ferritin-like protein; SH3: SH3-domain-containing protein; LysW: Lysine synthesis protein W. OxyR: hydrogen peroxide-inducible genes activator.

CHAPTER 4
A COMBINED NETWORK AND METAGENOMICS APPROACH TO ENABLE RECOVERY OF NOVEL CONSERVED ADAPTATION-RELATED GENES

4.1 Introduction

Novel, uncultured phyla, like members of the Candidate Phyla Radiation [213] or Candidatus Abyssubacteria [214] act as key drivers of global nitrogen, carbon, and sulfur cycling across a range of environments, including marine, terrestrial subsurface, ocean, and groundwater. These novel organisms, also called MDM, are thus important microbial players that possess unique characteristics necessary for survival and adaptation to a wide range of extremophilic conditions. In the previous chapter, we demonstrated that MDM are frequently hubs in extremophilic aquatic microbial communities, suggesting that these organisms, by nature of their hub score, high interconnectedness, and high prevalence, may be particularly adapted to survive and thrive within such inhospitable conditions. Due to their unique ability to adapt to a wide range of ecosystems, MDM may thus be an important source of novel genes, particularly of genes with functions related to stress. Identifying and characterizing this novel gene content is of great interest to researchers in microbial ecology and medicine alike, as these genes may help illuminate key questions related to evolution and adaptation of life and may also be important for antibiotic or other biomedical therapies necessary for treatment and prevention of disease. However, studying unknown genes of unknown microorganisms is especially challenging due to the low amount of available genomic and genetic information.

Up until recently, microbial functional diversity was traditionally achieved through whole genome shotgun (WGS) sequencing. Yet, with the affordable nature and plethora of existing bioinformatics tools for the analysis of 16S rRNA amplicon datasets, researchers are now using the taxonomic abundance results derived from 16S sequencing as starting points to infer the functional potential of microbial communities, through ancestral reconstruction based tools like PiCRUSt [207], Tax4Fun [**?**], and Vikodak [215], which rely on the assumption that the total genetic functional capacity of a microbial community is simply a reflection of the total relative microbial abundances of that same community. Though PiCRUst, Tax4Fun, and Vikodak have successfully been able to predict function retroactively of metagenomics datasets like the HMP

(Human Microbiome Project) and Tara Ocean datasets, these tools have major limitations and cannot be applied to most large-scale datasets or to organisms that lack genomic representation, leaving the genetic and functional content of MDM organisms unexplored. The requirement of a complete and well annotated reference genome to perform accurate functional prediction for these 16S reconstruction-based tools makes it impossible to perform functional prediction of the predominantly unknown taxa of extreme environments. Despite these drawbacks, the potential gene content of unknown species may be of particular interest for improving understanding of adaptation mechanisms, developing effective antibiotics impervious to bacterial resistance [216], and illuminating microbial evolution and ecology. For these reasons, functional inference of novel genes, particularly those belonging to MDM, is of extreme importance to advance microbial research and requires a methodology independent of reference genomes.

Here, we propose a novel strategy to infer gene function, centered on using the 16S rRNA sequences of unknown species as probes, via BLAST search, to find novel genes within functionally annotated scaffolds of a manually-curated comprehensive metagenome database. In the first step of this large-scale strategy, we use the hub score metric to rank and select promising top-scoring hub unknown organisms to drive the functional characterization of unknown genes. Due to their integral network position, we believe that these hubs may possess optimized adaptation strategies and therefore are ideal candidates to search for novel biological pathways of survival. Although gene function of unknown species cannot be predicted from 16S metagenomics alone when no reference genomes are available, neighboring known and unknown genes to the query sequence present within a metagenome scaffold can aid understanding and inference of the potential functional properties of MDM by the guilt-by-association clause. Building on these principles, we present a computational approach that uses the fasta sequences of particularly relevant hub MDM components as biomarkers against publicly available metagenomes to enable identification and predicted functional characterization of novel genes particularly relevant for extremophilic adaptation.

Hundreds of novel genes and operons within gene-rich metagenome scaffolds that matched with high sequence similarity to the query 16S sequences were able to be identified and functionally characterized. Comparable and large numbers of hypothetical and adaptation-related genes and putative adaptation-related operons were detected, regardless of environment or hub type, showing that our approach is robust and effective for well-characterized and uncultivated organisms alike. Afterwards, sequence similarity network ,genome neighborhood network, and gene ontology (GO) annotation analyses were conducted to find frequently occurring, evolutionary conserved novel genes and their functional properties. A substantial proportion of uncharacterized genes, particularly genes annotated as domains of unknown function (DUFs), were shown to be evolutionarily conserved, clustering with high sequence similarity to Uniprot genes. We also show, through the use of combined GO graphs, case study examples of frequently occurring, evolutionarily conserved novel genes whose functional roles appear to be predominantly active in extremophilic adaptation.

Our approach enables highly accurate novel gene recovery and characterization while circumventing limitations of traditional microbial functional inference strategies. Furthermore, by the use of a searchable local 13 terabyte metagenome database, our approach is able to accomodate the computational power and memory required of large-scale datasets, widening the applicability of this methodology to future microbial environmental meta-analyses. In sum, our approach aids large-scale screening of novel gene functions and pathways in poorly characterized genomes and may thereby prove useful for understanding microbial adaptation to extreme conditions, within and beyond Earth.

4.2 Results

4.2.1 Overall Strategy to Detect and Functionally Characterize Novel Genes from Metagenomes

Our overall strategy to detect novel genes from metagenomes (Fig. 4-1) centered on using the FASTA sequences of top-scoring hub unknown microorganisms from our previous work as probes against a large metagenome database that consisted of 6,396 publicly available metagenomes from IMG/M. A blastn search was performed against the metagenome database to

retrieve metagenome scaffolds that matched at high confidence ($\geq$ 95 % identity) to each of the 20 known and unknown query hub FASTA sequences of hot springs, hypersaline, deep sea, and polar microbial communities. The scaffold hits were screened to meet a minimum gene count of 30 genes. From these remaining high-confidence gene-rich scaffolds, we then identified putative operon structures (outlined in red) where uncharacterized genes with no known functions (in black) were assembled together in close proximity to genes annotated with adaptation-related keywords. Lastly, comparative genomics and bioinformatics strategies (i.e. sequence similarity networks and gene ontology (GO) annotation analyses) were implemented to identify patterns of gene conservation and infer gene function of particularly interesting novel genes.

4.2.2 Hypothetical Genes, Adaptation-related Genes, and Adaptation Operons Constitute Substantial Proportions Across Metagenome Scaffolds

4.2.2.1 Promising Hub Blast Results Return Thousands of High-Confidence Scaffold Hits

First, to ensure that our strategy was manageable and possible, we evaluated the utility of top 20 unknown top hub taxa of the hot springs, hypersaline, deep sea, and polar microbial communities as probes in identifying novel genes from high-confidence gene-rich scaffolds. A blastn search was conducted for each top hub sequence (see Methods), keeping all unique results that matched with at least 95 % sequence similarity and at least 95% sequence coverage (query cover). As a comparison, the top 20 known hub taxa of each community were also evaluated and used as probes in the blastn search against the metagenome databases. The blastn search proved successful in retrieving a total of 14,762 metagenome scaffold hits with at least 95 % sequence similarity to the top hub taxa of all four extreme environments.

The results of the blastn analysis returned up to 4088 (polar-known) high-confidence metagenome scaffolds hits across query sequences, demonstrating the utility and success of using hub sequences as probes. A similar total number of hits was retrieved for hubs from deep sea and hot springs microbial communities,regardless of known or unknown status, suggesting that the approach is robust and successful for recovering genomic information for unknown and well-characterized organisms alike. On the other hand, up to ten times the number of

high-confidence scaffold matches were able to be returned for hypersaline known than unknown hub sequences, due to the lack of publicly available metagenomes for hypersaline conditions in comparison to the better studied hot springs and deep sea conditions. The majority ($\geq$ 93 %) of the 14,762 scaffold hits contained less than 30 genes and were thus discarded from further analysis. The evaluation of the remaining scaffolds that met both the percent identity and gene count thresholds revealed an average gene count ranging from 34 (hypersaline, unknown) to 346 (hot springs, unknown) genes per scaffold (Table 4-1), enabling identification of operonic structures to be possible for all examined sequences.

4.2.2.2 Large and Comparable Percentages of Hypothetical Genes and Operons Found Across Metagenomes

Functional annotation of these predicted genes through prodigal [217] and eggnog-mapper [218] revealed that hypothetical genes (genes with no ascribed functions, grouped into COG category S) made up between 12 to 29 percent of all genes present per scaffold (Table 4-1), with the highest total number of hypothetical genes found for hot springs unknown and deep sea known query hits. The similar and high average percentage of hypothetical genes found within each metagenome, despite initial differences in environment, hub status, or total scaffold hits across all query combinations, demonstrates that novel gene recovery is both possible and successful using unknown hubs of different extreme aquatic environments and that this approach can be applied to any hub organism.

Further examination of the reduced set of metagenome scaffolds confirmed the high abundance of hypothetical genes, demonstrating that for each environment and hub type, most genes found within metagenomes belong to COG category S (no known functions), COG category E (amino acid metabolism and transport), COG category C (energy production and conversion), and COG category J (translation) (Fig. 4-2A). The dominance of hypothetical genes became even more apparent when examining each metagenome scaffold separately (Fig. 4-2B), with over 500 genes of COG category S identified in some metagenome scaffolds. The finding of an abundance of hypothetical genes across all metagenomes or individual metagenomes, suggests

that a substantial fraction of metagenome scaffolds is currently unknown. Thus, further investigation and characterization of these unknown genes is needed to better understand the full functional properties of a given species and the environment overall.

Having confirmed the high abundance of hypothetical genes, we next wanted to identify if any of these hypothetical genes were found alongside genes with known functions, in operonic structures. Hierarchical clustering of genomic distance was performed for pairs of genes present on a given strand of a metagenome scaffold to identify closely related genes, with any ten genes within 5000 bp to one another considered to belong to the same putative operon. If targeted genes were among the ten closest neighbors (i.e., less than 5000 bp away) to a hypothetical gene, we defined this set of hypothetical and neighboring genes as a putative hypothetical operon.

We then investigated the potential of operon containing unknown genes (OCUGs) to represent functions for adaptation to harsh conditions. A text mining approach was used to compare the gene annotations in OCUGs to a list of 87 adaptation-related genes extracted from the literature. By this manner, adaptation-related genes were able to be identified and were found to constitute a similar, substantial fraction (between 7 to 9 percent of all genes) across all metagenome scaffolds, regardless of environment or classification status of the query sequence. Again,the comparable and large fractions of hypothetical and adaptation-related genes detected across metagenome scaffolds demonstrated that, despite discrepancies in the number of total successful hits. our approach of using hub sequences as probes within metagenomes is robust and effective for well-characterized and uncultivated organisms alike. Furthermore, identification of hypothetical operons containing adaptation terms revealed a total of 77 putative adaptation operons for all queries, with the largest share (344 operons) detected for hot springs unknown queries. With such a large number of operons identifed, a computationally efficient method that would enable a systematic overview of the functional properties of each operon was necessary.

4.2.3 Evaluation of Biological and Functional Properties of Hypothetical Operons Reveals Strong Link to Stress Response

To gain a more comprehensive understanding of the biological processes and molecular mechanisms potentially at play across all hypothetical operons, we decided to perform a GO annotation analysis through Blast2GO [219] using the GO terms of adjacent, well-characterized genes to infer function of the novel genes present in each putative operon. As shown in Figure 4-3A, although most biological processes made up less than 3 percent of all annotations, a large proportion of genes were found to be related to stress response, cellular response to stimuli or abiotic stimuli, and oxidation-reduction through their GO terms, suggesting that novel genes in these operons may predominantly be active in survival and adaptation to extreme stress conditions. A deeper examination of the molecular functions (Fig. 4-3B) revealed that oxidoreductase activity and transmembrane transporter activity accounted for 10 % and 12 % of all possible functions across all hypothetical operons, reaffirming a strong connection to stress response mechanisms among uncharacterized genes.

Next, delving deeper into the unknown, we examined hypothetical operons of DUF genes, whose properties were completely uncharacterized and most mystifying. Altogether, 324 operons of the 663 putative adapation operons identified consisted of DUF genes, meaning that about half of all operon structures identified in our analysis were of completely uncharacterized, novel genes. Evaluation of the score distributions of GO annotations of DUF operon genes demonstrated, once more, a large proportion of biological processes and molecular functions dedicated to stress response and oxidative reduction. In fact, oxidation-reduction was the most dominant (10 %) of all biological processes for DUF operons (Fig. 4-3C), followed by DNA metabolic processes and cellular reponse to DNA damage stimulus (6 %), suggesting a more pronounced stress response role for DUF genes than for any other hypothetical genes. An overview of the score distribution of molecular functions of DUF operons suggested, by the fractions of oxidoreductase activity, transition metal ion binding, ATPase activity, and transferase activity that made up all molecular mechanisms (Fig. 4-3D), that DUF genes are most likely catalytic binding proteins, in contrast to

the total set of hypothetical genes, which were shown to have a large proportion of transferase, lyase, and transmembrane transporter activity functions (Fig. 4-3B). The results of these GO annotation analyses thus confirm a strong likelihood of hypothetical genes as active participants in stress response mechanisms. Survival to extremophilic condition may be a dominant functional property of DUF genes, as these genes showed a more pronounced proportion of biological and mechanistic processes related to oxidative reduction and response to DNA damage or abiotic stimuli. Thus, further characterization of DUF genes may be of most help in illuminating the adaptation mechanisms necessary for microbial life to persist in otherwise extreme conditions.

4.2.4 Identification of Frequently Occurring, Conserved Hypothetical Genes

Having established global functional traits of hypothetical genes from poorly characterized genomes, we next wanted to concentrate on identifying particularly important, frequently occurring, and evolutionarily conserved novel genes among this complete set of hypothetical genes. To identify genes which were most likely to be important to extremophilic adaptation, we evaluated the frequency of observing each hypothetical gene within operonic structures across metagenome scaffolds, using the description of each gene to identify individual unknown genes. In total, 547 genes were observed across two or more metagenomes and 102 hypothetical genes were observed across five or more metagenomes. Discarding any genes with too broad functions, such as genes described as ”protein conserved in bacteria”, ”membrane transporter”, or ”metallo-beta-lactamase superfamily”, left 94 of the 102 frequently reoccurring hypothetical genes to examine in further detail, to identify a subset of genes that were not only common but conserved among genomes. The ten hypothetical genes with the highest number of observations across metagenomes are shown in Table 4-2. Surprisingly, three out of the top 10 most frequently observed genes were UPF (Unknown Protein Family) or DUF genes (UPF0235, DUF1178 and DUF1801), demonstrating that novel genes with potential adaptation properties are particularly abundant among scaffolds associated with unknown hub taxa of extremophilic aquatic environments.

To evaluate the conservation of these common hypothetical genes, the sequences of each of the 94 genes were clustered with Uniprot sequences of the same descriptions (annotations) by an all-by-all blast to generate protein sequence similarity networks. Afterwards, an alignment score between 60-70 % sequence similarity was chosen to generate the final sequence similarity network (SSN). Hypothetical genes that were found in clusters with Uniprot genes were considered to be highly conserved, due to their high ($\geq$ 65 %) shared sequence identity with proteins found in the well-annotated Uniprot repository. A large proportion (75 %) of the 94 hypothetical proteins were found to be both common and conserved, sharing high sequence similarity (clustering well) with Uniprot genes of the same description. In fact, six of the top ten reoccurring hypothetical genes clustering with high sequence similarity to Uniprot genes of the same name and description (Table 4-2). Of the hypothetical proteins designated as DUFs (34 proteins total), 62 % of all DUFs were found to cluster with high sequence similarity to Uniprot genes, again demonstrating that these novel, potentially essential genes deserve to be studied in further detail due to their high abundance and sequence conservation.

Further gene neighborhood conservation analysis through EFI-GNT of the SSN results (particular cluster numbers in which these hypothetical proteins were found) for the top abundant genes and DUF genes present in at least 2 distinct metagenome scaffolds showed that 98 % of all genes with high sequence conservation also had high genome conservation and recurring gene neighnorhood association among various organisms. For instance, conserved gene neighborhood associations were observed for 83 (DUF302) to 1559 (DUF305) organisms with available genomic information (Supplementary Table 4-1) of a subset of interesting, adaptation-related DUF genes identified through this analysis. The strong gene neighborhood conservation of these DUF genes validated the putative operon structures identified previously in our analysis and also suggested that the functional properties of neighboring genes would be a good proxy for inferring gene function, where, by the guilt-by-association clause of neighboring genes with known functions, we could better understand the ecological roles and particular conditions of particular DUF genes.

Object 4-1. Link for Supplementary Table 4-1

4.2.5 Case Study Examples of Novel Conserved Genes Involved in Adaptation Response

The high proportion of common, conserved genes found using our approach led us to believe that these novel, conserved genes may be particularly important players in processes related to survival and adaptation of extremophilic conditions. To test out this theory, the GO combined graph outputs of the GO annotations of neighboring genes of a select group of conserved, reoccurring hypothetical proteins were evaluated to further unravel the potential functional properties of these seemingly important genes. Here we show two case study examples of two novel genes, DUF1150 and DUF1178, that were each shown to harbor unique functions related to stress and adaptation.

As shown in Figure 4-4, functional characterization of DUF1150 through examination of SSN conservation (Fig. 4-4A) and GO annotation results of neighboring genes revealed, by the presence of heat shock proteins in various DUF1150 operons (Fig. 4-4B) and the high proportion of genes involved in heat shock and copper ion stress response (Fig. 4-4C), that this novel gene may aid survival and adaptation of thermophilic, copper-enriched habitats. On the other hand, the common and conserved DUF1178 gene may be involved in microbial adaptation to oxidative stress conditions, phosphorylation, and generation of biosynthetic products. The neighboring genes of DUF1178 included glutaredoxin and nitrilase genes (Fig. 4-5B) and investigation of the GO annotation terms of neighboring genes in DUF1178 operons revealed a high proportion of genes involved in oxidation-reduction processes, phosphorylation, biosynthetic processes, and DNA repair (Fig. 4-5C). The high proportion of functions related to protein disulfide oxidoreductase activity and glutathione binding (Fig. 4-5D), as well as the presence of glutaredoxins in the conserved gene neighborhood associations of DUF1178 across 183 organisms suggest that DUF1178 may play a strong role in sulfur oxidation adaptation. From these analyses, the two studied genes appear to have distinct roles and may be uniquely suited to particular extremophilic habitats. The biological processes and molecular functions associated with DUF1150 operonic genes suggest that DUF1150 may be particularly ecologically relevant in

extremophilic, copper-rich habitats while DUF1178 may be most active in sulfur and oxygen-enriched environments.

Consequently,these two case study examples show how SSN and GO annotation analyses may be used to provide a global overview of the biological and molecular functional properties to aid characterization of novel genes associated with ecologically relevant unknown microorganisms. Including the case study examples described above, novel functional information (biological processes,molecular functions, associated GO terms) as well as sequence and gene neighborhood conservation were found for 20 frequently occurring DUF genes retrieved from blast search of hub organisms from deep sea, hot springs, hypersaline, and polar microbial communities (Supplementary Table 4-1). By this manner, both novel genes and their predicted functional properties were able to be identified, and the list of these DUF genes and their functional information provides a new avenue for further characterization studies. The results of these analyses suggest that our overall pipeline successfully enables large-scale identification and characterization of novel, essential genes from poorly characterized genomes.

4.3 Discussion

Microbial life is found all over the Earth, even at locations with such extreme environmental conditions that other forms of life are not viable. Understanding how microbes adapted to these conditions is not only of critical importance to understand life and evolution on Earth, but also may be particularly valuable to aid colonization efforts beyond our planet. However, most species in these extreme ecosystems are unknown and/or uncultivable, making them part of what is known as Microbial Dark Matter (MDM) [37]. Furthermore, the vast majority of genes remain uncharacterized for cultivable and novel microorganisms alike, presenting a grand challenge to microbial ecology research. While reconstructed genomes produced from metagenomics and single-cell sequencing data have helped reveal some of the functional diversity present within both known and novel bacterial and archaeal phyla [11, 137, 12, 220], the total diversity of microbial gene products remains challenging due to the fact that most genomes (particularly MAGs and SAGs) suffer from incomplete annotation, varying quality, and contamination

[221, 222]. Prediction of gene function from amplicon sequencing data, while now possible thanks to ancestral reconstruction-based tools [207], is limited to organisms with existing and well-annotated genome annotations, leaving the gene diversity of novel organisms a mystery.

The genomes available today, for MDM and well-characterized species alike, are rife with unknown, poorly annotated genes with unknown functions. A significant proportion of genes in sequenced genomes encode conserved hypothetical proteins [223] and many hypothetical proteins are believed to play essential roles. Yet, functional, phylogenetic, and metabolic annotation of these hypothetical proteins is still extremely challenging. The latest technologies, like PiCRUst [207], paprica [55], or MetaPhlAn [224], cannot identify or annotate novel genes without a complete, well-annotated reference genome. Lastly, though proteins involved in adaptation, including high arsenic resistance [45], have been found within hypothetical proteins from extreme environments, the majority of MDM proteins remain to be functionally annotated or even identified due to the clear lack of representation of most bacterial species in genome sequence databases. In sum, the ecological role and functional properties of MDM and their gene products are great mysteries that must be solved to better understand MDM and Earth's ecosystems as a whole.

In this work, we developed an integrative genomics approach that enabled identification and characterization of novel genes from poorly characteriazed microorganisms by the use of 16S rRNA sequences as probes within metagenome database to recover gene content while bypassing the need for genome representation. Implementation of this bioinformatic pipeline demonstrated that: 1) metagenome scaffolds with high gene content and high sequence identity to query 16S rRNA FASTA sequences could be successfully retrieved for unknown hub microbial taxa; 2)similar and large proportions of hypothetical genes and putative hypothetial operons could be identified; 3) a subset of essential, strongly conserved hypothetical genes could be identified from this group by evaluating the frequency and sequence conservation of this hypothetical gene set; 4) the potential functional properties of essential and conserved genes could be revealed with the help of genome neighborhood information and GO annotation of neighboring well-characterized

genes; 5) further investigation into the specific niches of novel, conserved, essential genes in a computationally efficient and large-scale manner is possible using this approach; and 6) the novel functional and conserved properties for 20 DUF genes acts as a starting point for further exploratory and experimental characterization studies of particularly promising conserved,abundant adaptation-related novel genes.

The successful identification of several highly conserved and frequently occurring novel genes that predominantly function in stress response, including the two case study examples of DUF1150 and DUF1178 genes demonstrate the effectiveness and utility of our approach towards directed characterization efforts of unknown microorganisms. Large-scale functional diversity investigation of poorly characterized taxa is possible through our integrative comparative genomics and network-based strategy. Furthermore, the subset of candidate conserved essential hypothetical genes and their potential functional properties that we were able to identify through our comparative genomics approach will help provide valuable insight on the ecological role of MDM, clarify some of the key functions encoded by these organisms' novel genes, and aid future work in gene inference of novel phyla. The contribution of a subset of essential, conserved adaptation genes may significantly advance research of MDM by helping to clarify in what ways (by which genes or pathways) unknown microbes shape other microbial members, their respective environments, and Earth as a whole. Consequently, conclusions derived from the proposed research may one day even aid detection of other life forms and effective colonization strategies of other planets, thus helping to fundamentally advance both the fields of microbial ecology and space biology exploration.

In summary, by bypassing the need for reference genomes and incorportating the use of a large, curated metagenome database, the computational approach here can not only shed light on the gene diversity present in ecologically relevant unknown bacteria, but may one day help illuminate gene diversity for all microorganisms, including microbial eukaryotes or viruses. Additionally, the list of candidate conserved genes associated with top hub taxa can be used in future analyses, such as integrative multi-omics analyses, to provide more insight into the

activities and dynamics of unknown microorganisms to a diverse array of ecosystems, conditions and perturbations, thus illuminating the roles and attributes of the hidden microbial world while simultaneously improving knowledge of microbial ecology and evolution.

4.4 Methods

4.4.1 Metagenome Hub Blast Analysis

4.4.1.1 16S hub data

The 16S rRNA fasta sequences of the top 20 known and unknown hubs (found by calculating hub score of each node present in the 'Original network') of the hot springs, hypersaline, deep sea, and polar microbial co-occurrence networks (created in our previous analysis) were retrieved using the subseq function from the SEQTK toolkit (https://github.com/lh3/seqtk) and the fasta file of complete sequences (new_refseqs.fna) produced by the QIIME pick_open_reference_otus.py script.

4.4.1.2 Metagenome data

8,365 assembled metagenome files and other corresponding metagenome information for 6,396 publicly available metagenomes were retrieve from the IMG/M server using the JGI Genome Portal API and a custom python script, with all resultant information stored to HiPerGator for further processing. All assembled.fna metagenome files were concatenated and assembled into two databases using the makeblastdb function from the ncbi_blast module (v.2.2.30) (parameter -dbtype equal to nucl).

4.4.1.3 Retrieval of metagenome scaffolds with high-confidence match to 16S hubs

All known and unknown hub sequences for each environment were blasted against each of the two metagenome databases using the blastn function, with the percent identity parameter (perc_identity) set to 95 to recover metagenome scaffolds that had at least 95 percent sequence similarity to the 16S rRNA hub fasta sequence query. The blast output format (outfmt) parameter was set to include query sequence id (qseqid), subject sequence id (sseqid), percent identity (pident), length mismatch, open gap (gapopen), query start position (qstart), sequence start position (sstart), sequence end position (send), e-value (evalue), bitscore, and query coverage

(qcovs) to identify scaffolds that had high query coverage. All sequences of the high-confidence metagenome scaffold hits from the blast output were then retrieved from each of the two metagenome databases using the blastdbcmb function.

4.4.2 Scaffold Protein Identification and Annotation

Prodigal (v.2.6.3) was used to predict the gene and protein content of each metagenome scaffold, using the prodigal function and -p meta parameter for metagenomes. The prodigal output returned, for each scaffold, a list of predicted proteins with corresponding start codon, GC content, start coordinate, end coordinate, and strand position for each protein. All predicted proteins present in each scaffold were then functionally annotated using the eggnog-mapper tool (v.2.0.1), which uses the eggNOG public database of orthology relationships, gene evolutionary histories and functional annotations of bacteria, archaea, eukaryotes, and viruses and pfam (protein families) database as reference guides. The total protein content of all scaffolds was calculated and only high-confidence scaffolds consisting of at least 30 proteins were used in further analyses.

4.4.3 Identification of Hypothetical Proteins, Adaptation-related Proteins, and Putative Adaptation Operons

All corresponding protein coordinate and annotation files produced from the prodigal and eggnog-mapper analyses were imported into R (v.3.6.3) for further analysis using the rbindlist() function from the data.table (v.1.12.8) package. The COG category of each predicted protein was used to differentiate 'hypothetical proteins' (any protein with a COG category of S or R(with no known function or only domain function) from proteins with known functions. A list of keywords related to metabolic and extreme environmental stress response functions retrieved from literature was used to parse for adaptation gene matches among all genes present on each strand (Supplementary Table 4-1). To identify closely related genes, hierarchical clustering was performed based on the distance between gene start and end coordinate positions of each gene pair using the packages ape and dendextend, and the function hclust()in base R. Closely clustered genes on a single branch represented putative operons and any ten genes within 5000 bp or less to

one another were considered to belong to one operon. If targeted genes were among the ten closest neighbors (i.e., less than 5000 bp away) to a hypothetical gene, we defined this set of hypothetical and potentially extreme stress adaptation-related neighboring genes as a putative adaptation operon.

4.4.4 Sequence Similarity Conservation Analysis of Hypothetical Proteins

4.4.4.1 Selection of hypothetical proteins

The frequency of hypothetical protein occurrence in putative adaptation operons was calculated in R using the dplyr package, with hypothetical protein frequency across metagenomes visualized as heatmaps and networks using the packages ggplot2 and igraph. Networks of the protein neighborhood of each hypothetical gene were also created using the package igraph. The protein sequences of all hypothetical proteins that reappeared at least five times in putative adaptation operons across metagenome scaffolds were retrieved from the original prodigal output for further sequence similarity analysis to identify a list of candidate proteins that were common and highly conserved.

4.4.4.2 Retrieval of uniprot proteins

The output from eggnog provided for each gene, regardless of its COG category, a predicted description of its function. Thus, even for hypothetical proteins (proteins with an S COG category), a domain of unknown function (DUF), protein of unknown function, or superfamily was described in the description category. The description of each frequently occurring hypothetical protein was saved to a text file and used as input to retrieve the protein sequences of all the UNIPROT terms that had a matching description in fasta format, through the python requests package and the UNIPROT API. For the purposes of the sequence similarity analysis, all UNIPROT-derived proteins will be used as the reference 'known' group and the hypothetical protein sequences from the metagenome scaffold analysis will be referred to as the 'hypothetical' group.

4.4.4.3 Sequence similarity network analysis

For each hypothetical protein description, the UNIPROT protein sequences and the hypothetical protein sequences (labeled with the corresponding environment and hub type the metagenome scaffold matched to in the metagenome blast analysis) were concatenated into one fasta file to be used as input for the creation of a sequence similarity network (SSN). 114 total files were created. The Enzyme Similarity (EFI-EST) tool was used to perform an all-vs-all blast on all sequences found in each fasta file to create the sequence similarity network (SSN). An alignment score threshold equivalent to a minimum protein sequence similarity of 60-65 % was used to draw edges connecting proteins in the final SSN. The EFI Color SSN utility and Genome Neighborhood Tool features were then used to color protein clusters and identify neighboring genes and operon structures for each protein cluster present in the SSN.

Each SSN was first visualized in Cytoscape to find clusters containing both known UNIPROT proteins and hypothetical proteins. The protein sequences of each cluster were saved and exported for multiple sequence alignment. Any hypothetical protein found within a large cluster of known proteins was considered to be conserved and highly similar.

Each SSN was also imported into R using the rvest package (version 0.3.5) to facilitate automated cluster identification by alignment score and sequence description (label). A text file of all node ids per cluster was also imported into R to identify clusters containing hypothetical genes (hypothetical gene cluster) and the sqlite output file of the Genome Neighborhood Tool was imported to R using the RSQLite (v.2.2.0) and DBI packages (v.1.1.0) to enable identification of all genes located in the same genomic neighborhood as the hypothetical gene cluster.

4.4.5 GO Annotation Analysis of Hypothetical Proteins

For each operon, all available gene ontology (GO) terms for genes with existing GO information were used to provide a functional characterization of each hypothetical gene. An .annot file was created, using each hypothetical gene description as the gene id, for Blast2GO [219] functional annotation. Then, the combined graph option within Blast2GO was selected to map the biological processes and functional properties (molecular functions) of all genes found

within operons containing a particular hypothetical gene. Multi-level pie charts and bar charts were then created for the biological process and molecular function combined graphs to provide a general description of the functional distribution of all genes.

Table 4-1. Overview of blast and functional annotation results

Query Type	Total Metagenome Hits	Total Gene-rich Hits	Average Gene Count	Average Hypothetical Gene Count (%)	Average Adaptation Gene Count (%)	Total Adaptation Operon Count
HS-K	411	14	238	17.8	8.6	75
HS-UNK	589	47	346	17.1	8.2	344
HY-K	2435	21	103	15.4	8.6	52
HY-UNK	22	2	34	11.8	8.8	2
DS-K	3401	26	237	19.8	8.0	174
DS-UNK	2518	16	145	15.9	8.3	72
PO-K	4088	10	199	19.9	8.0	56
PO-UNK	1298	2	71	28.8	7.1	2

HS: Hot Springs; HY: Hypersaline; DS:Deep Sea; PO:Polar; K:Known; UNK:Unknown

Table 4-2. Top reoccurring hypothetical genes among metagenomes

Gene Rank	Gene Name	Gene Description	Number of Observations	Conserved
1	NA	Activator of HSP90 ATPase homolog 1-like protein	15	No
2	NA	Tripartite tricarboxylate transporter family receptor	14	No
3	YggT	YggT family	14	Yes
4	RamA	Nitrilase cyanide hydratase and apoliprotein N-acyltransferase	13	Yes
5	NA	UPF0235 family	13	Yes
6	NA	DUF1801	12	No
7	NA	Esterase-like activity of phytase	12	No
8	NA	DUF1178	12	Yes
9	LemA	LemA family	11	Yes
10	TctB	Tripartite tricarboxylate transporter TctB family	11	Yes

NA: No Annotation Available

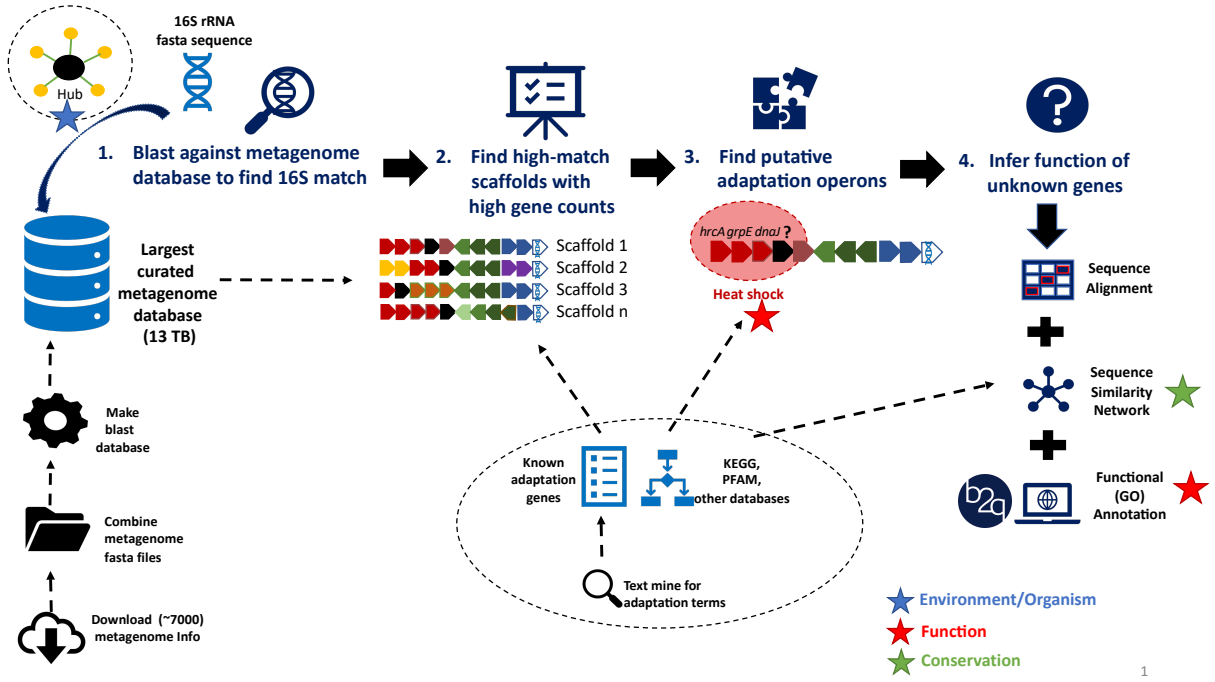


Figure 4-1. Overview of the hub blast pipeline. A blastn search is performed on the 16S rRNA fasta sequence of a microorganism identified as a top hub by a previous network analysis against a large curated metagenome database (made from publicly available IMG/M metagenomes) to identify high-confidence scaffolds. Gene and protein content of each scaffold hit is predicted, using a combination of existing databases and tools and a list of adaptation terms from literature, to retain high-confidence scaffolds with a sufficiently large number of genes, including hypothetical genes with no known functions (black). Putative adaptation operons consisting of both known and hypothetical genes found within a close distance (5000 bp) to one another are identified. The functions of the neighboring genes in the operon and further bioinformatics analyses(i.e. sequence similarity networks and multiple sequence alignments) enable prediction of the functional properties and identification of the conservation of each hypothetical gene. Stars highlight contexts of gene importance (environment/organism, stress-related function, and conservation).

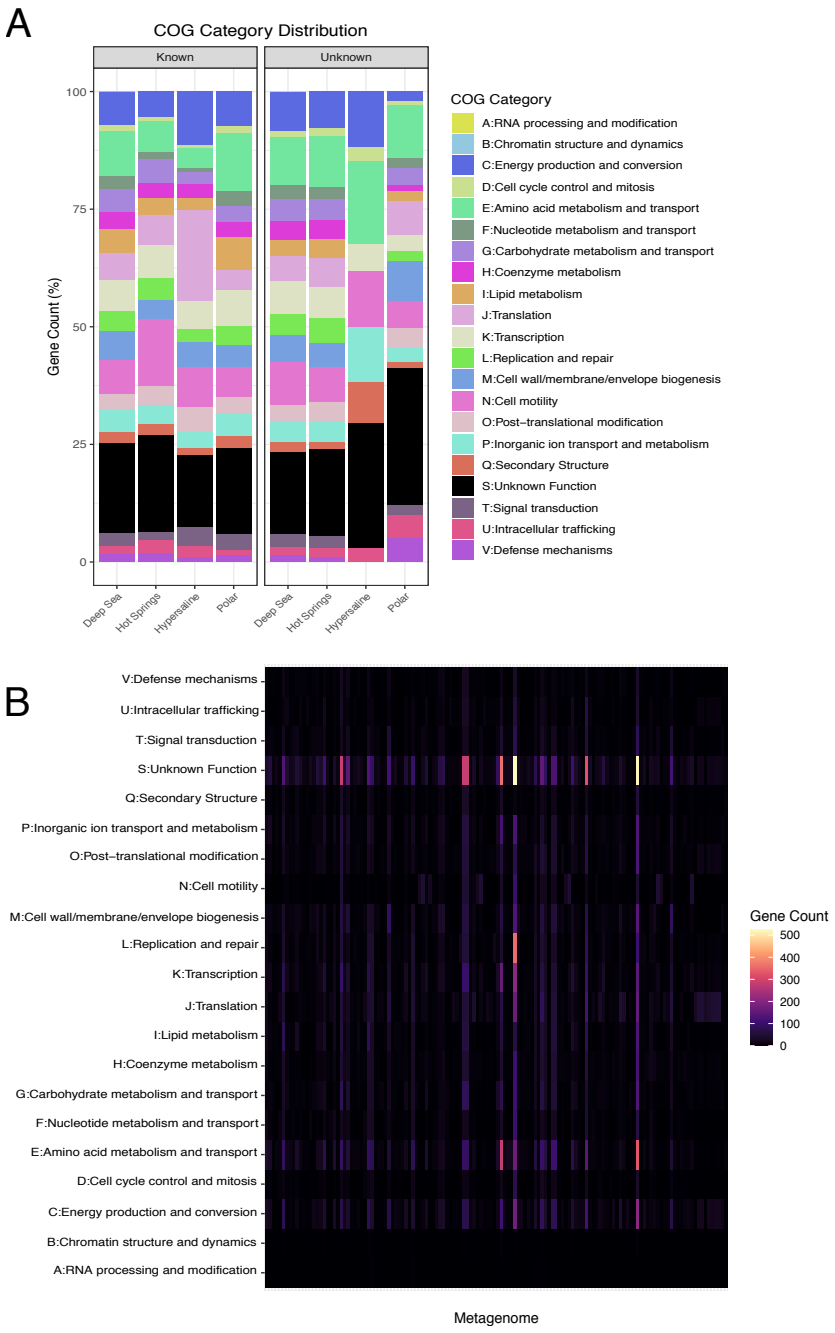


Figure 4-2. Global and individual COG category distribution among metagenomes. A. The proportion of genes belonging to each Cluster of Orthologous Groups (COG) category is shown as a percentage per environment and classification status (Known or Unknown) of all hub blast query sequences. Colored bars represent COG categories. B. Heatmap of the gene abundance (in absolute counts) of each COG category for each high-confidence gene-rich metagenome, using a blue-yellow color-scale mapping to represent low (blue) and high (yellow) gene count respectively.

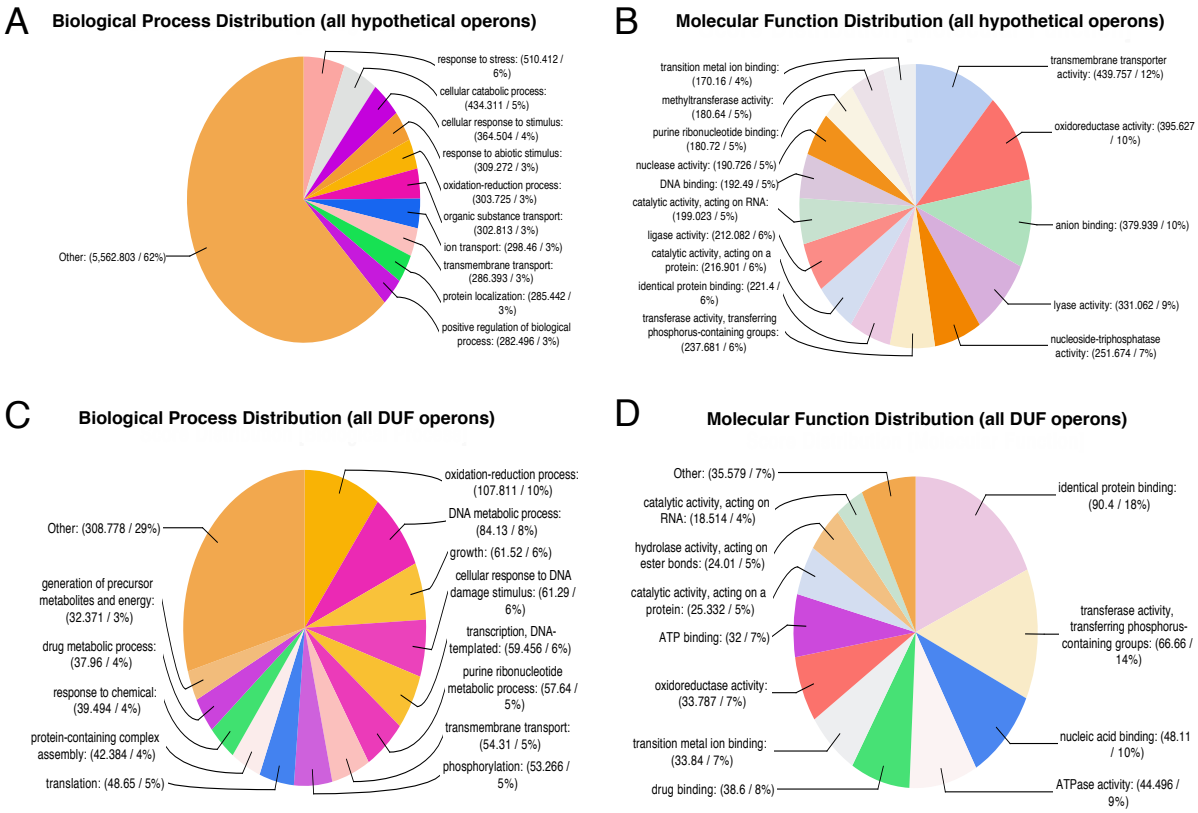


Figure 4-3. Overview of functional properties of hypothetical and DUF operons. A. Pie chart of the distribution of biological processes inferred from GO annotation among all hypothetical operons. B. Pie chart of the distribution of molecular functions inferred from GO annotation among all hypothetical operons. C. Pie chart of the distribution of biological processes inferred from GO annotation among DUF operons only. D. Pie chart of the distribution of molecular functions present inferred from GO annotation among DUF operons only.

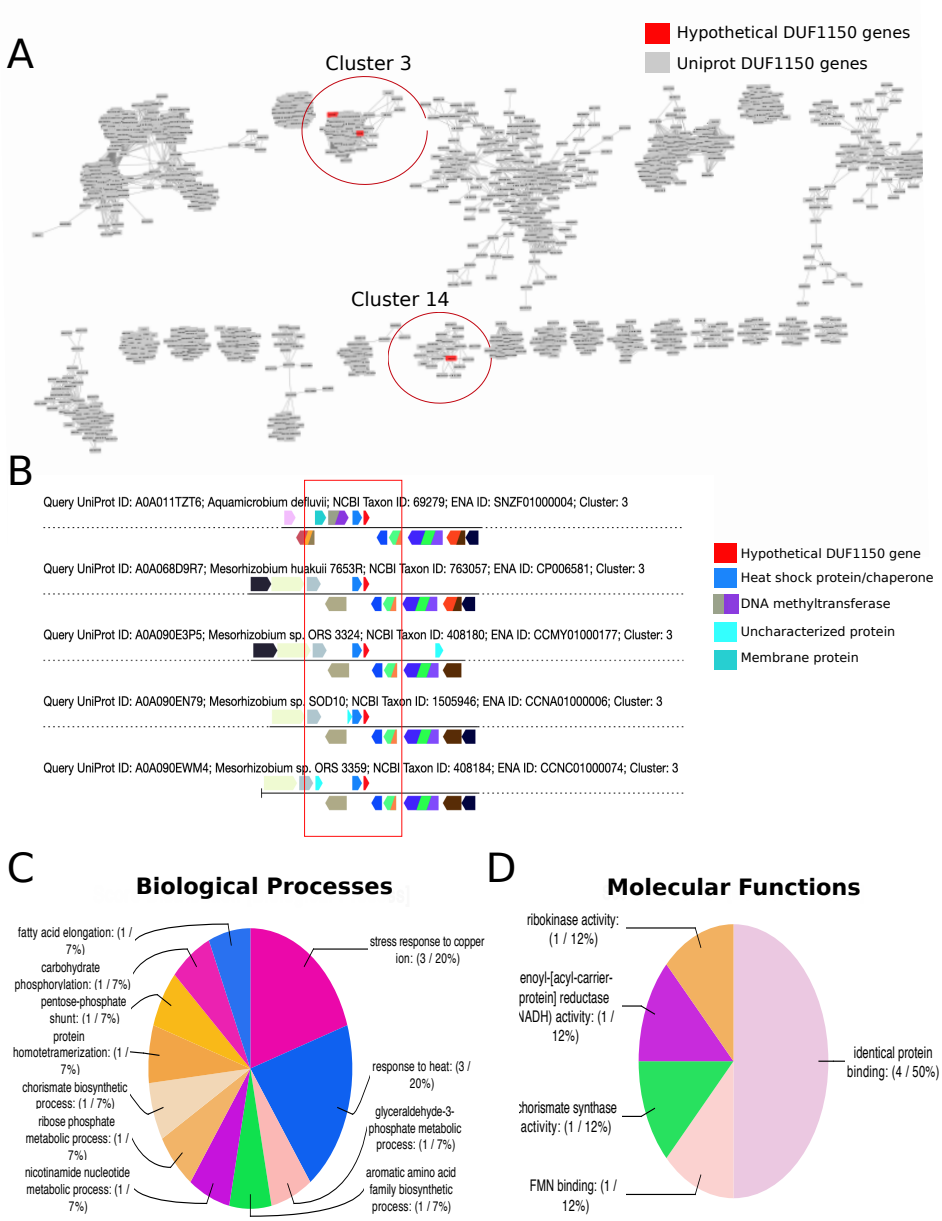


Figure 4-4. Functional characterization of DUF1150. A. Sequence Similarity Network of DUF1150 Uniprot (gray) and hypothetical (red) genes produced from our hub blast approach. The largest clusters are shown, with clusters containing hypothetical genes outlined in red and annotated. B. Genome neighborhood diagrams of Cluster 3 DUF1150 genes found in the genomes of five different organisms. DUF1150-containing operons are outlined in red, with the function of the neighboring genes of DUF1150 described in the color legends. C. Pie chart of the score distribution of biological processes of hypothetical DUF1150 operons. D. Pie chart of the score distribution of molecular functions of hypothetical DUF1150 operons.

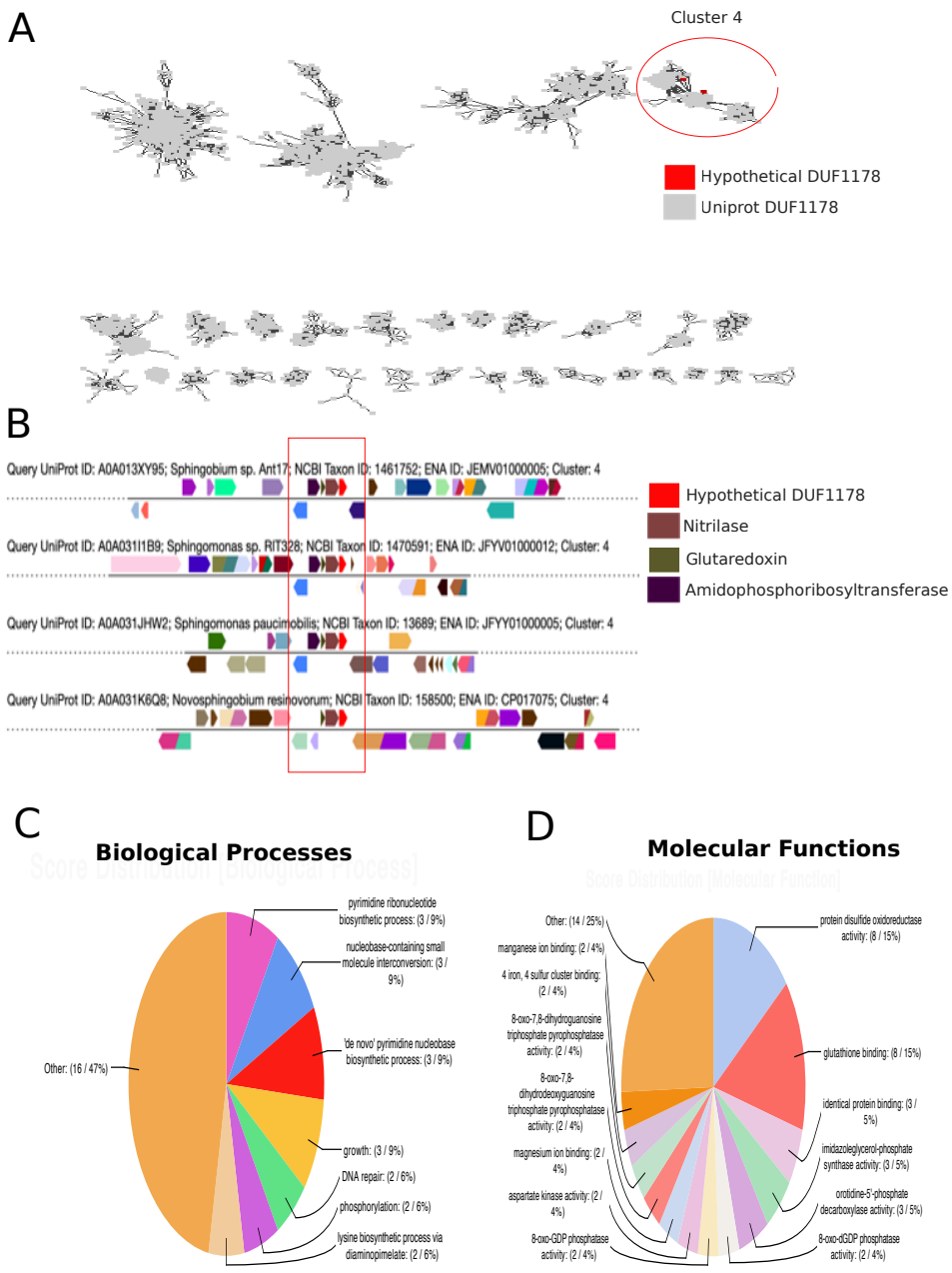


Figure 4-5. Functional characterization of DUF1178. A. Sequence Similarity Network of DUF1178 Uniprot (gray) and hypothetical (red) genes produced from our hub blast approach. The largest clusters are shown, with clusters containing hypothetical genes outlined in red. B. Genome neighborhood diagrams of Cluster 4 DUF1178 genes found in the genomes of five different organisms. DUF1178-containing operons are outlined in red, with the function of the neighboring genes described in the color legends. C. Pie chart of the score distribution of biological processes of hypothetical DUF1178 operons. D. Pie chart of the score distribution of molecular functions of hypothetical DUF1178 operons.

CHAPTER 5
SUMMARY AND CONCLUSIONS

Due to improvements in next-generation sequencing technologies and the efforts of integrated metagenomics and single-cell genomics analyses, a plethora of novel genomes have been produced (reconstructed), revealing several unique, novel bacterial and archaeal phyla [225, 101, 59] in the process. However, this tremendous growth of new genomic sequencing information has only increased the fraction of microbial dark matter (MDM) for researchers to analyze, as the majority of the recovered novel organisms and gene products remain poorly annotated and uncharacterized. In fact, as much as 80% of all archaeal genes encode 'hypothetical' proteins or genomic 'dark matter' [226]. The substantial fraction of unknown genes found in all microbial genomes is only expected to grow with increased genomic resolution [227], causing functional characterization of these unknown components to be an urgent concern. The finding that a portion of these unknown genes, particularly families of genes called 'domains of unknown function', are conserved [223] and potentially essential for microbial survival and adaptation [226] only escalates the need to develop strategies that enable functional characterization in a cost-effective, efficient, and scalable manner. Microbial ecology research in particular stands to benefit from advancements in functional characterization efforts, as complete understanding of the origin and evolution of microbial life on Earth is currently impeded by the lack of knowledge concerning the biological and functional impact of MDM on ecosystem function.

Unraveling the role of MDM, particularly MDM microbes from extreme conditions, is key to uncover the relevance of these microbes to their communities and subsequently unlock the adaptation strategies needed to survive harsh conditions within and even beyond Earth. Consequently, there is a crucial need to ascertain the adaptive properties and roles MDM possess to adapt to extreme environmental conditions, without which we will never be able to complete the puzzle on the origins and evolution of life. Accordingly, our objective in this work was to enhance understanding of the ecological and functional role of MDM in extreme aquatic environments by means of a computationally intensive approach.

We first developed a robust and scalable network-based computational pipeline to determine the local and global impact of MDM. Our results showed that the network metrics degree, betweenness, closeness, and hub score can each be used to effectively quantify the importance of unknown taxa. Furthermore, validation of this methodology on four different network estimation strategies (SpiecEASI, SparCC, CCLasso, and Pearson) and different sample prevalence thresholds (ranging from 20 to 40 percent prevalence) showed that changes in network estimation or sample prevalence highly affect the resultant network structure and topology, yet produce similar results regarding the measured impact of unknown taxa by different network metrics. Consequently, we found that differences in correlation estimation methodologies or sample prevalence should not statistically significantly impact the relative ecological relevance of unknown taxa, making this approach both sound and applicable to future analyses, including to large-scale meta-analyses and to studies that implement improved, state-of-the-art microbial network construction methodologies.

Application of this network-based approach to a diverse 16S 1086-sample dataset encompassing four different extreme environments (hot springs, hypersaline, deep sea and polar communities) demonstrated that 1) unknowns are as prevalent and abundant as known microorganisms and therefore merit inclusion in microbial association networks; 2) unknown organisms play central and integral roles within extreme environmental networks, highly impacting both network structure and taxa co-occurrence by significantly decreasing overall degree and betweenness values when removed; 3) unknown organisms at genus level that predominantly featured as top hubs across all environments are particularly ecologically relevant and adapted to their extreme environments due to their high abundance and associations with other taxa; and 4) network metrics, like hub score, can be used to prioritize important unknown taxa for further characterization. This approach resulted in a comprehensive list of top-scoring hub candidate microorganisms to target for future characterization efforts.

We hypothesized that these top hub unknown microorganisms are equipped with an abundance of unique and novel adaptation mechanisms to survive and thrive within otherwise

inhospitable conditions. To evaluate this theory, we used 20 of the top-scoring hub known and unknown taxa from each of the four extreme environments studied previously as probes to identify and characterize novel adaptation genes via an integrated metagenomics and comparative genomics approach, thus bypassing the requirement of a reference genome that limits the capacity of other existing gene prediction pipelines to infer function of novel taxa. Our bioinformatics pipeline proved successful in retrieving hundreds of metagenome scaffolds that matched at high sequence identity to the query hub 16S sequences. Though a majority of these scaffolds contained few genes, the scaffolds that met our 30 gene count threshold contained large proportions (hundreds to thousands) of hypothetical genes (designated as functional COG category S). Regardless of the original environment or classification status of the query hub sequence, a similar and relatively high percentage of hypothetical genes, adaptation-related genes, and putative hypothetical operons was detected, suggesting that our approach may be useful for discovery and functional inference of both well-characterized microorganisms and novel microorganisms that lack genomic representation. Evaluation of the GO terms associated to genes found in all hypothetical operons and DUF-specific operons revealed a strong presence of genes involved in stress response mechanisms and oxidoreductase, transmembrane transporter, and transition metal ion binding processes. The large proportion of GO terms found to be associated to biological processes and molecular functions implicated in stress and the strong clustering at high sequence identity to existing hypothetical, DUF Uniprot genes within sequence similarity networks demonstrate that the hypothetical genes associated to ecologically relevant microorganisms are indeed conserved and essential for microbial adaptation to extremophilic conditions. As a result of this approach, we provide a subset of frequently occurring and conserved hypothetical genes for continued characterization studies.

We showed, as proof-of-concept, that functional characterization at the level of an individual hypothetical gene is possible through the case study examples of DUF1150 and DUF1178 genes. The results showed that conserved, essential hypothetical genes play distinct ecological roles due to differences in the proportion of genes assigned to certain biological

processes and molecular functions. In our examples, the large percentage of genes attributed to heat shock and copper ion stress response within DUF1150 operons and the large percentage of oxidation-reduction and protein disulfride oxidoreductase activity binding proteins found within DUF1178 genes suggest that DUF1150 genes play active roles in adaptation to thermophilic, copper-enriched habitats while DUF1178 genes function primarily in adaptation to oxidative stress conditions. Consequently, both global and protein-specific functional characterization of hypothetical genes can be achieved with the help of genome neighborhood information, GO annotation of neighboring well-characterized genes and SSN evaluation. Large-scale functional diversity investigation of poorly characterized taxa is thus possible through our integrative comparative genomics and network-based strategy.

In summary, the bioinformatics pipelines generated in this work enable large-scale discovery and ecological and functional inference of novel organisms currently lacking genomic representation. Unlike most MDM studies to-date, focus was placed on the community position and relationships of MDM and functional inference was performed by computational analysis of metagenome databases instead of reference bacterial genomes alone. To the best of our knowledge, our 16S rRNA dataset was unprecedented in size (over 1000 samples) and scope (four distinct extreme environments), thereby providing a more global and thorough overview on the ecological niches and functional roles associated with MDM. The subset of ecologically relevant candidate bacteria (identified from our network-based approach) and the subset of conserved hypothetical genes with essential roles in stress adaptation (identified from the metagenome hub blast analysis) will aid continued characterization efforts of MDM, thereby helping to bridge the gap between the number of sequenced and cultivated organisms. We showed that MDM do significantly impact various extreme aquatic environments and that a large number of conserved hypothetical genes essential for microbial survival and adaptation can be identified by using top hub MDM as probes. This work provides valuable insight into the ecological role of MDM, helps clarify key functions encoded by novel genes through two examples, and aids future work in gene inference of novel phyla by the contribution of a subset of candidate taxa and genes. The results

of this study establish a methodology to further clarify how and why unknown microbes shape and adapt to ecosystems within and beyond Earth. Incorporation of this computationally intensive approach to future large-scale, integrative, and multi-omics analyses may thereby improve existing knowledge of microbial biodiversity and evolution, illuminating the currently hidden microbial world.

[1] Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 6578–83 (1998). URL http://www.ncbi.nlm.nih.gov/pubmed/9618454http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC33863.

[2] Pedrós-Alió, C. & Manrubia, S. The vast unknown microbial biosphere (2016).

[3] Campbell, B. J., Yu, L., Heidelberg, J. F. & Kirchman, D. L. Activity of abundant and rare bacteria in a coastal ocean. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 12776–81 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21768380http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3150899.

[4] Bull, A. T. *et al.* High altitude, hyper-arid soils of the Central-Andes harbor mega-diverse communities of actinobacteria. *Extremophiles* **22**, 47–57 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29101684http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5770506.

[5] Fernandez, A. B. *et al.* Microbial diversity in sediment ecosystems (evaporites domes, microbial mats, and crusts) of Hypersaline Laguna Tebenquiche, Salar de Atacama, Chile. *Frontiers in Microbiology* (2016).

[6] Baricz, A. *et al.* Spatio-temporal insights into microbiology of the freshwater-to-hypersaline, oxic-hypoxic-euxinic waters of Ursu Lake. *Environmental Microbiology* (2019). URL http://www.ncbi.nlm.nih.gov/pubmed/31894632.

[7] Rashid, M. & Stingl, U. Contemporary molecular tools in microbial ecology and their application to advancing biotechnology. *Biotechnology Advances* **33**, 1755–73 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26409315.

[8] Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, N.Y.)* **304**, 66–74 (2004). URL http://www.ncbi.nlm.nih.gov/pubmed/15001713.

[9] Castelle, C. J. *et al.* Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nature communications* **4**, 2120 (2013).

[10] Hedlund, B. P., Dodsworth, J. A. & Staley, J. T. The changing landscape of microbial biodiversity exploration and its implications for systematics. *Systematic and Applied Microbiology* **38**, 231–236 (2015).

[11] Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015). nature14486.

[12] Hedlund, B. P., Dodsworth, J. A., Murugapiran, S. K., Rinke, C. & Woyke, T. Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter" (2014). URL https://pubmed.ncbi.nlm.nih.gov/25113821/.

[13] Dodsworth, J. A. *et al.* Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nature Communications* **4**, 1854 (2013). URL http://www.nature.com/articles/ncomms2884.

[14] Stubbendieck, R. M., Vargas-Bautista, C. & Straight, P. D. Bacterial Communities: Interactions to Scale. *Frontiers in Microbiology* **7**, 1234 (2016). URL http://journal.frontiersin.org/Article/10.3389/fmicb.2016.01234/abstract.

[15] Banerjee, S., Schlaeppi, K. & van der Heijden, M. G. Keystone taxa as drivers of microbiome structure and functioning (2018). URL www.nature.com/nrmicro.

[16] Fisher, C. K. & Mehta, P. Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression. *PLoS ONE* **9**, e102451 (2014). URL http://dx.plos.org/10.1371/journal.pone.0102451.

[17] Agler, M. T. *et al.* Microbial Hub Taxa Link Host and Abiotic Factors to Plant Microbiome Variation. *PLOS Biology* **14**, e1002352 (2016). URL https://dx.plos.org/10.1371/journal.pbio.1002352.

[18] Berry, D. & Widder, S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology* **5**, 219 (2014). URL http://journal.frontiersin.org/article/10.3389/fmicb.2014.00219/abstract.

[19] Vigneron, A. *et al.* Contrasting winter versus summer microbial communities and metabolic functions in a permafrost thaw lake. *Frontiers in Microbiology* **10**, 1656 (2019). URL http://www.ncbi.nlm.nih.gov/pubmed/31379798http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6646835.

[20] Briggs, B. R. *et al.* Seasonal patterns in microbial communities inhabiting the hot springs of Tengchong, Yunnan Province, China. *Environmental microbiology* **16**, 1579–1591 (2014). URL http://doi.wiley.com/10.1111/1462-2920.12311.

[21] Comeau, A. M., Harding, T., Galand, P. E., Vincent, W. F. & Lovejoy, C. Vertical distribution of microbial communities in a perennially stratified Arctic lake with saline, anoxic bottom waters. *Scientific Reports* **2**, 604 (2012). URL http://www.nature.com/articles/srep00604.

[22] KLEPAC-CERAJ, V. *et al.* Microbial diversity under extreme euxinia: Mahoney Lake, Canada. *Geobiology* **10**, 223–235 (2012). URL http://doi.wiley.com/10.1111/j.1472-4669.2012.00317.x.

[23] Lugo-Martinez, J., Ruiz-Perez, D., Narasimhan, G. & Bar-Joseph, Z. Dynamic interaction network inference from longitudinal microbiome data. *Microbiome* **7**, 54 (2019). URL https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0660-3.

[24] Thompson, A. W. *et al.* Dynamics of Prochlorococcus Diversity and Photoacclimation During Short-Term Shifts in Water Column Stratification at Station ALOHA. *Frontiers in*

Marine Science **5**, 488 (2018). URL https://www.frontiersin.org/article/10.3389/fmars.2018.00488/full.

[25] Hua, Z. S. *et al.* Insights into the ecological roles and evolution of methyl-coenzyme M reductase-containing hot spring Archaea. *Nature Communications* **10**, 1–11 (2019).

[26] Delgado-Baquerizo, M. *et al.* A global atlas of the dominant bacteria found in soil. *Science (New York, N.Y.)* **359**, 320–325 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29348236.

[27] Mandakovic, D. *et al.* Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. *Scientific Reports* **8**, 5875 (2018). URL http://www.nature.com/articles/s41598-018-23931-0.

[28] Cruaud, P. *et al.* Comparative Study of Guaymas Basin Microbiomes: Cold Seeps vs. Hydrothermal Vents Sediments. *Frontiers in Marine Science* **4**, 417 (2017). URL http://journal.frontiersin.org/article/10.3389/fmars.2017.00417/full.

[29] Glass, J. B. *et al.* Adaptations of Atribacteria to life in methane hydrates: hot traits for cold life. *bioRxiv* 536078 (2019). URL https://www.biorxiv.org/content/10.1101/536078v1.abstract.

[30] Uritskiy, G. & Di Ruggiero, J. Applying genome-resolved metagenomics to deconvolute the halophilic microbiome (2019).

[31] Colson, P., La Scola, B., Levasseur, A., Caetano-Anollés, G. & Raoult, D. Mimivirus: leading the way in the discovery of giant viruses of amoebae. *Nature Reviews Microbiology* **15**, 243–254 (2017). URL http://www.nature.com/articles/nrmicro.2016.197.

[32] Kauffman, K. M. *et al.* A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature* **554**, 118–122 (2018).

[33] Hug, L. A. *et al.* Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. *Environmental Microbiology* **18**, 159–173 (2016). URL http://doi.wiley.com/10.1111/1462-2920.12930.

[34] Hugenholtz, P., Pitulle, C., Hershberger, K. L. & Pace, N. R. Novel division level bacterial diversity in a Yellowstone hot spring. *Journal of bacteriology* **180**, 366–76 (1998). URL http://www.ncbi.nlm.nih.gov/pubmed/9440526http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC106892.

[35] Dudek, N. K. *et al.* Novel Microbial Diversity and Functional Potential in the Marine Mammal Oral Microbiome. *Current biology : CB* **27**, 3752–3762.e6 (2017).

[36] Eloe-Fadrosh, E. A. *et al.* Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs (2016).

[37] Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013). URL https://pubmed.ncbi.nlm.nih.gov/23851394/.

[38] Adam, P. S., Borrel, G., Brochier-Armanet, C. & Gribaldo, S. The growing tree of Archaea: New perspectives on their diversity, evolution and ecology (2017).

[39] Chrismas, N. A. M., Barker, G., Anesio, A. M. & Sánchez-Baracaldo, P. Genomic mechanisms for cold tolerance and production of exopolysaccharides in the Arctic cyanobacterium Phormidesmis priestleyi BC1401. *BMC Genomics* **17**, 533 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27485510http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4971617http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2846-4.

[40] Singh, A. H., Doerks, T., Letunic, I., Raes, J. & Bork, P. Discovering functional novelty in metagenomes: Examples from light-mediated processes. *Journal of Bacteriology* **91**, 32–41 (2009).

[41] Lobb, B., Kurtz, D. A., Moreno-Hagelsieb, G. & Doxey, A. C. Remote homology and the functions of metagenomic dark matter. *Frontiers in Genetics* **6** (2015).

[42] Knapik, K., Becerra, M. & González-Siso, M.-I. Microbial diversity analysis and screening for novel xylanase enzymes from the sediment of the Lobios Hot Spring in Spain. *Scientific Reports* **9**, 11195 (2019). URL http://www.nature.com/articles/s41598-019-47637-z.

[43] Butterfield, C. N. *et al.* Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. *PeerJ* **4**, e2687 (2016). URL https://peerj.com/articles/2687.

[44] Zhang, X. *et al.* Comparative genomics unravels metabolic differences at the species and/or strain level and extremely acidic environmental adaptation of ten bacteria belonging to the genus Acidithiobacillus. *Systematic and Applied Microbiology* **39**, 493–502 (2016). URL https://www.sciencedirect.com/science/article/pii/S0723202016300923.

[45] da Costa, W. L. O. *et al.* Functional annotation of hypothetical proteins from the Exiguobacterium antarcticum strain B7 reveals proteins involved in adaptation to extreme environments, including high arsenic resistance. *PLOS ONE* **13**, e0198965 (2018). URL https://dx.plos.org/10.1371/journal.pone.0198965.

[46] Ramadan, E. *et al.* Molecular Adaptations of Bacterial Mercuric Reductase to the Hypersaline Kebrit Deep in the Red Sea. *Applied and environmental microbiology* **85**, e01431–18 (2019). URL http://www.ncbi.nlm.nih.gov/pubmed/30504211.

[47] Fierer, N. & Jackson, R. B. The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 626–31 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/16407148http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1334650.

[48] Zinger, L. *et al.* Global Patterns of Bacterial Beta-Diversity in Seafloor and Seawater Ecosystems. *PLoS ONE* **6**, e24570 (2011). URL http://dx.plos.org/10.1371/journal.pone.0024570.

[49] Gaidos, E., Rusch, A. & Ilardo, M. Ribosomal tag pyrosequencing of DNA and RNA from benthic coral reef microbiota: community spatial structure, rare members and nitrogen-cycling guilds. *Environmental Microbiology* **13**, 1138–1152 (2011). URL http://doi.wiley.com/10.1111/j.1462-2920.2010.02392.x.

[50] Wong, H., Ahmed-Cox, A. & Burns, B. Molecular Ecology of Hypersaline Microbial Mats: Current Insights and New Directions. *Microorganisms* **4**, 6 (2016). URL http://www.mdpi.com/2076-2607/4/1/6.

[51] Finstad, K. M. *et al.* Microbial Community Structure and the Persistence of Cyanobacterial Populations in Salt Crusts of the Hyperarid Atacama Desert from Genome-Resolved Metagenomics. *Frontiers in microbiology* **8**, 1435 (2017).

[52] Andrei, A. *et al.* Hypersaline sapropels act as hotspots for microbial dark matter. *Scientific Reports* **7**, 6150 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28733590http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5522462.

[53] Nobu, M. K. *et al.* Microbial dark matter ecogenomics reveals complex synergistic networks in a methanogenic bioreactor. *ISME Journal* **9**, 1710–22 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25615435http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4511927.

[54] Lambrechts, S., Willems, A. & Tahon, G. Uncovering the Uncultivated Majority in Antarctic Soils: Toward a Synergistic Approach. *Frontiers in Microbiology* **10**, 242 (2019). URL https://www.frontiersin.org/article/10.3389/fmicb.2019.00242/full.

[55] Bowman, J. S. Identification of Microbial Dark Matter in Antarctic Environments. *Frontiers in Microbiology* **9**, 3165 (2018). URL https://www.frontiersin.org/article/10.3389/fmicb.2018.03165/full.

[56] Farag, I. F., Davis, J. P., Youssef, N. H. & Elshahed, M. S. Global patterns of abundance, diversity and community structure of the aminicenantes (Candidate Phylum OP8). *PLoS ONE* **9**, e92139 (2014). URL http://dx.plos.org/10.1371/journal.pone.0092139.

[57] Becraft, E. D. *et al.* Rokubacteria: Genomic giants among the uncultured bacterial phyla. *Frontiers in Microbiology* **8**, 2264 (2017). URL https://www.frontiersin.org/articles/10.3389/fmicb.2017.02264/full. bioRxiv.

[58] Youssef, N. H., Couger, M. B., McCully, A. L., Criado, A. E. G. & Elshahed, M. S. Assessing the global phylum level diversity within the bacterial domain: A review. *Journal of advanced research* **6**, 269–82 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26257925http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4522544.

[59] Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* **2**, 1 (2017). URL https://www.nature.com/articles/s41564-017-0012-7.

[60] Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications* **7** (2016).

[61] Castelle, C. J. & Banfield, J. F. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* **172**, 1181–1197 (2018).

[62] Steen, A. D. *et al.* High proportions of bacteria and archaea across most biomes remain uncultured. *ISME Journal* **13**, 3126–3130 (2019).

[63] Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J. & Crosby, L. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems* **3**, e00055–18 (2018). URL http://msystems.asm.org/lookup/doi/10.1128/mSystems.00055-18.

[64] Wrighton, K. C. *et al.* RubisCO of a nucleoside pathway known from Archaea is found in diverse uncultivated phyla in bacteria. *The ISME journal* **10**, 2702–2714 (2016).

[65] Jaffe, A. L., Castelle, C. J., Dupont, C. L. & Banfield, J. F. Lateral Gene Transfer Shapes the Distribution of RuBisCO among Candidate Phyla Radiation Bacteria and DPANN Archaea. *Molecular biology and evolution* **36**, 435–446 (2019). URL http://www.ncbi.nlm.nih.gov/pubmed/30544151http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6389311.

[66] Lannes, R., Olsson-Francis, K., Lopez, P. & Bapteste, E. Carbon Fixation by Marine Ultrasmall Prokaryotes. *Genome biology and evolution* **11**, 1166–1177 (2019). URL http://www.ncbi.nlm.nih.gov/pubmed/30903144http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6475129.

[67] Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME Journal* **9**, 207–216 (2015).

[68] Ghylin, T. W. *et al.* Comparative single-cell genomics reveals potential ecological niches for the freshwater acI Actinobacteria lineage. *The ISME Journal* **8**, 2503–2516 (2014). URL http://www.nature.com/articles/ismej2014135.

[69] Kirk Harris, J. *et al.* Phylogenetic stratigraphy in the Guerrero Negro hypersaline microbial mat. *ISME Journal* **7**, 50–60 (2013).

[70] Hawley, A. K. *et al.* Diverse Marinimicrobia bacteria may mediate coupled biogeochemical cycles along eco-thermodynamic gradients. *Nature Communications* **8**, 1507 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/29142241http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5688066.

[71] Ju, F., Wang, Y. & Zhang, T. Bioreactor microbial ecosystems with differentiated methanogenic phenol biodegradation and competitive metabolic pathways unraveled with genome-resolved metagenomics. *Biotechnology for Biofuels* **11**, 135 (2018). URL https://biotechnologyforbiofuels.biomedcentral.com/articles/10.1186/s13068-018-1136-6.

[72] Castelle, C. J. *et al.* Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nature Reviews Microbiology* **16**, 629–645 (2018).

[73] Baker, B. J. *et al.* Enigmatic, ultrasmall, uncultivated Archaea. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 8806–8811 (2010).

[74] Comolli, L. R. & Banfield, J. F. Inter-species interconnections in acid mine drainage microbial communities. *Frontiers in Microbiology* **5** (2014).

[75] Golyshina, O. V. *et al.* 'ARMAN' archaea depend on association with euryarchaeal host in culture and in situ. *Nature Communications* **8**, 60 (2017). URL http://www.nature.com/articles/s41467-017-00104-7.

[76] Delafont, V., Samba-Louaka, A., Bouchon, D., Moulin, L. & Héchard, Y. Shedding light on microbial dark matter: a TM6 bacterium as natural endosymbiont of a free-living amoeba. *Environmental Microbiology Reports* **7**, 970–978 (2015). URL http://doi.wiley.com/10.1111/1758-2229.12343.

[77] He, X. *et al.* Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 244–9 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25535390http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4291631.

[78] Gong, J., Qing, Y., Guo, X. & Warren, A. "Candidatus Sonnebornia yantaiensis", a member of candidate division OD1, as intracellular bacteria of the ciliated protist Paramecium bursaria (Ciliophora, Oligohymenophorea). *Systematic and Applied Microbiology* **37**, 35–41 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24231291http://linkinghub.elsevier.com/retrieve/pii/S0723202013001574.

[79] Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nature reviews. Microbiology* **10**, 538–50 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22796884http://www.nature.com/articles/nrmicro2832.

[80] Friedman, J. & Alm, E. J. Inferring Correlation Networks from Genomic Survey Data. *PLoS Computational Biology* **8**, e1002687 (2012). URL https://dx.plos.org/10.1371/journal.pcbi.1002687.

[81] Kelder, T., Stroeve, J. H. M., Bijlsma, S., Radonjic, M. & Roeselers, G. Correlation network analysis reveals relationships between diet-induced changes in human gut

microbiota and metabolic health. *Nutrition & diabetes* **4**, e122 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24979151http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4079927.

[82] Mainali, K., Bewick, S., Vecchio-Pagan, B., Karig, D. & Fagan, W. F. Detecting interaction networks in the human microbiome with conditional Granger causality. *PLoS computational biology* **15**, e1007037 (2019).

[83] Faust, K., Lahti, L., Gonze, D., de Vos, W. M. & Raes, J. Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current Opinion in Microbiology* **25**, 56–66 (2015). URL https://www.sciencedirect.com/science/article/pii/S1369527415000478?via{%}3Dihub.

[84] Ma, B. *et al.* Genetic correlation network prediction of forest soil microbial functional organization. *The ISME Journal* **12**, 2492–2505 (2018). URL http://www.nature.com/articles/s41396-018-0232-8.

[85] Zhou, J., Deng, Y., Luo, F., He, Z. & Yang, Y. Phylogenetic molecular ecological network of soil microbial communities in response to elevated CO2. *mBio* **2** (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21791581http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3143843.

[86] Gilbert, J. A. *et al.* Defining seasonal marine microbial community dynamics. *The ISME Journal* **6**, 298–308 (2012). URL http://www.nature.com/articles/ismej2011107.

[87] Hurwitz, B. L., Westveld, A. H., Brum, J. R. & Sullivan, M. B. Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 10714–10719 (2014).

[88] Lima-Mendez, G. *et al.* Ocean plankton. Determinants of community structure in the global plankton interactome. *Science (New York, N.Y.)* **348**, 1262073 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25999517.

[89] Ghasemi, M., Seidkhani, H., Tamimi, F., Rahgozar, M. & Masoudi-Nejad, A. Centrality measures in Biological Networks. Tech. Rep. (2014). URL https://pdfs.semanticscholar.org/7696/8224a83df88454ad0b236ff3a502b02d24b2.pdf.

[90] Gosak, M. *et al.* Network science of biological systems at different scales: A review. *Physics of Life Reviews* **24**, 118–135 (2018). URL https://www.sciencedirect.com/science/article/pii/S1571064517301501?via{%}3Dihub{#}se0040.

[91] Proulx, S. R., Promislow, D. E. & Phillips, P. C. Network thinking in ecology and evolution (2005).

[92] Das, S., Meher, P. K., Rai, A., Bhar, L. M. & Mandal, B. N. Statistical Approaches for Gene Selection, Hub Gene Identification and Module Interaction in Gene Co-Expression Network Analysis: An Application to Aluminum Stress in Soybean (Glycine max L.).

PLOS ONE **12**, e0169605 (2017). URL http://dx.plos.org/10.1371/journal.pone.0169605.

[93] Widder, S. *et al.* Fluvial network organization imprints on microbial co-occurrence networks. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 12799–804 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/25136087http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4156742.

[94] Banerjee, S. *et al.* Network analysis reveals functional redundancy and keystone taxa amongst bacterial and fungal communities during organic matter decomposition in an arable soil. *Soil Biology and Biochemistry* **97**, 188–198 (2016). URL https://www.sciencedirect.com/science/article/pii/S0038071716300268.

[95] Steele, J. A. *et al.* Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *The ISME Journal* **5**, 1414–1425 (2011). URL http://www.nature.com/articles/ismej201124.

[96] Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nature Microbiology* **3**, 870–880 (2018). URL http://www.nature.com/articles/s41564-018-0190-y.

[97] Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/26863193http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4851848.

[98] Cram, J. A. *et al.* Cross-depth analysis of marine bacterial networks suggests downward propagation of temporal changes. *The ISME Journal* **9**, 2573–2586 (2015). URL http://www.nature.com/articles/ismej201576.

[99] Crits-Christoph, A., Diamond, S., Butterfield, C. N., Thomas, B. C. & Banfield, J. F. Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. *Nature* **558**, 440–444 (2018). URL http://www.nature.com/articles/s41586-018-0207-y.

[100] Duran-Pinedo, A. E., Paster, B., Teles, R. & Frias-Lopez, J. Correlation Network Analysis Applied to Complex Biofilm Communities. *PLoS ONE* **6**, e28438 (2011). URL http://dx.plos.org/10.1371/journal.pone.0028438.

[101] Bernard, G., Pathmanathan, J. S., Lannes, R., Lopez, P. & Bapteste, E. Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. *Genome Biology and Evolution* **10**, 707–715 (2018). URL https://academic.oup.com/gbe/article/10/3/707/4840377.

[102] Nasir, A., Kim, K. M. & Caetano-Anollés, G. Lokiarchaeota: Eukaryote-like missing links from microbial dark matter? *Trends in Microbiology* **23**, 448–50 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26112912.

[103] Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).

[104] Saw, J. H. *et al.* Exploring microbial dark matter to resolve the deep archaeal ancestry of eukaryotes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, 20140328 (2015). URL http://rstb.royalsocietypublishing.org/lookup/doi/10.1098/rstb.2014.0328.

[105] Yau, S. *et al.* Virophage control of antarctic algal host-virus dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 6163–8 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21444812http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3076838.

[106] Zhou, J. *et al.* Three novel virophage genomes discovered from Yellowstone Lake metagenomes. *Journal of virology* **89**, 1278–85 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25392206http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4300641.

[107] Oh, S., Yoo, D. & Liu, W.-T. Metagenomics Reveals a Novel Virophage Population in a Tibetan Mountain Lake. *Microbes and environments* **31**, 173–7 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27151658http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4912154.

[108] Wu, D. *et al.* Stalking the Fourth Domain in Metagenomic Data: Searching for, Discovering, and Interpreting Novel, Deep Branches in Marker Gene Phylogenetic Trees. *PLoS ONE* **6**, e18011 (2011). URL http://dx.plos.org/10.1371/journal.pone.0018011.

[109] Diener, C., Gibbons, S. M. & Resendis-Antonio, O. MICOM: Metagenome-Scale Modeling To Infer Metabolic Interactions in the Gut Microbiota. *mSystems* **5** (2020). URL http://msystems.asm.org/.

[110] Gupta, R. S., Bhandari, V. & Naushad, H. S. Molecular signatures for the PVC clade (Planctomycetes, Verrucomicrobia, Chlamydiae, and Lentisphaerae) of bacteria provide insights into their evolutionary relationships. *Frontiers in Microbiology* **3**, 327 (2012). URL http://journal.frontiersin.org/article/10.3389/fmicb.2012.00327/abstract.

[111] Vanwonterghem, I. *et al.* Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. *Nature microbiology* **1**, 16170 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27694807.

[112] Raoult, D., Scola, B. L. & Birtles, R. The Discovery and Characterization of Mimivirus, the Largest Known Virus and Putative Pneumonia Agent. *Clinical Infectious Diseases* **45**, 95–102 (2007). URL https://academic.oup.com/cid/article-lookup/doi/10.1086/518608.

[113] Lemos, L. N. *et al.* Genomic signatures and co-occurrence patterns of the ultra-small Saccharimonadia (phylum CPR/Patescibacteria) suggest a symbiotic lifestyle. *Molecular Ecology* **28**, 4259–4271 (2019). URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.15208.

[114] Luef, B. *et al.* Diverse uncultivated ultra-small bacterial cells in groundwater. *Nature Communications* **6**, 6372 (2015). URL http://www.nature.com/articles/ncomms7372.

[115] Boyd, J. A. *et al.* Divergent methyl-coenzyme M reductase genes in a deep-subseafloor Archaeoglobi. *The ISME Journal* 1 (2019). URL http://www.nature.com/articles/s41396-018-0343-2.

[116] Salcher, M. M., Schaefle, D., Kaspar, M., Neuenschwander, S. M. & Ghai, R. Evolution in action: habitat transition from sediment to the pelagial leads to genome streamlining in Methylophilaceae. *The ISME Journal* 1 (2019). URL http://www.nature.com/articles/s41396-019-0471-3.

[117] Eisen, J. A. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Current Opinion in Genetics & Development* **10**, 606–611 (2000). URL https://www.sciencedirect.com/science/article/pii/S0959437X0000143X.

[118] Avcı, B., Krüger, K., Fuchs, B. M., Teeling, H. & Amann, R. I. Polysaccharide niche partitioning of distinct Polaribacter clades during North Sea spring algal blooms. *The ISME Journal* 1–15 (2020). URL http://www.nature.com/articles/s41396-020-0601-y.

[119] Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data* **5**, 170203 (2018). URL http://www.nature.com/articles/sdata2017203.

[120] Sberro, H. *et al.* Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. *Cell* (2019). URL https://linkinghub.elsevier.com/retrieve/pii/S0092867419307810.

[121] Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nature Communications* **10** (2019).

[122] Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **0** (2019). URL https://linkinghub.elsevier.com/retrieve/pii/S0092867419300017.

[123] West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C. & Banfield, J. F. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome research* **28**, 569–580 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29496730http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5880246.

[124] Wrighton, K. C. *et al.* Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. *The ISME Journal* **8**, 1452–1463 (2014). URL http://www.nature.com/articles/ismej2013249.

[125] Tan, S. *et al.* Insights into ecological role of a new deltaproteobacterial order Candidatus Acidulodesulfobacterales by metagenomics and metatranscriptomics. *ISME Journal* (2019).

[126] Dong, X. *et al.* Metabolic potential of uncultured bacteria and archaea associated with petroleum seepage in deep-sea sediments. *Nature Communications* **10**, 1816 (2019). URL http://www.nature.com/articles/s41467-019-09747-0.

[127] Zelezniak, A. *et al.* Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 6449–54 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/25941371http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4443341.

[128] Blattman, S. B., Jiang, W., Oikonomou, P. & Tavazoie, S. Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing. *Nature Microbiology* **5**, 1192–1201 (2020). URL https://doi.org/10.1038/s41564-020-0729-6.

[129] Karst, S. M. *et al.* Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. *Nature Biotechnology* **36**, 190–195 (2018).

[130] Röttjers, L. & Faust, K. From hairballs to hypotheses–biological insights from microbial networks. *FEMS Microbiology Reviews* **42**, 761–780 (2018). URL https://academic.oup.com/femsre/article/42/6/761/5061627.

[131] Ghurye, J. S., Cepeda-Espinoza, V. & Pop, M. Metagenomic Assembly: Overview, Challenges and Applications. *The Yale journal of biology and medicine* **89**, 353–362 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27698619http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5045144.

[132] Escobar-Zepeda, A. *et al.* Analysis of sequencing strategies and tools for taxonomic annotation: Defining standards for progressive metagenomics. *Scientific Reports* **8**, 12034 (2018). URL http://www.nature.com/articles/s41598-018-30515-5.

[133] Gu, C., Kim, G. B., Kim, W. J., Kim, H. U. & Lee, S. Y. Current status and applications of genome-scale metabolic models (2019).

[134] Chan, S. H. J., Simons, M. N. & Maranas, C. D. SteadyCom: Predicting microbial abundances while ensuring community stability. *PLOS Computational Biology* **13**, e1005539 (2017). URL http://dx.plos.org/10.1371/journal.pcbi.1005539.

[135] Solden, L., Lloyd, K. & Wrighton, K. The bright side of microbial dark matter: Lessons learned from the uncultivated majority. *Current Opinion in Microbiology* **31**, 217–226 (2016). URL https://www.sciencedirect.com/science/article/pii/S1369527416300558?via{%}3Dihub{#}bib0335.

[136] Catchpole, R. & Forterre, P. Positively twisted: The complex evolutionary history of Reverse Gyrase suggests a non-hyperthermophilic Last Universal Common Ancestor. *bioRxiv* 524215 (2019). URL https://www.biorxiv.org/content/10.1101/524215v1.abstract.

[137] Chen, I.-M. A. *et al.* IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic acids research* **47**, D666–D677 (2019). URL http://www.ncbi.nlm.nih.gov/pubmed/30289528http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6323987.

[138] Méheust, R., Burstein, D., Castelle, C. J. & Banfield, J. F. The distinction of CPR bacteria from other bacteria based on protein family content. *Nature Communications* **10**, 1–12 (2019). URL https://doi.org/10.1038/s41467-019-12171-z.

[139] Galperin, M. Y. & Koonin, E. V. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Research* **32**, 5452 (2004). URL http://www.ncbi.nlm.nih.gov/pubmed/15479782http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC524295.

[140] Antczak, M., Michaelis, M. & Wass, M. Investigating the unknown functions in the minimal bacterial genome reveals many transporter proteins. *bioRxiv* 381657 (2018). URL https://www.biorxiv.org/content/10.1101/381657v1.

[141] Thomas, A. M. & Segata, N. Multiple levels of the unknown in microbiome research (2019).

[142] Huang, S. Back to the biology in systems biology: what can we learn from biomolecular networks? *Briefings in functional genomics & proteomics* **2**, 279–97 (2004). URL http://www.ncbi.nlm.nih.gov/pubmed/15163364.

[143] Ma'ayan, A. Introduction to network analysis in systems biology. *Science signaling* **4**, tr5 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21917719http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3196357.

[144] Bauer, M. A., Kainz, K., Carmona-Gutierrez, D. & Madeo, F. Microbial wars: competition in ecological niches and within the microbiome. *Microbial Cell* **5**, 215–219 (2018). URL http://microbialcell.com/researcharticles/microbial-wars-competition-in-ecological-niches-and-within-the-microbiome/.

[145] Kurz, F. T. *et al.* Network dynamics: quantitative analysis of complex behavior in metabolism, organelles, and cells, from experiments to models and back. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **9**, e1352 (2017). URL http://doi.wiley.com/10.1002/wsbm.1352.

[146] Barberán, A., Bates, S. T., Casamayor, E. O. & Fierer, N. Using network analysis to explore co-occurrence patterns in soil microbial communities. *The ISME Journal* **6**, 343–351 (2012). URL http://www.nature.com/articles/ismej2011119.

[147] Xiao, J. *et al.* Predictive Modeling of Microbiome Data Using a Phylogeny-Regularized Generalized Linear Mixed Model. *Frontiers in microbiology* **9**, 1391 (2018). URL http://www.ncbi.nlm.nih.gov/pubmed/29997602http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6030386.

[148] Wuchty, S., Ravasz, E. & Barabasi, A.-L. The Architecture of Biological Networks. *Complex Systems in Biomedicine* (2003).

[149] Blüthgen, N., Fründ, J., Vázquez, D. P. & Menzel, F. What do interaction network metrics tell us about specialization and biological traits? *Ecology* **89**, 3387–99 (2008). URL http://www.ncbi.nlm.nih.gov/pubmed/19137945.

[150] Toju, H., Tanabe, A. S. & Sato, H. Network hubs in root-associated fungal metacommunities. *Microbiome* **6**, 1–16 (2018). URL https://doi.org/10.1186/s40168-018-0497-1.

[151] Shi, Y. *et al.* Abundance of kinless hubs within soil microbial networks are associated with high functional potential in agricultural ecosystems. *Environment International* **142**, 105869 (2020).

[152] He, X. & Zhang, J. Why Do Hubs Tend to Be Essential in Protein Networks? *PLoS Genetics* **2**, e88 (2006). URL https://dx.plos.org/10.1371/journal.pgen.0020088.

[153] Zhang, D.-Q., Zhou, C.-K., Chen, S.-Z., Yang, Y. & Shi, B.-K. Identification of hub genes and pathways associated with bladder cancer based on co-expression network analysis. *Oncology letters* **14**, 1115–1122 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28693282http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5494668.

[154] Yang, W. *et al.* Identification of hub genes and therapeutic drugs in esophageal squamous cell carcinoma based on integrated bioinformatics strategy. *Cancer Cell International* **19**, 142 (2019). URL https://cancerci.biomedcentral.com/articles/10.1186/s12935-019-0854-6.

[155] Cai, Y. *et al.* Identification of five hub genes as monitoring biomarkers for breast cancer metastasis in silico. *Hereditas* **156**, 20 (2019). URL https://hereditasjournal.biomedcentral.com/articles/10.1186/s41065-019-0096-6.

[156] Liu, N. *et al.* Functional metagenomics reveals abundant polysaccharide-degrading gene clusters and cellobiose utilization pathways within gut microbiota of a wood-feeding higher termite. *The ISME Journal* **13**, 104–117 (2019). URL http://www.nature.com/articles/s41396-018-0255-1.

[157] Brooks, B. *et al.* Strain-resolved microbial community proteomics reveals simultaneous aerobic and anaerobic function during gastrointestinal tract colonization of a preterm infant. *Frontiers in microbiology* **6**, 654 (2015).

[158] Fouhy, F., Clooney, A. G., Stanton, C., Claesson, M. J. & Cotter, P. D. 16S rRNA gene sequencing of mock microbial populations-impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiology* **16**, 123 (2016).

[159] Tremblay, J. *et al.* Primer and platform effects on 16S rRNA tag sequencing. *Frontiers in Microbiology* **6**, 771 (2015).

[160] Vollmers, J., Wiegand, S. & Kaster, A.-K. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! *PloS one* **12**, e0169662 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28099457http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5242441.1604.03071.

[161] Plummer, E. & Twin, J. A Comparison of Three Bioinformatics Pipelines for the Analysis of Preterm Gut Microbiota using 16S rRNA Gene Sequencing Data. *Journal of Proteomics & Bioinformatics* **8**, 283–291 (2015).

[162] Nilakanta, H., Drews, K. L., Firrell, S., Foulkes, M. A. & Jablonski, K. A. A review of software for analyzing molecular Sequences (2014).

[163] Schloss, P. D. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Computational Biology* **6**, 19 (2010).

[164] Edgar, R. C. Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. *PeerJ* **5**, e3889 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/29018622http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5631090.

[165] Edgar, R. C. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34**, 2371–2375 (2018).

[166] Claesson, M. J., Clooney, A. G. & O'Toole, P. W. A clinician's guide to microbiome analysis (2017).

[167] Knight, R. *et al.* Best practices for analysing microbiomes (2018).

[168] Yang, B., Wang, Y. & Qian, P. Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics* **17**, 135 (2016).

[169] Kurtz, Z. D. *et al.* Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology* **11**, e1004226 (2015). URL http://dx.plos.org/10.1371/journal.pcbi.1004226.

[170] Almaas, E. Biological impacts and context of network theory (2007). URL http://www.thebiogrid.org/.

[171] Joy, M. P., Brock, A., Ingber, D. E. & Huang, S. High-betweenness proteins in the yeast protein interaction network. *Journal of biomedicine & biotechnology* **2005**, 96–103 (2005). URL http://www.ncbi.nlm.nih.gov/pubmed/16046814http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1184047.

[172] Ashtiani, M. *et al.* A systematic survey of centrality measures for protein-protein interaction networks. *BMC Systems Biology* **12**, 80 (2018). URL https://bmcsystbiol.biomedcentral.com/articles/10.1186/s12918-018-0598-2.

[173] Kranthi, T., Rao, S. B. & Manimaran, P. Identification of synthetic lethal pairs in biological systems through network information centrality. *Molecular BioSystems* **9**, 2163–2167 (2013). URL https://pubs.rsc.org/en/content/articlehtml/2013/mb/c3mb25589ahttps://pubs.rsc.org/en/content/articlelanding/2013/mb/c3mb25589a.

[174] Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001). URL http://www.nature.com/articles/35075138.

[175] Fang, H., Huang, C., Zhao, H. & Deng, M. CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* **31**, 3172–3180 (2015). URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv349.

[176] Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336 (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20383131http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3156573http://www.nature.com/articles/nmeth.f.303.

[177] McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8**, e61217 (2013). URL https://dx.plos.org/10.1371/journal.pone.0061217.

[178] Holm, S. Board of the Foundation of the Scandinavian Journal of Statistics A Simple Sequentially Rejective Multiple Test Procedure A Simple Sequentially Rejective Multiple Test Procedure. Tech. Rep. 2 (1979).

[179] Marcy, Y. *et al.* 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 of the United States of America* **104**, 11889–94 (2007). URL http://www.ncbi.nlm.nih.gov/pubmed/17620602http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1924555.

[180] Schulz, F. *et al.* Towards a balanced view of the bacterial tree of life. *Microbiome* **5**, 140 (2017). URL http://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0360-9. 1708.02002.

[181] Marx, V. The big challenges of big data. *Nature* **498**, 255–260 (2013). URL https://www.nature.com/articles/498255a.

[182] Stephens, Z. D. *et al.* Big Data: Astronomical or Genomical? *PLOS Biology* **13**, e1002195 (2015). URL https://dx.plos.org/10.1371/journal.pbio.1002195.

[183] Shade, A. *et al.* Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *mBio* **5**, e01371–14 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/25028427http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4161262.

[184] Williams, R. J., Howe, A. & Hofmockel, K. S. Demonstrating microbial co-occurrence pattern analyses within and between ecosystems. *Frontiers in Microbiology* **5**, 358 (2014). URL http://journal.frontiersin.org/article/10.3389/fmicb.2014.00358/abstract.

[185] Gehlenborg, N. *et al.* Visualization of omics data for systems biology (2010). URL https://pubmed.ncbi.nlm.nih.gov/20195258/.

[186] Aylward, F. O. *et al.* Supplementary Appendix for " Microbial Community Transcriptional Networks are Conserved in Three Domains at Ocean Basin Scales " 3 Supplementary Dataset Legends (S1-S5) URL http://www.pnas.org/content/pnas/suppl/2015/03/06/1502883112.DCSupplemental/pnas.1502883112.sapp.pdf.

[187] Coutinho, F. H. *et al.* Niche distribution and influence of environmental parameters in marine microbial communities: a systematic review. *PeerJ* **3**, e1008 (2015). URL https://peerj.com/articles/1008.

[188] Dai, Y., Jiang, J.-B., Wang, Y.-L., Jin, Z.-T. & Hu, S.-Y. Functional and protein-protein interaction network analysis of colorectal cancer induced by ulcerative colitis. *Molecular medicine reports* **12**, 4947–58 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26239378http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4581825.

[189] Cardona, C., Weisenhorn, P., Henry, C. & Gilbert, J. A. Network-based metabolic analysis and microbial community modeling (2016).

[190] Beck, C., Knoop, H. & Steuer, R. Modules of co-occurrence in the cyanobacterial pan-genome reveal functional associations between groups of ortholog genes. *PLOS Genetics* **14**, e1007239 (2018). URL https://dx.plos.org/10.1371/journal.pgen.1007239.

[191] Corel, E. *et al.* Bipartite Network Analysis of Gene Sharings in the Microbial World. *Molecular Biology and Evolution* **35**, 899–913 (2018). URL https://academic.oup.com/mbe/article/35/4/899/4810447.

[192] Das, P., Ji, B., Kovatcheva-Datchary, P., Bäckhed, F. & Nielsen, J. In vitro co-cultures of human gut bacterial species as predicted from co-occurrence network analysis. *PLOS ONE* **13**, e0195161 (2018). URL https://dx.plos.org/10.1371/journal.pone.0195161.

[193] Li, J. *et al.* Application of Weighted Gene Co-expression Network Analysis for Data from Paired Design. *Scientific Reports* **8**, 622 (2018). URL http://www.nature.com/articles/s41598-017-18705-z.

[194] Pruesse, E. *et al.* SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35**, 7188–7196 (2007).

[195] Tipton, L. *et al.* Fungi stabilize connectivity in the lung and skin microbial ecosystems. *Microbiome* **6**, 12 (2018). URL https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0393-0.

[196] Liu, H., Roeder, K. & Wasserman, L. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. Tech. Rep. (2010). 1006.3316v1.

[197] Ju, F., Xia, Y., Guo, F., Wang, Z. & Zhang, T. Taxonomic relatedness shapes bacterial assembly in activated sludge of globally distributed wastewater treatment plants. *Environmental Microbiology* **16**, 2421–2432 (2014). URL http://doi.wiley.com/10.1111/1462-2920.12355.

[198] Cao, X. *et al.* Heterogeneity of interactions of microbial communities in regions of Taihu Lake with different nutrient loadings: A network analysis. *Scientific Reports* **8**, 8890 (2018). URL http://www.nature.com/articles/s41598-018-27172-z.

[199] Zhou, Z. *et al.* Genome- and Community-Level Interaction Insights into Carbon Utilization and Element Cycling Functions of Hydrothermarchaeota in Hydrothermal Sediment. *mSystems* **5** (2020). URL http://www.ncbi.nlm.nih.gov/pubmed/31911466http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6946796.

[200] Cardenas, J. P., Quatrini, R. & Holmes, D. S. Aerobic lineage of the oxidative stress response protein rubrerythrin emerged in an ancient microaerobic, (hyper)thermophilic environment. *Frontiers in Microbiology* **7**, 1822 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27917155http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5114695.

[201] Rappé, M. S. & Giovannoni, S. J. The Uncultured Microbial Majority (2003). URL https://pubmed.ncbi.nlm.nih.gov/14527284/.

[202] Becraft, E. D. *et al.* Single-cell-genomics-facilitated read binning of candidate phylum EM19 genomes from geothermal spring metagenomes. *Applied and Environmental Microbiology* **82**, 992–1003 (2016). URL /pmc/articles/PMC4751853/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4751853/.

[203] Bruno, A. *et al.* Exploring the under-investigated "microbial dark matter" of drinking water treatment plants. *Scientific Reports* **7**, 44350 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28290543http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5349567.

[204] Lv, X. *et al.* Strengthening Insights in Microbial Ecological Networks from Theory to Applications. *mSystems* **4**, e00124–19 (2019). URL http://www.ncbi.nlm.nih.gov/pubmed/31117020http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6529547.

[205] Meier, D. V. *et al.* Niche partitioning of diverse sulfur-oxidizing bacteria at hydrothermal vents. *The ISME journal* **11**, 1545–1558 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28375213http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5520155.

[206] Han, Y. *et al.* Hydrothermal chimneys host habitat-specific microbial communities: analogues for studying the possible impact of mining seafloor massive sulfide deposits. *Scientific Reports* **8**, 10386 (2018). URL http://www.nature.com/articles/s41598-018-28613-5.

[207] Langille, M. G. I. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* **31**, 814–821 (2013). URL http://www.nature.com/articles/nbt.2676.

[208] Casaburi, G., Goncharenko-Foster, I., Duscher, A. A. & Foster, J. S. Transcriptomic changes in an animal-bacterial symbiosis under modeled microgravity conditions. *Scientific Reports* **7** (2017). URL https://pubmed.ncbi.nlm.nih.gov/28393904/.

[209] Marx, V. Engineers embrace microbiome messiness. *Nature Methods* **16**, 581–584 (2019). URL http://www.nature.com/articles/s41592-019-0460-5.

[210] Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Research* **26**, 1612–1625 (2016).

[211] Overbeek, R. *et al.* The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic acids research* **42**, D206–14 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24293654http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3965101.

[212] Harrison KJ, Crécy-Lagard V, Z. R. Gene Graphics: a genomic neighborhood data visualization (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29228171.

[213] Danczak, R. E. *et al.* Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. *Microbiome* **5**, 112 (2017). URL http://www.ncbi.nlm.nih.gov/pubmed/28865481http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5581439.

[214] Momper, L., Aronson, H. S. & Amend, J. P. Genomic Description of 'Candidatus Abyssubacteria,' a Novel Subsurface Lineage Within the Candidate Phylum Hydrogenedentes. *Frontiers in Microbiology* **9**, 1993 (2018). URL https://www.frontiersin.org/article/10.3389/fmicb.2018.01993/full.

[215] Nagpal, S., Haque, M. M. & Mande, S. S. Vikodak - A Modular Framework for Inferring Functional Potential of Microbial Communities from 16S Metagenomic Datasets. *PLOS ONE* **11**, e0148347 (2016). URL https://dx.plos.org/10.1371/journal.pone.0148347.

[216] Ling, L. L. *et al.* A new antibiotic kills pathogens without detectable resistance. *Nature* **517**, 455–459 (2015). URL http://www.nature.com/articles/nature14098.

[217] Hyatt, D., Locascio, P. F., Hauser, L. J. & Uberbacher, E. C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**, 2223–2230 (2012).

[218] Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution* **34**, 2115–2122 (2017). URL https://pubmed.ncbi.nlm.nih.gov/28460117/.

[219] Conesa, A. *et al.* Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).

[220] Kamke, J. *et al.* The candidate phylum Poribacteria by single-cell genomics: new insights into phylogeny, cell-compartmentation, eukaryote-like repeat proteins, and other genomic features. *PloS one* **9**, e87353 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24498082http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3909097.

[221] Chen, L. X., Anantharaman, K., Shaiber, A., Murat Eren, A. & Banfield, J. F. Accurate and complete genomes from metagenomes (2020). URL http://www.genome.org/cgi/doi/10.1101/gr.258640.119.

[222] Xu, Y. & Zhao, F. Single-cell metagenomics: challenges and applications (2018). URL /pmc/articles/PMC5960468/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5960468/.

[223] Limam, R. D. *et al.* Members of the uncultured bacterial candidate division WWE1 are implicated in anaerobic digestion of cellulose. *MicrobiologyOpen* **3**, 157–167 (2014). URL http://doi.wiley.com/10.1002/mbo3.144.

[224] Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* **9**, 811–814 (2012). URL http://www.nature.com/articles/nmeth.2066.

[225] Banfield, J. F., Anantharaman, K., Williams, K. H. & Thomas, B. C. Complete 4.55-Megabase-Pair Genome of "Candidatus Fluviicola riflensis," Curated from Short-Read Metagenomic Sequences. *Genome announcements* **5** (2017).

[226] Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Towards functional characterization of archaeal genomic dark matter. *Biochemical Society Transactions* **47**, 389–398 (2019). URL http://www.biochemsoctrans.org/content/47/1/389.abstract.

[227] Blaser, M. J. *et al.* Toward a Predictive Understanding of Earth's Microbiomes to Address 21st Century Challenges. *mBio* **7**, e00714–16 (2016). URL http://www.ncbi.nlm.nih.gov/pubmed/27178263http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4895116.

BIOGRAPHICAL SKETCH

Tatyana Zamkovaya was born in Dnipro, Ukraine and moved to Florida when she was 5 years old. From an early age, Tatyana was interested in biology and had a passion for animals and medicine. An interdisciplinary high school research project on the biology, chemistry, and physics behind extremophile survival in hydrothermal vents incited a lifelong interest in microbiology for Tatyana. After finishing high school, she pursued a double bachelor's degree in microbiology and cell science and French and francophone studies at the University of Florida. The diverse coursework and mentorship she received in the Microbiology and Cell Science Department led her to pursue a PhD at her alma mater. Although she had taken some bioinformatics courses in her undergraduate studies, the R for genomics course and subsequent rotation in the Conesa lab on the investigation of type 1 diabetes responses in patients in her first year of her doctorate studies together inspired a new passion for computational biology and bioinformatics in Tatyana. Under the advisorship of Dr. Ana Conesa, she investigated the ecological and functional role of Microbial Dark Matter within extremophilic conditions, gaining valuable experience in analyzing omics data, various programming languages, and teaching during her studies. Her current research interests lie in metagenomics and other omics data analysis, Big Data analysis, network biology and their integration as a means to better understand ecology, gene function and diversity, disease, and evolution.