THE UNIVERSITY OF CHICAGO

GENE FAMILY PHYLOGEOGRAPHY EXPANDS INSIGHTS INTO MICROBIAL ECOLOGY BEYOND GENOME COLLECTIONS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON MICROBIOLOGY

BY
MATTHEW SPENCER SCHECHTER

CHICAGO, ILLINOIS
MARCH 2025

This thesis is dedicated to my family, whose relentless love and support are the reason I am here today. Thank you for everything, my love for you guys knows no bounds.

ABSTRACT

Advances in metagenomics and genome recovery from diverse habitats have dramatically expanded the availability of microbial genomes, unlocking unprecedented insights into microbial ecology and evolution. These discoveries are paving the way for transformative biotechnological and biomedical innovations. Despite this progress, final genome collections often provide an incomplete picture of microbial diversity, limiting our understanding of genome variation and excluding certain gene family orthologs. In contrast, metagenomic assemblies offer a more comprehensive view of microbial genomic diversity, capturing gene families lost during genome recovery and revealing broader ecological and variation of orthologs. This thesis introduces the EcoPhylo workflow, an open-source computational framework designed to track the phylogeography of gene families across environments. By leveraging single-copy core genes, such as ribosomal protein gene families, EcoPhylo enables benchmarking of genome recovery rates across environments, taxa, and recovery methods, providing a powerful tool for evaluating microbial genome datasets. Using this workflow, I benchmarked genome recovery rates in the human oral cavity and gut, uncovering variation across taxa and biomes, and offering valuable guidance for optimizing genome recovery strategies in future studies. The scalability of EcoPhylo is further demonstrated through its application to a biome-level genome collection from the global surface ocean, highlighting its ability to analyze large-scale datasets. Additionally, I applied EcoPhylo to investigate the evolutionary landscape of anaerobic flavin respiratory reductases across hundreds of gut microbes, illustrating its broader utility for exploring gene family phylogenetics and functional diversity. Overall, the EcoPhylo workflow provides a platform for linking microbial genomes and gene families across environmental metagenomics datasets and offers a new avenue for benchmarking genome recovery efforts and advancing our understanding of microbial ecology.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am beyond grateful for all the people who have supported me on my journey to becoming a scientist. I feel incredibly fortunate to have been surrounded by such intelligent, kind, and supportive individuals who have guided and inspired me along the way.

My thesis committee, thank you for your support through this Ph.D. journey. Your positivity, encouragement, and constructive feedback kept on track. Thank you!

Sam Light, thank you for accepting me in your lab during the Meren lab transition to Germany. You could have asked me to completely change projects, to start over. Instead, you creatively found a way to integrate my on-going work with your lab's research goals. Our collaboration on Little et al. 2024 exploring reductase active sites architecture was one of my favorite projects of my Ph.D.. I truly believe your lab is conducting some of the most exciting and groundbreaking work in the human gut microbiome. I wish you and everyone in the lab all the best as you continue pushing the boundaries of this fascinating field.

Meren, I am so lucky to have joined your lab. Your creativity, intellectual rigor, and ethics have changed the trajectory of my life. Whatever my next step will be, your supervision acts as a bright guiding light for my future endeavors. Thank you for challenging me to reach intellectual and professional heights I never thought possible, and for genuinely caring about my growth as both a scientist and a person. I hope we stay connected and can't wait to follow your future discoveries in biology.

Kenneth Nealson Ph.D., **Radu Popa Ph.D.**, **Antonio Fernández-Guerra Ph.D.**, thank you for giving me an opportunity, for showing me what science is all about, and for your encouragement to pursue a Ph.D.

CHAPTER 1

INTRODUCTION

1.1 Microbes Everywhere All at Once

Microbial life thrives across this planet. In fact, after a five-minute call with Dani Rojas from Ted Lasso, I can easily imagine that he would walk away shouting, "Microbiology is life!"

Microbial life is everywhere, and if life as we know it originated on this planet - it was likely microbial (Woolfson, 2015). Microbes flourish across the global oceans (Venter et al., 2004; Sunagawa et al., 2015), from the poles (DeLong et al., 1994) to hydrothermal vents, (Schrenk et al., 2004; Brazelton and Baross, 2009) and even the bottom of the ocean in sediment entombed for millions of years (Garber et al., 2024). They thrive in soils all over the world (Fierer and Jackson, 2006), in permafrost (Hultman et al., 2015), and in extreme environments such as hypersaline habitats (Fernández et al., 2014). Microbes also bloom on and inside plants (Berendsen et al., 2012) and animals (Simon et al., 2019; Qu et al., 2008) - yes, even jellyfish also have microbiomes (Tinta et al., 2019). Unsurprisingly, *Homo sapiens* (Gilbert et al., 2018) and other primates (Grieneisen et al., 2020, 2021) also host rich microbial communities.

Microbial life has contributed to global biogeochemistry since the origin of life on this planet ~3.5 billion years ago. The evolution of phototrophs radically changed global biogeochemistry by oxygenating the Earth's atmosphere. Marine microbes impact global carbon cycling by fixing carbon dioxide and settling on the bottom of the ocean, sequestrating carbon 1,300 Pg of Carbon per year (Nowicki et al., 2022) through the biological carbon pump (Siegel et al., 2023). Nitrogen-fixing microbes (diazotrophs) perform the incredible catalysis of breaking three covalent bonds in nitrogen gas, making bioavailable ammonia for the global oceans and soils (̃300 Tg nitrogen per year) (Hutchins and Capone, 2022; Gruber and Galloway, 2008). Microbes also play essential roles in most key elemental cycles that impact life, such as phosphorus (Duhamel et al., 2021), sulfur (Zhou et al., 2023), iron (Tagliabue et al., 2017), and man-

ganese (Aguilar and Nealson, 1994). All of these microbial powered, global biogeochemical processes are what make this planet habitable for us humans.

Plants and animals have co-evolved with microbes, developing symbioses essential for mutual ecological success, as well as parasitic relationships. The plant microbiome plays a variety of roles for its host, including the production of enzymes to break down nutrients in the surrounding soil (Backer et al., 2018), defense against plant pathogens (Pieterse et al., 2014; Trivedi et al., 2016), and critically creating bio-available nitrogen in the rhizosphere (Lynch et al., 2001). A classic example of an animal microbe symbiosis is the bioluminescent symbiont *Vibrio fischeri*, which helps camouflage the nocturnal Hawaiian bobtail squid (McFall-Ngai and Ruby, 1991). This relationship, which involves the squid hosting the bacterial partner in a dedicated light organ (Septer and Visick, 2024), has advanced our understanding of the evolution and biochemistry of host-microbe interactions and inspired microbiologists to discover quorum sensing (Nealson et al., 1970). Another example of an intimate relationship between microbes and the host is *Wolbachia*, a genus of obligate intracellular bacteria that resides in insect ovaries (Reveillaud et al., 2019). This symbiosis is being harnessed as a mosquito-control strategy to prevent the spreading of tropical diseases such as West Nile, Zika, dengue, and yellow fever (Shragai et al., 2017; Flores and O'Neill, 2018). Furthermore, ruminant animals have co-evolved with their gut microbiome, which breaks down complex carbohydrates such as cellulose to produce nutrients (Mizrahi et al., 2021). The impact of a microbiome on hosts varies, as some evidence suggests that "microbiome-free" hosts can ecologically succeed without a microbiome. In other cases, a host microbiome can be temporary or even parasitic (Hammer et al., 2019). Given the vast examples of host-microbe interactions, the term "holobiont" has been used to describe these relationships where evolution acts on both the host and associated microbiome as one entity. Interestingly, the term holobiont stems from the theoretical discussion of endosymbiosis (O'Malley, 2017), expanding the gene repertoire of the host cell. Overall, these intricate partnerships highlight the dynamic interplay between

hosts and their microbiomes, shaping the evolutionary trajectories of the host, microbes, and holobiont.

In Chapter 3 of this thesis, I leveraged a computational workflow I developed during my Ph.D., the anvi'o EcoPhylo workflow, to study the genome recovery rates and microbial ecology of three microbiomes: the human oral cavity, the human gut, and the surface ocean. Each environment provides unique evolutionary and ecological pressures on colonizing microbiomes.

The first environment I decided to explore with the EcoPhylo workflow was the human oral microbiome. Excitingly, plaque was one of the first microbial samples that Antonie van Leeuwenhoek's visualized underneath his single lens light microscope. The human oral cavity is a diverse environment defined by a range of colonization pressures. It has varied attachment substrates, including hard teeth and a soft tongue (Mark Welch et al., 2019). Additionally, the habitat contains multiple environmental gradients including oxygen, nutrients, and saliva (Simón-Soro et al., 2013). These ecological factors select for diverse microbial communities that form biofilms with distinct spatial structures (Mark Welch et al., 2016). Interestingly, the teeth have been hypothesized to be a transition zone for environmental microbes to adapt to a host-associated lifestyle (Shaiber et al., 2020). Expanding genome collections from this microbiome has direct implications for understanding oral health, and its accessibility for sampling makes it an intriguing habitat to study microbial ecology.

Among the broad anecdotes of microbes impacting the globe mentioned above, it can be argued that the human gut microbiome has been the most extensively studied from a scientific funding perspective. Research has shown that the human gut microbiome is vertically transmitted from mother to infant during birth (Ferretti et al., 2018). The infant microbiome transitions into an adult-like state over the first few years of life by incorporating microbes from the environment and increasing species richness (Bäckhed et al., 2015). Interestingly, this transfer includes not only microbes but also mobile genetic elements, such as *Bacteroides*

plasmids (Fogarty et al., 2024). The correlations between the gut microbiome and health or disease have further driven research efforts (Almeida et al., 2019; Parks et al., 2017; Pasolli et al., 2019). In fact, numerous microbiome surveys have sought to define a "healthy" gut microbiome to better detect disease states and develop medical interventions (Fan and Pedersen, 2020). Dysbiosis of the gut microbiome, characterized by reduced alpha diversity, has been linked to conditions such as Crohn's disease and Ulcerative Colitis (Kostic et al., 2014; Nagalingam and Lynch, 2012). Notably, studies have revealed differences between Western and non-Western gut microbiomes, underscoring the need to extract more human gut microbial genomes from underrepresented populations to explore the impact of industrialization on gut microbiomes (Pasolli et al., 2019).

The marine ecosystem hosts a remarkably diverse microbiome, spanning from the ocean floor to the sunlit surface. Circulating through this vast environment, the global ocean conveyor belt - a system of currents driven by temperature and salinity - takes approximately 1,000 years to move a single parcel of seawater through its entirety. In just 1 mL of seawater, there can be around 1 million plankton (Porter and Feig, 1980), and the abundance of phages infecting these microbes can reach up to $\sim 10^7$ (Suttle, 2005). Growing up near the ocean as a surfer and lifeguard, I swallowed gallons of seawater, along with many marine bacteria and viruses.

In Chapter 3 of this thesis, I focused on analyzing microbiome samples from the sunlit ocean, specifically analyzing samples from depths shallower than 30 meters. This biome is one of the most sampled in the global ocean due to its accessibility and role as the starting point for multiple biogeochemical processes (Falkowski et al., 2008; Falkowski, 2012). Sunlight energy permeates these shallow depths powering primary producers that fix CO_2 into organic matter through photosynthesis. Additionally, the input of Aeolian dust from the surrounding continents provides essential nutrients (e.g. iron and phosphorus) for Carbon fixation as well as Nitrogen fixation. These are driving factors in the sunlit ocean that shape marine microbial ecology, limiting where primary productivity can bloom as well as diazatroph activity (Tyrrell,

1999). Outside of nutrient limitations, temperature has been shown to correlate with microbial diversity in the global oceans (Sunagawa et al., 2015). The surface ocean hosts a vast array of biomes, but one Order of microbes, *Pelagibacterales* (SAR11), tends to dominate this habitat, ranging from 20 to 40% cells in seawater samples (Schattenhofer et al., 2009). In fact, SAR11 1a.3.V has been shown to recruit 1.5% of metagenomic reads from metagenomes representing an enormous proportion of DNA extracted from a sample (Delmont et al., 2019). Despite its ecological importance as a globally prevalent heterotroph that metabolizes low-molecular-weight dissolved organic matter across the global ocean, relatively few complete genome sequences are available. Challenging culture conditions for microbial isolation (Giovannoni, 2017) and fragmented assemblies from marine metagenomes (Chen et al., 2020a) limit genome recovery of this taxon. Ongoing efforts aim to address this gap and expand genomic resources for this taxon (Freel et al., 2024). Measuring this discrepancy between the global prevalence and lack of genomic representation of SAR11 is discussed further in Chapter 3.

A fundamental aspect of studying microbial communities is observing them in their natural environments, which can be approached through culture-dependent and culture-independent methods. Culture-dependent strategies involve isolating individual microbes or consortia from the environment to study them in the laboratory. While this approach enables direct experimentation with living organisms, it is slow and constrained by the need to develop specific growth conditions and isolation techniques, making it impractical to culture all microbes on Earth. In contrast, culture-independent methods detect microbes directly in their environments through DNA-based surveys, bypassing the limitations of culturing techniques. Among these, metagenomics has been transformative, enabling the sequencing of all DNA present in an environmental sample. This approach has generated the majority of the genomic data analyzed in this thesis and forms the foundation of my perspective on microbiology and microbial ecology. I will therefore introduce the history metagenomics in detail in the next section.

1.2 A brief history of metagenomics and its impact on the field of microbiology

This section is derived from the following blog post:

Matthew S. Schechter. The history of metagenomics: An incomplete summary. Retrieved from `https://merenlab.org/2020/07/27/history-of-metagenomics/`.

1.2.1 Back to the Origins

It was difficult to find a starting point for the history of metagenomics. As a primer, I thought it would be great to begin before the word metagenomics was invented, with arguably the most prominent figure of microbial ecology and evolution, Carl Woese. His philosophical and experimental contributions to 16S ribosomal RNA gene amplicon sequencing is an important catalyst that launched us into the era of metagenomics because it opened our eyes to the unimaginable diversity of microbes. Let's get started!

Carl Woese fundamentally changed our view of biological diversity by utilizing the 16S ribosomal RNA molecule as a phylogenetic marker. With this, he added to the tree of life 'archaea' as a new domain of life (Woese and Fox, 1977), and primed 16S rRNA gene to be later used as a tool in high-throughput amplicon sequencing studies to explore microbial ecology. The Carl Woese's review, "Bacterial Evolution", is a microbial perspective written over 30 years ago and yet still feels relevant today (Woese, 1987). The manuscript provided a breadth of reasoning and evidence for the utility of 16S rRNA gene as a phylogenetic marker in biology, as well as a source of wisdom and outlook on the field of microbiology. Although the majority of the review is describing arguments for the usage of 16S rRNA gene to explore bacterial evolution, I focused on the introductory sections that presented timeless ideas that impact the way I perceive metagenomics and microbiology today.

The "Perspective" chapter reads like a microbial op-ed, reflecting on how nucleic acid se-

quencing fundamentally changed our understanding of genetics and evolution, and how taxonomy before sequencing was, "a fruitless search". Woese has a few epic quotes in this section that would be a injustice to paraphrase and are worth discussing more in-depth:

> "The cell is basically a historical document, and gaining the capacity to read it (by sequencing of genes) cannot but drastically alter the way we look at all biology." (Woese, 1987)

This premise highlights how microbiology research has been fundamentally technology-driven. Our "capacity" to sequence genes drastically changed our understanding of diversity and protein evolution. What other technological innovations have pushed microbiology forward? Since the field of metagenomics rests on the technology of Next-Generation sequencing and long-read sequencing, I was inspired to review the origins of a few fundamental microbiology technologies.

I would argue that most knowledge about microbiology could not be discovered *a priori*. In fact, for 1000's of years humanity had only interacted with the miracles (e.g. bread and beer) and disasters (e.g. pandemics) of microbial life using senses that could not elucidate the perpetrator (can we hear microbes?). Yet, there was an "apple falling from the tree" moment when Antony van Leeuwenhoek made technological improvements to the single lens light microscope. This Dutch scientist is referred to by many as the "Father of Microbiology" and is credited for discovering bacteria and protists (Lane, 2015)!

The realization of the microbial world permanently changed our view of biological diversity; At the time, adding a new Kingdom. Yet, visually quantifying and categorizing microbes through microscopes had its limits. How could one assign metabolic abilities to a "species" when most microbes look the same in their natural community assemblages? To begin and ask questions about "What are they doing?" and "Who is there?" another innovation was required.

To address basic questions about microbial physiology and taxonomy, pure culturing techniques needed to be invented. Robert Koch is credited for developing the first microbial isolation techniques. He grew the first bacterial colonies on thin potato slices (Madigan et al. 2018)! Koch had the genius insight to assume that a colony was monoclonal and formed from a single colonizing bacteria. This catalyzed his research and allowed for the first pure culture (Reinkulturen) experiments. Koch's assistant, Julius Richard Petri, expanded on Koch's potato slices and invented, you guessed it, the Petri dish. This technological leap is responsible for the "Golden Age of Bacteriology" and an exponential increase in our knowledge of microbial diversity and function leading to discoveries such as antibiotics. Yet, as the century went on it became clear that culturing techniques were not keeping up with the immense diversity of microbes. Although there were many observations describing the discrepancy between the diversity of microbes that were able to be cultured versus those seen in microscopes, Staley et al. 1985 first deemed the problem as, "The Great Plate Count Anomaly" (Staley and Konopka 1985). We are still battling this culturing bias today and it has a new name, "The Uncultured Majority" (Michael S. RappÃľ and Stephen J. Giovannoni). Metagenomic sequencing would later provide a more unfiltered glimpse into the microbial world, regardless of culturing ability.

1.2.2 The first metagenomic sampling

16S rRNA gene sequencing studies expanded our understanding of microbial diversity and ecology, and ushered in the era of culture-independent studies. This era was catalyzed by improved PCR primers and economic Sanger sequencing. Although discoveries of novel microbial diversity seemed endless, limitations of amplicon sequencing began to arise. Cultured representatives could not keep up with the discovery of new phylotypes observed in the environment. This is a continued disconnect even today - it is very difficult to isolate ecologically relevant microbes and/or representatives of the total niche space in an environment. Due to this culturing bias it became hard to extrapolate metabolic capabilities and phenotypes of cul-

tured representatives to newly discovered sequence phylotypes; more genomic context was needed. As Carl Woese puts it, "In the extreme, interspecies exchanges of genes could be so rampant, so broad spread, that a bacterium would not actually have a history in its own right; it would be an evolutionary chimera, each with its own history" (Woese, 1987). One prominent example is *Staphylococcus aureus*. Strains that have identical marker genes, such as identical 16S rRNA genes, may have different antibiotic resistance phenotypes obtained from genomic islands via horizontal gene transfer. A new sequencing technique was needed to access more genetic information from environmental microbes to expand hypothesis-generating capabilities.

Stein et al. (1996) recognized the limitations of amplicon sequence and pushed the field forward with the first attempt of metagenomic sequencing in Hawaiian ocean water (although the name metagenomic sequencing had not been coined yet) [1]. Stein et al. (1996) posited

1. Where did the term "metagenomics" come from? The term metagenomics was first coined by Handelsman et al. (1998), a few years after Stein et al. (1996). Interestingly, metagenomics was first used to describe an approach for biosynthetic gene cluster (BGC) research. Natural product discovery via BGC research at this point had been driven by microbial isolation with selective media to search for specific antibiotic and metabolic phenotypes. Handelsman et al. (1998) recognized how using an entire sample's worth of DNA could be beneficial for accessing novel BGC loci (similar to how Stein et al. (1996) wanted to explore novel metabolisms in archaea described below). The method described in this paper included cloning potential BGC genomic fragments from an environmental sample into *E. coli* vectors and phenotype screening. BGC research would no longer be constrained by the Great Plate Anomaly!
The first sentence with the term metagenome:

> "The methodology [cloning of environmental DNA into E. coli for phenotype screening] has been made possible by advances in molecular biology and Eukaryotic genomics, which have laid the groundwork for cloning and functional analysis of the collective genomes of soil microflora, which we term the metagenome of the soil." Handelsman et al. (1998)

Considering this is the first use of the term metagenomics, I want to take a second to break down its etymology. The prefix "meta-" is Greek in origin and has a few applications including the inference of later, behind, or beyond in time and space (e.g. metacarpus). But when used as a prefix to a subject, it means the critical and/or abstract analysis of the subject (e.g. metaphysics, metalinguistics).
Genomics is the study of the genetic material of an organism. If genomics is a subject, the most direct interpretation of metagenomics would be, "to think abstractly about the genome." Yet in the case of Handelsman et al. (1998) and even todayâĂŹs field, this interpretation may not be appropriate. In fact, I believe meta- in the context of metagenomics is inferring "going beyond the genome." Microbiology ecology had been constrained by amplifying taxonomic markers and extrapolating the functional potential of microbes by linking them to known isolates. For example, in BGC research, you had to use 16S amplicon sequencing to see if potential microbes of interest (e.g. actinomycetes) were in a specific environment, then hope they could grow in standard culture techniques and be screened for specific phenotypes. This is essentially accessing one genome at a time. Handelsman et al. (1998) explains how the metagenome, "the collective genome," of the soil can now be accessed by cloning

that there is a need for culture-independent innovations because of the lack of cultured representatives of newly discovered phylotypes and examples of prokaryotic lineages having a large metabolic diversity of physiological capabilities. This manuscript looks to further culture-independent microbial ecology by investigating the marine Archaea clade, Creanrchaeota. They do this by going beyond surveying taxonomic markers of diversity via 16S and 18S rRNA amplicon sequencing. Instead, they extracted large fragments of genomic DNA (up to 40 Kb) from seawater and sequence via a Fosmid cloning approach [2].

At this point, it was known that there were abundant marine Archaea in the surface ocean (DeLong, 1992; Fuhrman et al., 1992; DeLong et al., 1994), but there was no cultured representative, only amplicon marker gene evidence. "Although the genotypic and phenotypic properties of marine pelagic archaea are unknown at present, their abundance and ecological distribution suggest that they may represent entirely new phenotypic groups within the domain Archaea" (Stein et al., 1996). With only marker genes, inferences of metabolic capabilities can only be made by looking at phylogenetically similar cultured representatives. Leveraging metagenomic sequencing of large genomic fragments, Stein et al. (1996) could use "genomic walking" (looking upstream and downstream on genomic fragments from a phylogenetic marker) in the marine environment. In essence, their goal was to collect large fragments from

---

genomic fragments into Fosmid vectors and expressing them in E. coli. I would argue that this is essentially a survey of the functional potential of a microbial community, an ecological perspective. Maybe metagenomics was not the best word to describe our field? Regardless, it stuck and I do not have a problem with it.

2. Whoa! Metagenomic sequencing is different now... Shotgun metagenomic sequencing with Illumina methods have come a long way since the time of this paper. Stein et al. (1996) could not just perform an Illumina library preparation. Instead, they had to use an *E. coli* Fosmid cloning vector to generate their library of large DNA fragments from the marine environment (30 liters of seawater were filtered!). Furthermore, prior to their library prep, they performed a standard 16S rRNA screen on a subset of each sample to confirm there was Creanrchaeota genomic material in the sample. This makes sense considering how labor intensive and expensive this sequencing must have been at the time. Next, the DNA was digested using restriction enzymes to achieve smaller fragments and then cloned into the Fosmid vector library. The clones were then screened for specifically archaeal DNA fragments by amplifying 16S rRNA genes content from the fragments. Fragments that did not contain a part of the Crenarchaeota ribosomal subunit were not further sequenced. The clones that passed this stage were then prepared for sequencing. Final DNA sequences were analyzed for homology upstream and downstream of the 16S phylogenetic anchor against the NCBI NR protein database using BLAST (Altschul et al., 1990). Glad we don't have to do that anymore!

picoplankton genomic DNA in an attempt to extract the genomic DNA from Crenarchaeota. By "walking" upstream and downstream from the phylogenetic marker, they hoped to obtain physiological insights.

Phylogenetic analysis revealed that 33% of genes found on the 40 kb fragments had significant homology to previously sequenced genes. Additionally, most genes located on the fragments were related to ribosomal proteins. Other notable genes included glutamate 1-semialdehyde transferase and RNA helicase, both of which are core microbial functions. Finally, there were examples of the entire 16S-23S operons found on fosmids. This cutting-edge environmental sequencing study at the time was unfortunately constrained by the use of the 16S phylogenetic marker of Crenarchaeota. The genomic neighborhood that was explored by "genomic walking" only allowed the discovery of neighboring genes of core functions of Archaea and a highly studied area of genomes. If there was another phylogenetic marker of Archaea more related to metabolism, there could have been a chance to explore a more novel genomic neighborhood away from the 16S region. Although the findings of the paper did not reveal novel Archaeal metabolisms for the time, it laid groundwork for a new era of environmental shotgun sequencing.

1.2.3 Order to chaos - Extracting genomes from metagenomes

Jumping ahead to another phase of the metagenomics progression, the era of genome-resolved metagenomics. Now that we had the ability to access the "collective genome" of an environmental sample, research was constrained by the immense amount of data. To address this data analysis challenge there were numerous data analysis innovations in comparative metagenomics from clustering orthologs (Yooseph et al., 2007) to gene catalogs (Turnbaugh et al., 2007). Yet, as Katherine McMahon puts it, "Genes are expressed within cells, not in a homogenized cytoplasmic soup." (McMahon, 2015) In other words, the first comparative metagenomics projects were genome agnostic. It was like comparing a Jackson Pollock paint-

ing of functional potential to a Jean-Paul Riopelle painting of short reads.

(Tyson et al., 2004) is the first example of genome-resolved metagenomics. The questions being explored in this paper were both methodologically and ecologically based. The technical question was whether it was possible to assemble and extract groups of metagenomic contigs that represent a collective genome from a population of similar microbes? The ecological question in this paper was what is the functional potential of the individual microbial community members in a low-diversity biofilm found in an acid mine drainage (pH 0.83)? Understanding life in extreme environments is important for fundamental questions in regards to astrobiology and geobiology. Although there were cultured representatives of extreme acidic environments at the time, it was unclear if the naturally occurring community of this acidic biofilm mirrored the isolates taxonomically and metabolically.

To address these questions, (Tyson et al., 2004) used unprecedentedly deep shotgun metagenomic sequencing (103,462 reads) to access the functional repertoire of this gnarly biofilm. Interestingly, they used the same method for DNA extraction as Stein et al. (1996) discussed above! Some other key strategies they implemented to prepare their data for genome reconstruction attempt were: (1) adjusting their assembly strategy to not penalize non-uniform read depth and (2) allowed read mapping of 95% identity. These two adjustments allowed for more contigs to be assembled from the environmental sample and would incorporate more variation in the final contigs.

Once contigs were assembled, (Tyson et al., 2004) grouped them together ("binned") using GC content then subsequently refined the groups ("bins") using differential coverage of short reads against the reference contigs. A bimodal distribution of contigs arose when statistics from their metagenomic contigs were visualized (Visualization of metagenomic bins has developed over the last 15 years, check out its progression in Figure 1.1). Tyson et al. (2004) highlighted that their success in reconstructing genomes from metagenomes in this environment was contingent upon the low diversity in the biofilm. The results of their binning yielded

two "nearly complete" genomes and 3 partially recovered metagenomes (aka metagenomic assembled genomes - MAGs). This data analysis strategy began the transition of the field of metagenomics from primarily read-based based approaches to a genome-centric phase.

1.2.4 Leveraging Genome-resolved Metagenomics

Genome-resolved metagenomics has revolutionized our understanding of uncultured microbes and catalyzed unprecedented discoveries that have impacted multiple fields, from biogeochemistry to evolutionary biology. To me, metagenomics is a fantastic hypothesis generator, and at this point in this thesis introduction, I wanted to highlight some discoveries made possible by genome-resolved metagenomics.

Bioavailable nitrogen in the global oceans is a hot topic because of its impact on the carbon cycle and primary productivity. For how critical nitrogen is to the ocean ecosystem, a surprisingly small group of Cyanobacteria (e.g., UCYN-A and Trichodesmium) are theorized to be responsible for the majority of nitrogen fixation in the global oceans. Other Bacterial phyla such Proteobacteria and Firmicutes have been linked to nitrogen fixation via PCR amplicon studies of the *nifH* gene. Although, surveys showed that these nitrogen fixers (diazotrophs) tend to be from the rare biosphere (in very low abundance) with only a few cultured representatives. This left a significant knowledge gap in terms of the function potential of these enigmatic diazotrophs and how they impact biogeochemistry. Delmont and Eren (2018) leveraged metagenomic binning and extracted around 1,000 MAGs from the Tara Oceans Project (A global ocean microbiome sampling effort with unprecedentedly deep sequencing) (Sunagawa et al., 2015). Some of these newly extracted MAGs contained nitrogen-fixing capability! An interesting highlight from this collection of MAGs was identifying diazotrophic capabilities in Planctomycetes for the first time. Genome-resolved metagenomics permitted access to uncultured microbes with key biogeochemical capabilities and revealed the rest of their metabolic potential. By putting contigs with nitrogen-fixing genes into genomic context, we can now

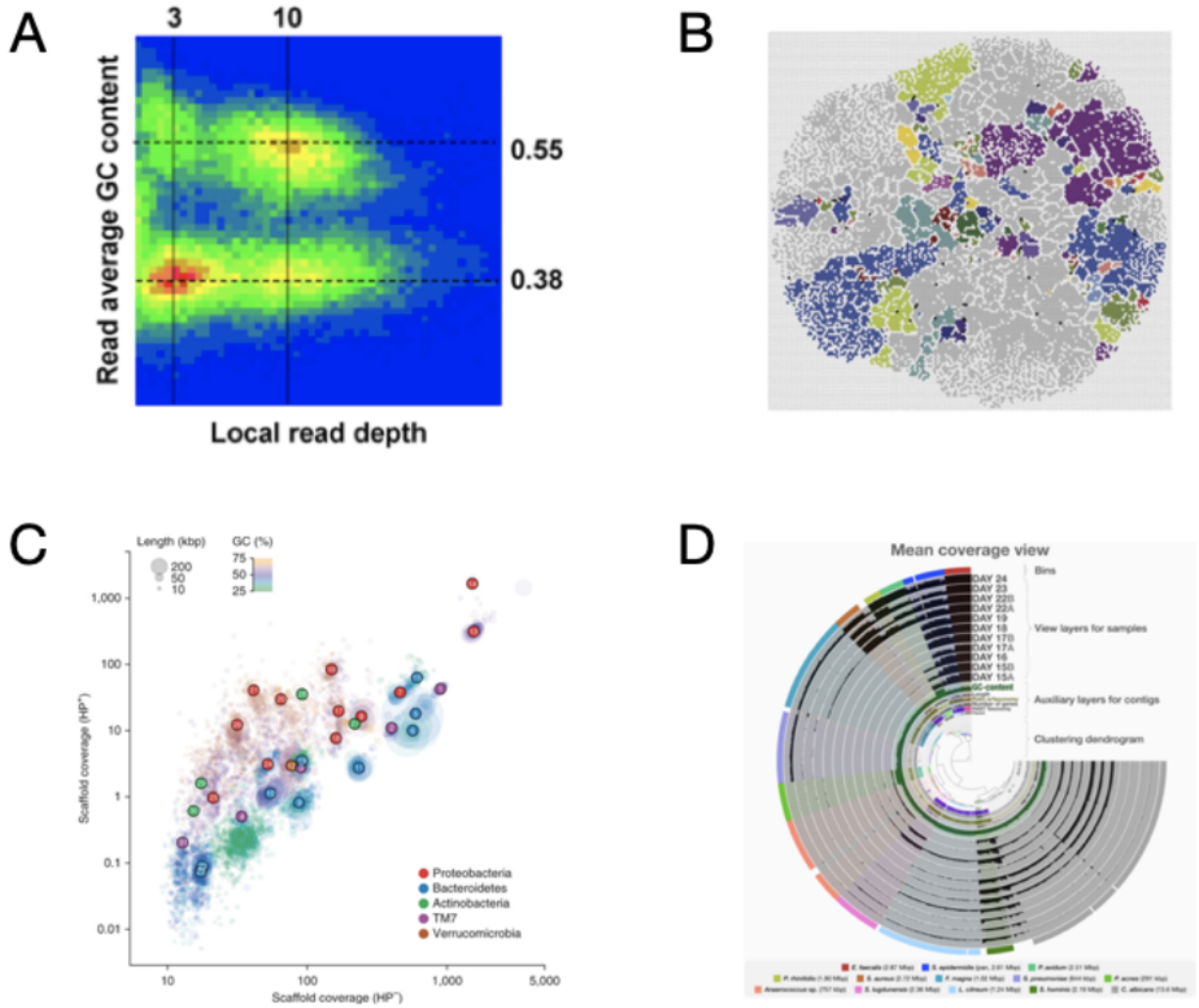

Figure 1.1: **Progression of genome-resolved metagenomics visualizations** Progression of genome-resolved metagenomics visualizations. (A) The first visualization of binned contigs using GC content and read depth coverage (Tyson et al., 2004). (B) Self-organizing map approach based on tetranucleotide z-scores (Weber et al., 2011). (C) Clustering of contig scaffolds based on contig read depth from two separate DNA extraction protocols, with and without hot phenol (HP^+ and HP^-), and colored by GC content (Albertsen et al., 2013). (D) Anvi'o interactive interface with contig clustering based on tetranucleotide content and differential read coverage from co-assembled metagenomic samples (Eren et al., 2015).

create targeted culturing campaigns to isolate these novel diazotrophs.

Genome-resolved metagenomics has catalyzed bacterial phylogenomics of uncultured microbes. Brown et al. (2015) leverage MAGs from deep aquifer groundwater to unveil more than 35 new bacterial Phyla deemed candidate phyla radiation (CPR). This unveiled extraordinary microbial diversity that was never shown using 16S amplicon surveys because the 16S rRNA gene sequences these enigmatic CPR were too divergent for standard bacterial primers. Metagenomic sequencing is a less biased sequencing approach compared to PCR-based approaches because there is no amplification step, thus allowing for more microbial biodiversity to be uncovered. Hug et al. (2016) leveraged this new set of CPR genomes and re-calculated the tree of life leveraging phylogenomics, which included all Domains. Without MAGs, the genes necessary to make a concatenated single-copy core gene alignments would not have been possible.

MAGs have re-rooted our interpretation of the origin of Eukaryotes. MAGs from the bottom of the ocean (really, the bottom of the ocean in a hydrothermal vent revealed a whole new class of archaea that contain eukaryotic-like genes (Asgard archaea named after Scandinavian gods). Spang et al. (2015) proposed that the tree of life has two, not three, branches. In other words, Eukaryotes branched off from a more ancient archaeal branch. This is a highly debated topic where other groups have claimed there are still 3 branches (Da Cunha et al., 2017). These analyses were only possible through putting contigs in a genomic context through binning which allowed for a phylogenomic approach to place these undiscovered genomes on a provocative branch of the tree of life.

The discoveries described above have all relied upon putting metagenomic assembled contigs into the genomic context of a microbial population genome. Although binning metagenomic contigs has the potential to yield a mosaic mixture of microbial populations, careful curation can yield MAGs with high probability of accuracy. In essence, MAGs have given scientists access to the genomic potential of microbes that are not yet cultured and maybe, the unculturable.

able! Although there are many inherent limitations to binning, MAGs are a great hypothesis-generation tool that can inspire new cultivation experiments that could have never been imagined. Imachi et al. (2020) spent 12 years enriching Asgard archaea from marine sediment 2,533 m below sea level (using powdered milk as one of their media ingredients!). This archaea named 'Candidatus Prometheoarchaeum syntrophicumthus' was a literal missing-link for evidence of the endosymbiotic theory. I would argue that motivation for a decade-long enrichment experiment would have dwindled if it were not for discoveries such as Spang et al. (2015).

1.3 Benchmarking the breadth of large genome datasets to uncover novel microbiology

Building collections of microbes is not a modern idea but actually a classical activity in microbiology. Microbiologists have been isolating microorganisms before DNA was discovered as the hereditary material. This is demonstrated by the second requirement for identifying a pathogen in Koch's postulates:

"The microorganism must be isolated from a diseased organism and grown in pure culture." - Robert Koch (Brock, 1999)

Recently, MAGs were reconstructed from $\sim$100,000 metagenomic samples spanning multiple biomes and yielded over 1 million MAGs (Schmidt et al., 2023). This remarkable achievement underscores the high-throughput nature of genome recovery and firmly establishes that we are in an era of big data in microbiology. Genomes are critical for novel discoveries in fundamental microbiology and serve as vast resources for novel proteins and natural products (e.g. enzymes, biosynthetic gene clusters) with broad implications for medicine and biotechnology (Paoli et al., 2022; Chen et al., 2024). Additionally, genome collections serve as reference sequences for homology detection in new sequencing data, crucial for characterizing previously

unexplored microbiomes. As the corpus of genomic DNA expands, researchers are increasingly using these collections to train foundation DNA large language models (Nguyen et al., 2024). Overall, these expanded applications of genome collection provide impetus to continue extracting microbial genomes from the environment.

In the past few decades, the explosion of microbial genomes in genome collections can be attributed to technological innovations in genome recovery. Even the classical strategy of microbial cultivation has seen innovations (Lewis et al., 2021). For example, Cross et al. (2019) leveraged an innovative strategy called "reverse genomics" to create antibodies that target specific taxa and isolate them with flow cytometry for genome sequencing. Additionally, microfluidic platforms have been leveraged for high-throughput cultivation, enabling the growth of slow-growing anaerobic microbes (Watterson et al., 2020). However, despite innovative cultivation techniques, microbiology remains limited by the unique growth conditions, media requirements, and symbiotic relationships, preventing us from isolating all microbes on Earth.

Genome-resolved metagenomics has been the main culture-independent technological driver for increasing the size of genome collections in recent decades. Specifically, innovations such as open-source bioinformatics tools for metagenomic assembly (Li et al., 2015) and binning (Kang et al., 2015; Pan et al., 2022), combined with workflow management software like Snakemake and Nextflow (Koster and Rahmann, 2012), which connect these data processing steps (Shaiber et al., 2020), have made genome recovery methods both highly accessible and scalable. Due to the high throughput of genome-resolved metagenomics, it will likely be the predominant genome recovery method for the near future.

Another genome recovery method that has gained traction in recent years is single amplified genomics. This approach isolates individual microbial cells from environmental samples using microfluidics and performs whole-genome amplification. It offers key advantages over genome-resolved metagenomics, which often struggles to recover genomes for rare microbes because of low sequencing depth and populations with high degrees of polymorphism pro-

ducing fragmented assemblies (Chen et al., 2020b). Unlike metagenomics, single amplified genomics is limited by the number of cells in a sample, making it particularly valuable for studying microbial populations that are rare in the biosphere or resistant to recovery through traditional metagenomic assembly (Hosokawa and Nishikawa, 2024). Additionally, the method can associate mobile genetic elements, such as plasmids, with genomes since it amplifies each genome separately. Unfortunately, a technological pitfall that holds back this method is the low completion and contamination of genomes after whole genome amplification, limiting the number of high-quality genomes that can be produced from this method.

Numerous large-scale genome sampling projects have produced enormous biome-specific datasets of MAGs (Almeida et al., 2019; Parks et al., 2017; Pasolli et al., 2019; Delmont and Eren, 2018; Almeida et al., 2020), and SAGs (Pachiadaki et al., 2019; Kawano-Sugaya et al., 2024) profiling microbiomes from across the planet including the oceans, soils, and those associated with animals. These datasets have added a significant amount of new diversity to the tree of life and provide a way forward to explore the uncultured majority of microbes and their functions. Furthermore, large-scale multi-biome resources are now available with multiple kinds of genome recovery methods, including EBI MAGnify (Richardson et al., 2023), SPIRE (Schmidt et al., 2023), and the DOE's IMG/M (Nayfach et al., 2021). Importantly, the Genome Taxonomy Database (GTDB) has curated a large genome to calculate a phylogenomic tree for the archaea and bacteria domains providing standardized phylogeny-based taxonomies (Parks et al., 2018). With these genome resources and domain scale phylogenies, we can now address questions about microbial functions on the tree-of-life scale. For example, we can identify microbial functions with broad and limited phylogenetic breadth across bacteria e.g. what are the phylogenetic constraints of a particular metabolic pathway?

Another step forward in large-scale genome recovery efforts is experimentally validating genes and pathways discovered in novel genomes. Paoli et al. (2022) used their global ocean genome dataset to identify biosynthetic gene clusters in genomes reconstructed from the deep

ocean (x > 2,000 m) and explored their enzymology in the lab. In another example of leveraging ocean genome collections, Chen et al. (2024) tested the gene editing efficiency of a novel CRISPR system in laboratory conditions. Additionally, Durrant et al. (2022) surveyed publically available genomes for novel serine recombinases and experimentally validated them in the lab to create innovative strategies for genome editing. Overall, this workflow of extracting genomes from the environment and mining them for novel biotechnology and medical applications is highly valuable and should be further pursued. However, as the field continues to expand these genome collections, a fundamental question in microbial ecology arises, how well do these collections represent the phylogenetic diversity and ecology of microbes in the environment?

In this thesis, I attempt to address this question by taking a step back in the genome recovery process, the metagenomic assembly.

1.4 An open-source workflow to explore the phylogeography of gene families

Methods to assess the phylogenetic and ecological diversity of genome collections are essential for optimizing genome acquisition strategies, both to achieve broad sampling of microbiomes and target specific taxa. For decades, microbiology has documented a persistent disconnect between the diversity of taxa detected in the environment and the rates of successful culturability (Staley and Konopka, 1985) or genome extraction through metagenomics (Chen et al., 2020b). A notable example is the lack of representative genomes for the widespread marine bacterial order SAR11. Genome recovery for this taxon is challenging because its specialized growth requirements hinder easy isolation (Giovannoni, 2017). Additionally, the extensive microdiversity within SAR11 populations leads to fragmented metagenomic assemblies, reducing the ability of binning tools to reconstruct complete genomes from contigs (Chen et al., 2020b). Current methodologies to benchmark genome recovery involve using metage-

nomic read recruitment statistics to assess what proportion of reads align with genomes. This approach assumes that the fraction of reads mapped to a genome collection serves as an indicator of how well the collection represents the microbial community in the environment. Although this is an efficient method, it has two pitfalls, including (1) an inability to track the microbial diversity found in the unmapped fraction and (2) no direct link to contigs for targeted genome recovery.

Although genome collections are valuable, they do not contain all the genetic information present in the environment. The ground truth genomic diversity of a microbiome is in the environment itself. If we could directly count all microbes and individually sequence their genomes, then the story of microbial ecology would stop here. Unfortunately, every step of the genome recovery process has potential to lose information. For example, microbes with cell walls resistant to DNA extraction kits, populations that are difficult to assemble in metagenomes, and populations that are difficult to bin (reconstruct genomes from metagenomes) all can lower genome recovery from various taxa. Stepping back in the genome recovery process, metagenomic assemblies present an opportunity to identify genomic elements assembled from metagenomic reads that are not included in the final genome collection. Furthermore, gene families that efficiently assemble from metagenomes and act as proxies for biological phenomena can provide more genomic information about microbial ecology beyond genome collections.

In this thesis, I focus on analyzing the distribution of ribosomal proteins, a subset of SCGs that assemble efficiently in metagenomes, to track the distribution of genome recovery of taxa in a metagenomic assembly. During the writing of this thesis, Wu et al. (2025) measured genome recovery by measuring the phylogenetic diversity (PD) (Wu et al., 2009) of MAGs and isolate genomes compared to SCGs from metagenomic assemblies. While this is an efficient method for explaining the contributions of various genome sampling methods to the tree of life, it does not account for ecological insights gained from metagenomic read recruitment. In fact,

the combination of phylogenetics and biogeography is a powerful data integration strategy to track microbial ecology (Ustick et al., 2023; Gaia et al., 2023). In Chapter 3 of my thesis, I describe the anvi'o EcoPhylo workflow, which tracks the phylogeography of gene families to explore microbial ecology across environments and track genome recovery rates of taxa and genome recovery methodologies.

1.5 Summary of thesis topics

This thesis presents the development and application of a novel computational workflow, the anvi'o EcoPhylo workflow, to track the phylogeography of gene families across microbial communities in diverse environments while highlighting my contributions to open-source software supporting 'omics analyses in microbiology. At its core, this work explores the premise that new discoveries lie beyond existing genome collections and that metagenomic data, both within assemblies and raw reads, can provide deeper insights into microbiomes. A key focus is assessing whether the vast microbial genome collections comprehensively capture the full breadth of microbial diversity and associated genes in the environment. To address this, I demonstrate that tracking the phylogeography of gene families, which serve as proxies for microbial diversity and function, extends our understanding beyond genome collections to metagenomic assemblies across environments and experiments. Furthermore, this thesis details the implementation of open-source bioinformatics tools and workflows to empower microbiologists and catalyze future discoveries, ensuring these technical advancements remain accessible and scalable for the broader scientific community.

To begin my thesis, Chapter 2 describes my open-source software contributions, bioinformatics tools, and workflows to catalyze the era of data-intensive microbiology. In Chapter 3, I highlight the keystone scientific contribution of my dissertation, a study that applies my computational workflow, the anvi'o EcoPhylo workflow, to track genome recovery rates of microbes through the phylogeography of ribosomal proteins. The combination of tracking both

the phylogenetic and biogeographical relationships of microbial communities offers a powerful platform to contextualize microbes without genome representation and benchmark genome collections. Finally, in Chapter 4, I present a composite chapter of further applications of the Ecophylo workflow. This includes surveying the evolutionary landscape of anaerobic flavin respiratory reductases in gut microbes. Additionally, I showcase EcoPhylo as an efficient method for detecting the presence of genomes in metagenomic data.

Overall, this dissertation provides insights into exploring microbial ecology beyond genome collections, focusing on metagenomic assemblies to uncover expanded collections of gene families.

CHAPTER 2

OPEN-SOURCE SOFTWARE EMPOWERS SCIENTISTS IN THE ERA OF BIG DATA AND MICROBIOLOGY

2.1 Introduction

This section is derived from the following section of a publication:

A. Murat Eren, Evan Kiefl, Alon Shaiber, Iva Veseli, Samuel E. Miller, **Matthew S. Schechter**, et al. "Community-led, integrated, reproducible multi-omics with anvi'o." *Nat. Microbiol.* **6**, 3–6 (2021). https://doi.org/10.1038/s41564-020-00834-3.

Novel data analysis strategies are essential to continuously extract new insights from the growing repository of publicly available genomes and metagenomes deposited from microbiome studies. Due to the intense influx of new data and data types, many bioinformatics tools are being developed to process, visualize, and store the data (Callahan et al., 2018). Eren et al. (2021) makes the argument that these programs fall into two categories: (1) 'essential tools' and (2) workflows. Bioinformatics tools are small programs designed to perform specific tasks and process 'omics data and workflows string together tools to efficiently analyze these massive datasets to produce summary tables and figures.

Traditional bioinformatics pipelines often produce static visualizations, such as bar charts or dimensionality-reduction plots, which fall short of enabling researchers to interact dynamically with the complexities of 'omics data. To address these limitations, interactive visualization tools, like those offered by the anvi'o platform (Eren and Delmont, 2024), empower researchers to explore their data more thoroughly. For example, the anvi'o platform enables users to interactively visualize read recruitment across various data types - genomics, metagenomics, and transcriptomics - alongside nucleotide variance patterns like single-nucleotide, single-codon, and single-amino acid variants. Live interaction with these datatypes has led to numerous discoveries that might have been missed without this access (Reveillaud et al., 2019; Fogarty

et al., 2024).

During my Ph.D., one of my overarching goals was to improve access and education to computational biology and analysis of 'omics data in microbiology. To accomplish this, one of the bi-products of my research endeavors was a variety of open-source software contributions to the microbiology community. I chose to contribute these pieces of software to the anvi'o platform Eren et al. (2021) due to its focus on interactive visualization, open-source philosophy, and commitment to open science and teaching. This section of my thesis details a subset of these contributions. I start by discussing my first solo contribution to anvi'o, 'anvi-analyze-synteny'. This program was implemented to explore synteny conservation across the hypervariable gene content of *B. fragilis* capsular polysaccharide utilization loci. Next, I document a suite of programs to annotate carbohydrate-active enzymes (CAZymes), including 'anvi-setup-cazymes' and 'anvi-run-cazymes', which leverage the publically available database of protein family CAZymes (Yin et al., 2012) to annotate genes predicted from genomes and metagenomes in anvi'o. Finally, I discuss 'anvi-export-locus,' a powerful program that can cut genomic loci out of a given genome or metagenomic assemblies and scale to genome collections of thousands of genomes, such as The Genome Taxonomy Database (Parks et al., 2022).

Subsequently, in this chapter, I highlight the workflows I contributed to the anvi'o platform. Workflows are similar to bioinformatics pipelines but differ in their scalability on high-performance computing clusters and their modularity to be customized to specific science questions. The first workflow I document is the anvi'o 'sra-download' workflow, a scalable workflow leveraging NCBI SRA programs (https://github.com/ncbi/sra-tools) to efficiently download paired-end metagenomic reads. Next, I document the keystone contribution of my Ph.D, the anvi'o EcoPhylo workflow. This workflow analyzes gene family phylogenetic and biogeographic distribution patterns across samples using metagenomic read recruitment. It features an interactive interface that allows users to seamlessly switch between co-occurrence and phyloge-

netic perspectives of gene family distribution. The applications of this workflow and its scientific findings are discussed in Chapters 3 and 4.

To enhance my software contributions and improve accessibility in computational biology and the anvi'o environment, I have documented these tools and workflows through detailed tutorials, as well as comprehensive code and artifact documentation for developers. This will allow these software contributions to be robust additions to the platform, allowing users and developers to leverage these tools and build on top of them long into the future. In addition to software development, my colleague Iva Veseli, Ph.D. and I designed and led multiple in-person lecture series covering core concepts in computational microbiology, including metagenomics and metabolomics data analysis in 2022 `https://merenlab.org/tutorials/dfi-metagenomics-workshop/` and 2023 `https://merenlab.org/tutorials/dfi-metagenomics-workshop/`. Additionally, I lectured on the EcoPhylo workflow at the 2021 workshop on Emerging Bioinformatics Applications for Microbial Ecogenomics (https://maignienlab.gitlab.io/ebame6/) as well as TA'd the associated anvi'o tutorials along with my colleague Florian Trigodet, Ph.D. The majority of the publically available resources I created during my Ph.D. can be found at https://anvio.org/people/mschecht/.

2.2 Exploring n-grams of syntenous genes

2.2.1 anvi-analyze-synteny

Comparing the differences in gene content and synteny of microbial loci between genomes or locus copies within the same genome can be a challenging task if a locus is hypervariable, i.e., gaining or losing gene content quickly over time. Pangenomics offers an elegant solution to compare gene content efficiently by creating gene clusters, but it does not account for synteny. Furthermore, traditional contig alignment strategies do not resolve open-reading frames and microbial functions. In the context of an operon, synteny is essential to explore because the

order of gene transcription can impact final protein products and function. Additionally, synteny can help elucidate which operons are more related to each other. In the field of Natural Language Processing, n-grams are used to track sequences of adjacent words and can be used to compare distances between texts by employing n-gram decomposition. The program ‘anvi-analyze-synteny‘ leverages this concept by breaking down microbial loci into blocks of syntenic genes with user-defined ranges.

The webpage https://anvio.org/help/main/programs/anvi-analyze-synteny/ serves documentation for this program. Briefly, ‘anvi-analyze-synteny‘ counts n-grams by converting contigs into strings of open reading frame annotations for a given user-defined gene annotation source. Using a source annotation for functions, anvi’o will use a sliding window of size `N` to deconstruct the loci of interest into n-grams and count their frequencies across a given set of contigs. The program ‘anvi-analyze-synteny‘ exports a count table of n-grams, allowing users to visualize n-gram distributions of function across microbial loci.

2.2.2 Exploring synteny in the Bacteroides fragilis polysaccharide utilization loci

This section is derived from the following section of a publication *in prep*:

Abigail C. Schmid, **Matthew S. Schechter**, et al., A multi-omics survey of *Bacteroides fragilis* hypervariable genomic islands, *in prep*.

Introduction

Capsular polysaccharides (CPS) form a protective capsule surrounding bacteria, allowing them to endure numerous stresses, including phage predation and the immune system (Whitfield et al., 2020). The microbe *Bacteroides fragilis*, an obligate anaerobe and prevalent member of the human gut microbiome, is famous for having eight phase-variable CPS loci (PSA-H),

some of which have been implicated with different impacts on the human host. The *B. fragilis* isolate NCTC 9343 PSA capsule is immunomodulatory in the human gut, with some research showing it causes abscesses in mice (Coyne et al., 2001) . In contrast, other PSX have been shown to protect mice from induced colitis (Surana and Kasper, 2012). Overall, these loci are difficult to compare due to their hypervariable synteny and gene content (Patrick et al., 2010). To investigate this, we used n-grams to explore their distribution of syntenous genes of *B. fragilis* PSX.

Synteny-aware characterization of CPS genetic structure

We implemented an n-gram decomposition strategy to identify conserved functional associations between groups of genes for each CPS class. This strategy quantifies the frequency of a set of 'n' collinear genes that occur in the same order across different sequences for a given CPS class. Using the NCBI's Clusters of Orthologous Groups (COGs) as the functional annotation source (Tatusov et al., 2000), we first deconstructed CPS sequences into collinear groups of 2 to 22 genes. Next, we counted the occurrence of unique n-grams across the entire CPS sequence database and within individual CPS classes. Conservation of n-grams across all CPS sequences was low, with a steep decrease of n-gram frequency as a function of increasing n-gram size (Figure 2.1). In fact, the most frequent n-grams consisted of 2-grams and 3-grams, indicating an overall lack of notable structure of synteny based on functional annotations alone. Not surprisingly, the most frequent n-gram was a 2-gram, which consisted of the CPS regulatory elements upxY and upxZ and occurred in 99.3% of all CPS in the dataset (the remaining 0.4% of the CPS sequences lacked this 2-gram due to missing HMM hits for the upxY (COG0250). The most common 3-gram across all CPS loci was upxY-upxZ-COG1209, which occurred in 26.9% of the CPS sequences but never in PSD and PSH. This 3-gram is an extension of the regulatory 2-gram upxY-upxZ followed by COG1209, a dTDP-glucose pyrophosphorylase, the key enzyme for the formation of dTDP-L-rhamnose, the precursor to a

core component of O-lipipolysaccharides, L-rhamnose (Mistou et al., 2016). Even though this set of genes was the most frequent 3-gram, its complete absence in PSD and PSH and its partial occurrence in sequences from other CPS classes suggest that its global significance to CPS function is limited. The most frequent 4-gram was an extension to this 3-gram with COG1898, a function related to the synthesis and secretion of O-specific LPS found in the rfb operon, and was found only in 15.6% of CPS sequences. The remaining n-grams lacked notable occurrence, revealing the absence of any set of collinear functions that were globally conserved across all CPS.

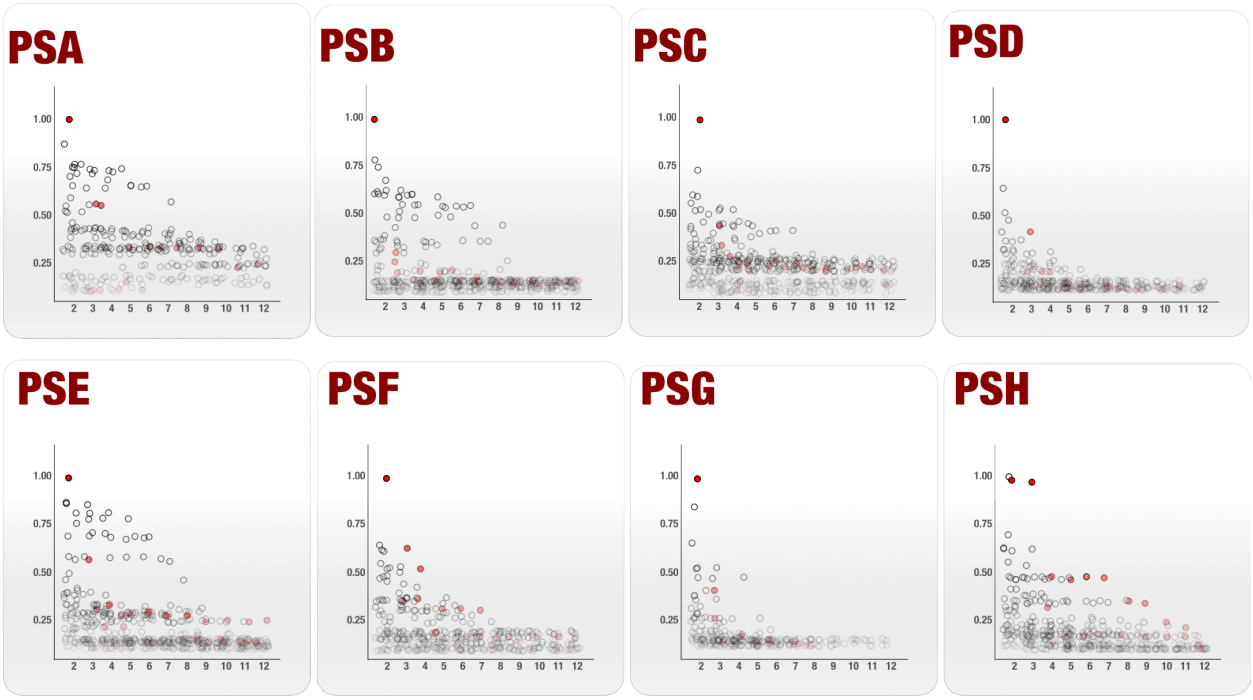

Figure 2.1: **CPS locus n-gram decomposition to explore gene synteny.** Each CPS class panel shows the percent counts of the n-grams. The x-axis is n-gram size and the y-axis is the percent occurrence of an n-gram within a CPS class. Red points represent n-grams that contain CPS regulatory genes (upxY, upxZ).

The most frequent n-gram across all CPS classes was a 3-gram that occurred in 97.5% of PSH sequences and included upxY and upxZ, as well as a gene that is involved in peptidoglycan-anchored antigen-related synthesis of carbohydrates (COG0472). The longest n-gram in more

than 50% of a CPS loci class was 7 genes long, and occurred in PSA, PSB, and PSE, each with unique gene content. On average, 95.5% of all n-grams was shared by less than 25% of sequences for any CPS class, highlighting the striking level of diversity without any conserved synteny patterns. PSD was the most variable CPS class also based on the n-gram decomposition, with only three n-grams occurring in more than 50% of the PSD sequences. Even the most frequent 3-gram across all CPS loci (upxY-upxZ-COG1209) occurred in more than 50% of sequences only in PSE and PSF, and did not show a uniform frequency pattern across CPS classes.

Conclusion

In summary, our analyses of the genetic structure of CPS sequences, based on the conservancy of collinear functions shed light on the hypervariability of these loci (Figure 2.1). We identified n-grams of gene blocks unique to individual CPS locus subtypes and shared between subtypes. Furthermore, we identified that PSD was the most variable CPS locus. At the same time, PSA, PSB, PSC, and PSF had more substantial signs of synteny conservation based on n-gram decomposition, suggesting differential selection pressures for these loci subclasses. If true, differential selection pressures raise an interesting question regarding whether there is a strict coupling of CPS classes with distinct types of environmental stresses that are yet to be answered. Our results also show that the visualization and the n-gram decomposition can infer the overall structural distribution of variability and quantitatively infer functional conservancy through recurring gene blocks for any hypervariable genomic island. Future directions should explore these patterns over a larger collection of *B. fragilis* genomes and perform quality control steps to filter out the surrounding genomic context. Additionally, a synteny-aware, graph-based pangenome strategy would combine the benefits of pangenomics and synteny analyses to further elucidate these hypervariable loci.

Material and Methods

Genomes and Metagenomes

We downloaded 113 *Bacteroides fragilis* isolate genomes from the National Center for Biotechnology Information (NCBI) Reference Sequence Database (RefSeq) and 542 human gut metagenomes that represented 159 individuals who lived in the US (Human Microbiome Project Consortium, 2012), 74 who lived in China (Qin et al., 2012), 36 who lived in Peru (Obregon-Tito et al., 2015). In addition, we acquired the *B. fragilis* clinical isolate genome 'p215' from (Vineis et al., 2016) the NCBI Sequence Read Archive dbGaP accession ID phs000262. To calculate the coverage of the *B. fragilis* clinical isolate genome we downloaded the anvi'o merged profile of the patient metagenome reported by Vineis et al. from doi:10.6084/m9.figshare.3851364.v1.

Extracting CPS loci from genomes and metagenomes

We used anvi'o v5.5 (Eren et al., 2015) and the anvi'o workflows that rely on Snakemake (Koster and Rahmann, 2012) to generate CONTIGS databases. We first used the anvi'o contigs workflow to create a CONTIGS database for each isolate genome and metagenome co-assembly. Briefly, this workflow runs (1) 'anvi-gen-contigs-database' on sequences, which uses Prodigal v2.6.3 (Hyatt et al., 2010) to identify open reading frames, (2) 'anvi-run-ncbi-cogs' on resulting anvi'o contigs databases to annotate functions using the NCBIs Clusters of Orthologous Groups database (Tatusov et al., 2000), and (3) runs 'anvi-run-hmms' to perform hidden Markov model (HMM) search on genes using HMMER 3.2.1 (Eddy, 1998). To identify upxZ genes, we used 'anvi-run-hmms' with a custom model that included a single HMM from the Pfam database (PF06603.10). We then used the anvi'o program 'anvi-export-locus' to extract CPS loci from the genomes and metagenomes. We ran 'anvi-export-locus' with the following parameters: '–use-hmm' (use anvi'o HMM framework to search for the user terms); '–search-term upxZ' (find upxZ genes)); '–num-genes 1,21 (mark one gene upstream

and 21 genes downstream of the upxZ as the target locus); and '–remove-partial-hits' (exclude candidate loci that do not include all genes of interest). Running 'anvi-export-locus' with these parameters on anvi'o contigs databases of genomes and metagenomes spawned smaller anvi'o contigs databases we could seamlessly use for all downstream analyses. A tutorial for 'anvi-export-locus' is available at the URL http://merenlab.org/export-loci.

Calculating Phylogenies of upxZ

We calculated a phylogenetic tree using the amino acid sequences for the upxZ genes in all isolate genomes. We recovered the sequences with the program 'anvi-get-sequences-for-hmm-hits' specifying '-hmm-source' and the parameters '–gene-name UpxZ' and '–get-aa-sequences'. We then used MUSCLE v3.8.425 (Edgar, 2004) with default parameters to align the amino acid sequences. Next, we computed the phylogenetic tree using IQ-TREE v1.6.6 (Nguyen et al., 2015) with '-bb 1000' and the substitution model JTT+R3 as determined by ModelFinder (Kalyaanamoorthy et al., 2017). Finally, we visualized the tree with anvi'o. We also computed two more phylogenetic trees with amino acid sequences, one with all upxZ genes in the loci we identified and one with the refined set of upxZ sequences (see Refining and classifying locus collection), in the manner described above but using the substitution model JTT+R5 (Jones et al., 1992; Yang, 1995; Soubrier et al., 2012) for all upxZ genes and JTT+R3 (Jones et al., 1992; Yang, 1995; Soubrier et al., 2012) for the refined set, both as determined by ModelFinder.

Refining and classifying locus collection

To ensure our collection of newly extracted loci only came from B. fragilis, we compared their upxZ genes to those of the eight characterized loci from the NCTC9343 B. fragilis strain. To do this, we first note that the phylogeny of upxZ sequences from isolate genomes clustered in eight groups, each of which contained one upxZ sequence from the reference NCTC9343.

Thus, we can use NCTC9343 as a single reference to verify that a given upxZ sequence belongs to B. fragilis. Using this, we recovered the upxZ sequences from the NCTC9343 genome contigs database with the program ‘anv-get-sequences-for-hmm-hits‘ (parameters: ‘–gene-name upxZ‘ and ‘–hmm-source‘). Then, we used the program ‘makeblastdb‘ with the parameter ‘-dbtype nucl‘ to create a blast database with the eight known upxZ sequences from NCTC9343 (Altschul et al., 1990). Next we used the program ‘blastn‘ to align the upxZ DNA sequences from the newly extracted loci to the NCTC9343 upxZ sequences. Any loci from isolates and metagenomes with upxZ genes that did not align to the NCTC9343 isolate upxZ genes were removed. To further scrutinize the loci from metagenomes, we blasted each full locus sequence using web-based blastn. We removed any loci where the top hit was not at least 20% alignment length to a B. fragilis genome. Finally, we classified loci as PSA through PSH by aligning the upxZ gene from a given locus to the upxZ sequences from the B. fragilis reference genome NCTC9343. Collapsing redundancy of CPS. We collapsed redundancy in our locus collection based on average nucleotide identity (ANI) and hierarchical clustering. To compute ANI for the collection of loci, we used the process described above (see Isolate genome all vs all ANI). The results were used to cluster sequences with hierarchical clustering via the hclust function in the R package ‘stats‘ (v3.4.1) with average agglomeration method. Groups were determined by the program ‘cutree‘ in the R package ‘dendextend‘ v1.8.0 (Galili, 2015) with a stringent height of .003. This height corresponds to loci that are extremely similar in percent identity and have extremely high alignment coverage because we use the element-wise product of the percent identity and alignment coverage matrices computed by pyANI. To preserve the structure of our dataset, we chose a representative locus sequence for each group alphabetically from the source (isolate or metagenome) which made up the larger proportion of the group.

N-gram analysis

We calculated n-gram occurrence patterns across all and within CPS classes using the anvi'o tool 'anvi-analyze-synteny'. Briefly, this tool deconstructs loci into windows of collinear functions (n-grams) based on a user defined functional annotation. In this analysis, we used both gene-clusters and COGs to define n-grams. The output of anvi-analyze-synteny was then post-processed and visualized in R using the Tidyverse (Wickham et al., 2019).

2.3 Annotating genomes and metagenomes with carbohydrate-active enzymes

Carbohydrate-active enZYmes (CAZymes) are proteins that either synthesize or break down carbohydrates. These enzymes are critical for microbes to help them build structural components such as capsular polysaccharides and cell walls (Wardman et al., 2022). Additionally, they help heterotrophs break down complex carbohydrates from the environment for energy conservation. Studies of CAZymes in the surface ocean have focused on their role in the carbon cycle, as heterotrophic bacteria utilize these enzymes to break down sugars released during algal blooms (Krüger et al., 2019). In general, protein annotation databases contain broad microbial functions, making it challenging to identify specific microbial functions in high resolution. The dbCAN2 CAZyme database provides a solution to this with a massive database of CAZyme protein HMMs (Yin et al., 2012). To leverage this resource and provide it to the community, I developed the programs 'anvi-setup-cazymes' and 'anvi-run-cazymes' to efficiently annotate these genes on contigs.

2.3.1 anvi-setup-cazymes

The program 'anvi-setup-cazymes' downloads and organizes a local copy of the data from dbCAN2 CAZyme HMMs (Yin et al., 2012) for use in function annotation. This program gen-

erates a cazyme-data artifact, which is required to run the program ‘anvi-run-cazymes‘. The webpage https://anvio.org/help/main/programs/anvi-setup-cazymes/ serves documentation for this program.

2.3.2 anvi-run-cazymes

The program ‘anvi-run-cazymes‘ annotates genes in an anvi’o contigs-db with functions using dbCAN CAZyme HMMs. The parameter ‘–noise-cutoff-terms‘ can be used to filter out HMM-hits with low homology. The default value is ‘–noise-cutoff-terms -E 1e-12‘. If you want to explore filtering options, check out the help menu of the underlying hmm program you are using e.g. ‘hmmsearch -h‘. The webpage https://anvio.org/help/main/programs/anvi-run-cazymes/ serves documentation for this program.

2.3.3 Case study with anvi-run-cazymes: exploring CAZyme contributions in Alteromonas genomes

This section is derived from the following section of a publication:

Iva Veseli, Michelle A. DeMers, Zachary S. Cooper, **Matthew S. Schechter**, Samuel Miller, Laura Weber, Christa B. Smith, Lidimarie T. Rodriguez, William F. Schroer, Matthew R. McIlvin, Paloma Z. Lopez, Makoto Saito, Sonya Dyhrman, A. Murat Eren, Mary Ann Moran, Rogier Braakman. Digital Microbe: a genome-informed data integration framework for team science on emerging model organisms. *Sci Data* 11, 967 (2024). https://doi.org/10.1038/s41597-024-03778-z

The development of the ‘anvi-run-cazymes‘ program was driven by community input from The National Science Foundation Science and Technology Center, The Center for Chemical Currencies of a Microbial Planet (C-CoMP). Scientists at C-CoMP in the Data Integration Working Group, expressed the need for a tool to annotate CAZymes from genomes with the

anvi'o software platform (Eren et al., 2021) to explore a genome collection of the marine heterotroph *Alteromonas*, a model organism used to study carbon cycling. After a few meetings with the team, I developed the program and made it publicly available within a month. This highlighted a dynamic collaboration between scientists and scientific programmers.

Veseli et al. (2024b) swiftly utilized 'anvi-run-cazymes' to annotate evolutionary patterns in *Alteromonas* carbohydrate utilization, revealing how carbohydrate heterotrophy may have shaped whole-genome evolution. For instance, some CAZyme phylogenies mirrored whole-genome phylogenies, suggesting vertical inheritance of these enzyme families. Conversely, other CAZyme phylogenies showed branching patterns distinct from whole-genome phylogenies, indicating potential horizontal gene transfer events to acquire specific enzyme families. Without a tool like 'anvi-run-cazymes', organizing these CAZyme annotations would have been significantly more time-consuming, potentially delaying scientific progress. This application underscores how open-source bioinformatics tools can be rapidly developed to accelerate scientific discoveries through collaboration.

2.4 Extracting loci in high throughput from genomes and metagenomic assemblies

Comparing operons and loci between microbial genomes is critical for understanding the functional and evolutionary relationships among genes and the organisms that harbor them. Operons, clusters of co-regulated genes transcribed together, often encode critical functions that can provide insights into adaptation (Coleman et al., 2006), niche specialization (Gushgari-Doyle et al., 2022), and metabolic capabilities of microbes (Crits-Christoph et al., 2020). By analyzing and comparing these features across microbial genomes, researchers can identify conserved and unique genetic elements, infer horizontal gene transfer events, and predict gene functions and interactions. There are examples of tools that extract subtypes of microbial

loci e.g. biosynthetic gene clusters (Kautsar et al., 2020). However, a flexible tool that extracts any kind of locus can empower microbiology research. To address this, I implemented the program ‘anvi-export-locus‘, which can extract microbial loci in high throughput across tree of life scale genome collections as well as metagenomic assemblies

2.4.1 anvi-export-locus

This program can cut a ’locus’ from a larger genetic context (e.g., contigs, genomes). By default, anvi’o will locate a user-defined anchor gene, extend its selection upstream and downstream based on user-defined parameters and then extract the locus to create a new anvi’o contigs database and associated FASTA file. The anchor gene can be identified with multiple protein annotation databases such as: Pfam (Finn et al., 2016), COGs (Tatusov et al., 2000), and CAZymes (Yin et al., 2012). A comprehensive description of the program can be found here: https://anvio.org/help/main/programs/anvi-export-locus/.

This section shows a tutorial I wrote to increase the accessibility of the tool ‘anvi-export-locus‘. I adjusted the post for readability in this thesis and the complete blog post can be found here: `https://merenlab.org/2019/10/17/export-locus/`

Introduction

Some genetic analyses require comparing specific genetic loci between genomes. For example, one may be interested in investigating evidence for adaptive evolution of the lac operon between different *E. coli* strains, and the first step to this analysis would be to extract the various lac operon from a collection of *E. coli* genomes.

To address this example and other genomic loci analyses alike, we present ‘anvi-export-locus‘, an anvi’o program that enables you to target regions of interest across genomes and/or metagenomic assemblies, and report sequences and/or anvi’o contigs databases for cut loci for downstream analyses.

Briefly, ‘anvi-export-locus‘ cuts out loci using two approaches: ‘default-mode‘ or what we call ‘flank-mode‘. In the ‘default-mode‘, the tool locates a designated anchor gene, then cuts upstream and downstream based on user-defined input. Notice that what is "upstream" and what is "downstream" is determined according to the direction of the anchor gene, i.e., if the anchor gene is in the reverse direction, then "upstream" would mean genes that have higher gene callers ids, and vice versa. On the other hand, ‘flank-mode‘ finds designated genes that define the left and right boundaries of the target locus, then cuts in between them. Genes to locate locus anchors or flanking genes are defined through their specific ids in anvi’o or through ‘search-terms‘ that query functional annotations or HMM hits stored in your contigs database.

The purpose of this article is to demonstrate the functionality of ‘anvi-export-locus‘ using a simple and reproducible example: extracting the lac operon from the larger genomic context of *E. coli* genomes.

Downloading *E. coli* genomes

First, let’s download Genbank files for a few representative *E. coli* strains. Please see the tutorial [here](http://merenlab.org/2019/03/14/ncbi-genome-download-magic/) describing how to automagically download genomes from NCBI and include them in anvi’o workflows. We’ll be using this tool to download a few *E. coli* genomes.

```
# Set working directory variable for later
WD=$(pwd)

# Download Genbank files
ncbi-genome-download bacteria \
                     --assembly-level chromosome,complete \
                     --genus Escherichia \
                     --metadata metadata.txt \
```

```
                    --refseq-category reference
```

Next, we'll make FASTAs, external gene calls, and functional annotations for all the Genbanks we just downloaded:

```
anvi-script-process-genbank-metadata -m metadata.txt \
                                     --output-dir ecoli \
                                     --output-fasta-txt fasta.txt
```

Just to have an idea about what is going on, please take a look at the output file 'ecoli.txt'. We will use this file to create our contigs databases for these genomes.

Generate contigs DBs

Now we need to get the fasta files into an anvi'o friendly format. There are many ways to turn your FASTA files into anvi'o contigs databases, but here we will follow our best practices and process all our files using the anvi'o contigs workflow.

First, make a config file for the anvio workflow called 'config.json':

```
anvi-run-workflow -w contigs --get-default-config config.json
```

Then run the contigs workflow!

This step may take a while depending on your computational resources. If you have any questions about running anvi'o workflows please refer to this tutorial here. If you access to an HPC or cluster computer, check out additional parameters here. HINT: you will probably need to edit 'config.json' by switching the rule 'anvi_gen_contigs_database' parameter '–ignore-internal-stop-codons' to 'true'.

```
anvi-run-workflow -w contigs \
                  -c config.json \
                  --additional-params \
                            --jobs 6 \
                            --resources nodes=6
```

Extract lac operon

We should now have contigs DBs for our genomes.

Now that we have six representative *E. coli* examples, we will use a combination of the 'default-mode' and 'flank-mode' to cut out the genomic neighborhood around the lac operon and then trim the contig to just contain the target operon.

First, we will use 'default-mode'. This requires the user to provide a '–search-term' and '–num-genes' parameter. The '–search-term' will act as an anchor gene to locate the locus within the contigs provided. In this case, we will use the lacZ gene to locate the lac operon in our *E. coli* genomes. Once the anchor gene is located '–num-genes X,Y' will instruct 'anvi-export-locus' to cut 'X' gene(s) upstream and 'Y' gene(s) downstream of the designated anchor gene.

Let's get cutting!

Default mode

First, we'll use 'default-mode' to extract the general genomic neighborhood around the lac operon. To cut the lac operon from a single genome, we would have run this command in this general form:

```
anvi-export-locus -c MY_GENOME.db \
                  --num-genes 10,10 \
                  --search-term "lacZ" \
                  -O MY_GENOME_lac_locus
```

But since we have multiple genomes we wish to study all at once, we will build a 'for' loop in BASH to make life easier (while making everything more reproducible, and less error-prone at the same time):

```
mkdir 03_LOCI
```

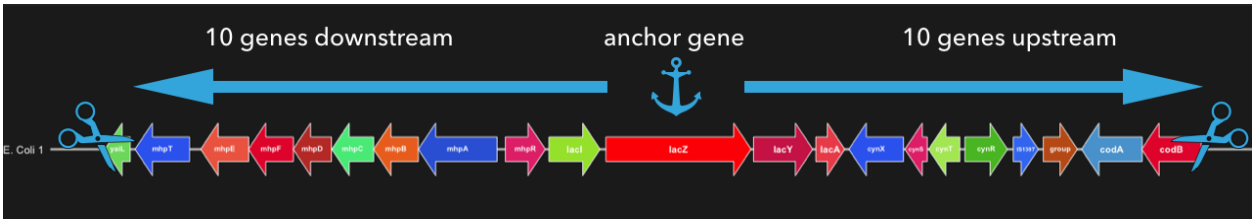

Figure 2.2:

```
cd 03_LOCI

for GENOME in `ls "${WD}"/02_CONTIGS/*-contigs.db`;
do
    FNAME=$(basename "${GENOME}" -contigs.db)
    anvi-export-locus -c "${GENOME}" \
                      --num-genes 10,10 \
                      --search-term "lacZ" \
                      -O "${FNAME}"_lac_locus;
done
```

Be aware that ‘–search-term‘ is NOT case sensitive unless you surround your term in quotes (e.g. ‘–search-term "lacZ"‘)

Here is a visual representation of how ‘anvi-export-locus‘ found the anchor gene "lacZ" the cuts 10 genes upstream and downstream.

Flank-mode

Awesome, now we have some smaller contigs that contain the lac operon. BUT, we also grabbed some extra genes that don’t belong to the operon. Let’s use ‘–flank-mode‘ to trim the loci to just contain the lac operon.

To do this, give ‘anvi-export-locus‘ two flanking search-terms: ‘lacI‘ and ‘lacA‘

```
for GENOME in `ls "${WD}"/03_LOCI/*.db`;
do
```

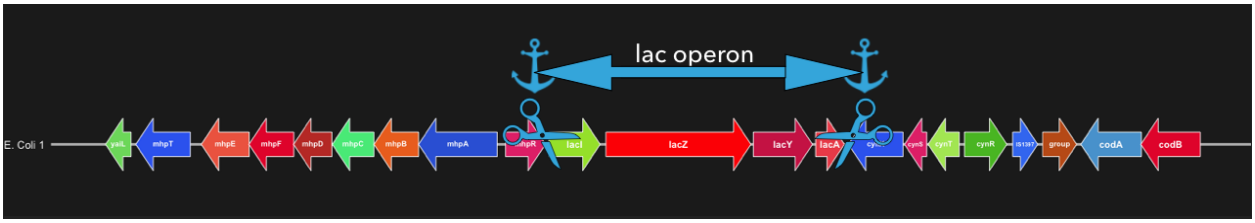

Figure 2.3:

```
    FNAME=$(basename "${GENOME}" _lac_locus_0001.db)
    anvi-export-locus -c "${GENOME}" \
                      --flank-mode \
                      --search-term "lacI","lacA" \
                      -O "${FNAME}"_lac_locus_clean;
done
```

Be aware that '–flank-mode' requires flanking genes to be single copies in the contig it's searching. If your locus of interest does not have fixed coordinates in your genomes or metagenomes, you may need to adjust the '-search-term's on a case-by-case basis.

Here is a visual representation of how 'flank-mode' cuts out a locus using flanking genes:

Conclusion

So there you have it, 'anvi-export-locus' is a flexible tool that allows you to extract genomic loci from genomes and metagenomes. In this tutorial, we looked at the classic lac operon in *E. coli* genomes, but this tool can also be unleashed on any loci in genomes or metagenomic assemblies. For example, one can download all genomes available for taxa with biosynthetic gene clusters, identify an anchor gene, then extract their loci for characterization!

2.5 Reproducible Bioinformatics Workflows in anvi'o

In the era of big data in microbiology, workflow technology catalyzes the ability to swiftly and reproducibly process terabytes of 'omics data at scale (Eren and Delmont, 2024). Snakemake,

workflow infrastructure, has been a game changer for the open-source bioinformatics community (Mölder et al., 2021) and leverages the popular coding language Python. Setting up a computational workflow provides numerous advantages for computational biologists, including (1) processing multiple data types in parallel, drastically decreasing compute time; (2) careful logging of all computational processes so that errors can be efficiently traced and resolved; (3) scaling a workflow to as many files as you need without changing the definition of the workflow; and finally, my favorite (4) restarting a workflow EXACTLY where it left off even if it prematurely stops.

The anvi'o platform integrates seamlessly with Snakemake workflows, reflecting its design philosophy of creating small, task-specific programs that work together to form a cohesive computational ecosystem for processing 'omics data. This compatibility allows Snakemake to organize and link these programs in a modular fashion, enabling the development of powerful and scalable computational pipelines tailored for microbiologists. Notably, the first anvi'o Snakemake workflows were specifically implemented to automate commonly used anvi'o pipelines for metagenomic data analysis, streamlining these processes and enhancing reproducibility and efficiency. The webpage https://anvio.org/help/main/programs/anvi-run-workflow/ serves documentation for the 'main' program from which anvi'o Snakemake workflows are run, and this webpage https://merenlab.org/2018/07/09/anvio-snakemake-workflows/ serves as an overarching tutorial for the most popular workflow in anvi'o.

2.5.1 The anvi'o sra-download workflow

Workflow description

The anvi'o 'sra-download' workflow is a computational workflow built with Snakemake (Mölder et al., 2021) that automates the process of downloading paired-end FASTQ files for a given list of SRA-accessions. Downloading paired-end FASTQ files from NCBI is common practice for

computational microbiologists to acquire publicly available sequencing data from microbiomes with multiple sequencing types, including 16S amplicon sequencing and shotgun metagenomic sequencing. Downloading data requires careful verification to ensure files are complete, uncorrupted, and suitable for downstream analysis. I became acutely aware of this while downloading and processing hundreds of metagenomes for my keystone paper in Chapter 3. Additionally, sequencing data packages can be terabytes in size, requiring days to download. Manually retrieving individual SRA accessions via an HPC terminal is both time-consuming and error-prone. The anvi'o 'sra-download' workflow provides numerous advantages compared to batch downloading or 'for' loops including (1) parallel downloading on high-performance computing clusters to drastically increase the download speed, (2) high-quality logging to carefully document when downloads fail, (3) restarting downloading workflows exactly where they left off.

Excitingly, this workflow has been featured in an anvi'o tutorial, showcasing its efficiency in downloading metagenomes to profile *Prochlorococcus* genomes in the surface ocean and demonstrating how anvi'o analyses can scale to terabytes of data: https://anvio.org/tutorials/scaling-up/. The webpage https://anvio.org/help/main/workflows/sra-download/ serves documentation for this program.

2.5.2 The anvi'o EcoPhylo workflow

The EcoPhylo workflow explores the ecological and phylogenetic relationships between a gene family and the environment. Briefly, the workflow extracts gene families orthologs from any set of FASTA files (e.g., isolate genomes, MAGs, SAGs, or simply assembled metagenomes) using a user-defined HMM, and displays an interactive interface highlighting the phylogeographic distribution of orthologs across environments. The EcoPhylo workflow can leverage any HMM that models amino acid sequences. If the user chooses an HMM for a single-copy core gene, such as ribosomal protein, the workflow will yield multi-domain taxonomic pro-

files of metagenomes *de facto*. Below, I highlight some key features of the workflow but a comprehensive description including command line examples can be found here: `https://anvio.org/help/main/workflows/ecophylo/`

The EcoPhylo workflow starts with a user-provided target gene family defined by an HMM and a list of assembled genomes and/or metagenomes. The final output is an interactive interface that includes (1) a phylogenetic analysis of all genes detected by the HMM in genomes and/or metagenomes and (2) the distribution pattern of each of these genes across metagenomes if the user-provided metagenomic short reads to survey.

The 'user-provided HMM' is passed to EcoPhylo via the hmm-list file, and the input assemblies of genomes and/or metagenomes to query using the HMM are passed to the workflow via the files external-genomes and metagenomes-txt, respectively. Finally, the user can also provide a set of metagenomic short reads via a samples-txt to recover the distribution patterns of genes across samples.

EcoPhylo first identifies homologous genes based on the input HMM, clusters matching sequences based on a user-defined sequence similarity threshold, and finally selects a representative sequence from each cluster that contains more than two genes. The final set of representative genes passes through a series of QC at multiple steps of the workflow, including (1) HMM alignment coverage, (2) open reading frame completeness, and (3) multiple sequence alignment filtering.

HMM alignment coverage filtering: The first step to removing bad hmm-hits is to filter out hits with low-quality alignment coverage. This is done with the rule 'filter_hmm_hits_by_model_coverage', which leverages anvi-script-filter-hmm-hits-table. This tool uses the output of 'hmmsearch' to filter out hits based on the model and/or gene coverage. We recommend 80% model coverage filter for most cases which can confidently identify HMM-hits while leaving room for open-reading frame length variation. However, exploring the distribution of model coverage with any new HMM is always recommended to help you determine a

proper cutoff. To adjust this parameter, go to the `filter_hmm_hits_by_model_coverage` rule and change the parameter `--min-model-coverage`. You can also adjust the gene coverage by change the parameter `--min-gene-coverage`. This can help filter out ORFs with outlier lengths, but its effectiveness entirely depends on the HMM in use. You can explore the distribution of alignment coverages before choosing HMM alignment coverage filtering values with a tool like Interproscan. Additionally, you can parse the domtblout files from hmmsearch to explore the distribution of these values within your own data.

Filtering open-reading frame completion values: Genes predicted from genomes and metagenomes can be partial or complete depending on whether a start and stop codon is detected. Even if you filter out hmm-hits with bad alignment coverage as discussed above, HMMs can still detect low-quality hits with good alignment coverage and homology statistics due to partial genes. Unfortunately, partial genes can lead to spurious phylogenetic branches and/or inflate the number of observed populations or functions in a given set of genomes/metagenomes. To remove partial genes from the EcoPhylo analysis, the user can assign `true` for `--filter-out-partial-gene-calls` parameter so that only complete open-reading frames are processed.

Trimming large multiple sequence alignments: One step of EcoPhylo involves performing a multiple sequence alignment of the recruited orthologs. Depending on the application and size of the input dataset, this process may involve thousands of ORFs, making the MSA particularly challenging. By default, the EcoPhylo is designed for quick insights, and thus the workflow-config file uses MUSCLE parameters to perform a large MSA, swiftly. The EcoPhylo workflow removes sequences that contain more than 50% gaps in the MSA to remove assembly artifacts that could add a lot of gaps to the MSA. Additionally, the workflow filters for columns with more phylogenetic information using trimAL with the '-gappyout' flag (Capella-Gutiérrez et al., 2009).

After quality control steps, the EcoPhylo workflow can continue with one of two modes de-

fined by the user in the workflow-config: tree-mode or profile-mode. In the tree-mode, the user must provide a hmm-list and metagenomes-txt and/or external-genomes, and the workflow will stop after extracting representative sequences and calculating a phylogenetic tree (without any insights into the ecology of sequences through a subsequent step of metagenomic read recruitment. In contrast, the profile-mode will require an additional file: samples-txt. In this mode, the workflow will continue with the profiling of representative sequences via read recruitment across user-provided metagenomes to recover and store coverage statistics. Completing the workflow will yield all files necessary to explore the results in downstream analyses to investigate associations between ecological and evolutionary relationships between target genes.

Future work

The anvi'o EcoPhylo workflow has been successfully applied in several projects to investigate the evolution of respiratory reductases (Little et al., 2024), detect genomes within large metagenomic datasets (Veseli et al., 2024b), and, as showcased in Chapter 3 of this thesis, track genome recovery rates using ribosomal proteins across metagenomic projects. Despite its current utility, the workflow has multiple potential avenues for future development.

One of the strengths of EcoPhylo is its modular design, which allows users to easily swap tools for key steps of the workflow. For example, the clustering algorithm currently relies on MMseqs2 (Steinegger and Soding, 2017), but users could replace it with CD-HIT (Li and Godzik, 2006) with minor code adjustments to the workflow. This flexibility ensures that the workflow can adapt to specific research needs or emerging methodologies.

Another promising direction for EcoPhylo is to expand beyond analyzing individual open reading frames to examine entire loci across genomes and metagenomes. For instance, users could leverage the anvi-export-locus command to extract a defined set of upstream and downstream genes around a target HMM hit. These loci could then be passed through the workflow, potentially enhancing its phylogenetic capabilities. This would be particularly valuable for

studying deep branching patterns of taxa without complete genomes by accessing ribosomal gene operons and performing phylogenetic analyses on multiple genes simultaneously.

Currently, the most significant bottleneck in the EcoPhylo workflow lies in the first three steps, which involve running anvi-run-hmms, filtering results with anvi-script-filter-hmm-hits-table, and extracting sequences using anvi-get-sequences-for-hmm-hits. These steps are handled by separate rules, each invoking a distinct program. Combining the functionality of these three steps into a single program would streamline the workflow and improve efficiency.

Finally, another enhancement would be implementing metrics to assess the diversity explained by categorical variables in the resulting phylogenetic trees. For example, Wu et al. (2025) evaluated genome recovery by measuring the phylogenetic diversity (PD) (Wu et al., 2009) of metagenome-assembled genomes (MAGs). Incorporating such measurements into EcoPhylo would provide users with valuable insights into the evolutionary diversity of gene families captured by their analyses.

CHAPTER 3

RIBOSOMAL PROTEIN PHYLOGEOGRAPHY OFFERS QUANTITATIVE INSIGHTS INTO THE EFFICACY OF GENOME-RESOLVED SURVEYS OF MICROBIAL COMMUNITIES

This section is derived from the following publication:

Matthew S. Schechter, Florian Trigodet, Iva A. Veseli, Samuel E. Miller, Matthew L. Klein, Metehan Sever, Loïs Maignien, Tom O. Delmont, Samuel H. Light, A. Murat Eren. "Ribosomal protein phylogeography offers quantitative insights into the efficacy of genome-resolved surveys of microbial communities." bioRxiv. https://doi.org/10.1101/2025.01.15.633187

3.1 Abstract

The increasing availability of microbial genomes is essential to gain insights into microbial ecology and evolution that can propel biotechnological and biomedical advances. Recent advances in genome recovery have significantly expanded the catalogue of microbial genomes from diverse habitats. However, the ability to explain how well a set of genomes account for the diversity in a given environment remains challenging for individual studies or biome-specific databases. Here we present EcoPhylo, a computational workflow to characterize the phylogeography of any gene family through integrated analyses of genomes and metagenomes, and our application of this approach to ribosomal proteins to quantify phylogeny-aware genome recovery rates across three biomes. Our findings show that genome recovery rates vary widely across taxa and biomes, and that single amplified genomes, metagenome-assembled genomes, and isolate genomes have non-uniform yet quantifiable representation of environmental microbes. EcoPhylo reveals highly resolved, reference-free, multi-domain phylogenies in conjunction with distribution patterns of individual clades across environments, providing a means to assess genome recovery in individual studies and benchmark biome-level genome

collections.

3.2 Introduction

Establishing comprehensive genome catalogues is a fundamental objective in microbiology as genomes are essential to develop insights into microbial life and to advance biotechnology and biomedicine (Eren and Banfield, 2024). Indeed, the rapidly increasing number of microbial genomes (1) provides an evolutionary framework to resolve the branches of the Tree of Life (Brown et al., 2015; Spang et al., 2015), (2) enables hypothesis generation and testing through comparative genomics (Paoli et al., 2022; Al-Shayeb et al., 2022; Durrant et al., 2022; Chen et al., 2024), (3) offers resources to search for novel biosynthetic capabilities and natural products (Paoli et al., 2022; Chen et al., 2024), (4) contributes to the body of nucleotide data used to train biological language models (Cornman et al., 2024; Nguyen et al., 2024; Hwang et al., 2024) and more, while well-structured databases aim to consolidate and give access to the outcomes of genome recovery efforts (Parks et al., 2022; Schmidt et al., 2023).

Increasing availability of microbial genomes is a result of multiple complementary breakthroughs that include (1) advances in high-throughput or targeted cultivation that enable the recovery of isolate genomes (Jiang et al., 2016; Watterson et al., 2020; Cross et al., 2019), (2) the use of environmental shotgun sequencing that enables the recovery of metagenome-assembled genomes (MAGs) (Chen et al., 2020b), and (3) the use of microfluidics and cell sorting that enables the recovery of single amplified genomes (SAGs) (Woyke et al., 2017). These strategies have not only been used in large-scale characterization of many of the Earth-âĂŹs biomes (Pasolli et al., 2019; Parks et al., 2017; Pachiadaki et al., 2019; Ma et al., 2023), but also have been applied to many specific questions or niche systems that span a wide range of research priorities, collectively resulting in over 500,000 non-redundant bacterial and archaeal genomes (Parks et al., 2022). The recovery of microbial genomes is now a relatively well-established practice, yet it is not straightforward to assess (1) how taxonomic or

biome-specific biases impact on genome recovery efforts, and (2) the ecological or evolutionary importance of unrecovered populations. As a result, individual studies that recover genomes, or efforts that curate biome-specific or global genomic collections, rarely offer quantitative insights into one of the key questions they aim to address: "how well do these genomes represent this environment?".

Attempts to benchmark genome recovery often rely upon metagenomic read recruitment statistics to quantify the fraction of reads that map to genomes with the assumption that the proportion of reads recruited by a genomic collection is a proxy for the degree to which a genome collection represents the genomic fragments found in a given environment. In individual studies that reconstruct genomes directly from environmental metagenomes, the proportion of metagenomic reads that are recruited by resulting MAGs can vary from as low as 7% in the surface ocean (Delmont et al., 2018a) to as high as 80% in the human gut (Carter et al., 2023). While read recruitment statistics are easy to generate and communicate, they fail to contextualize what is present in the unmapped fraction and thus leave considerable ambiguity about the microbial community. For instance, a large fraction of metagenomic reads not mapping to the genome catalogue could belong to a single organism or multiple taxonomically diverse microbes with critical ecological roles in the system. Furthermore, genome collections often systematically underrepresent certain portions of the tree of life, as the rate of genome recovery differs across taxa as a function of genome recovery methodology: while cultivation efforts often struggle to capture slow-growing organisms (Imachi et al. 2020) or those that depend on others for survival (He et al., 2015), genome-resolved metagenomics often struggle to reconstruct genomes from taxa that form highly complex populations (Giovannoni, 2017; Pachiadaki et al., 2019). Altogether, biological and non-biological factors confound accurate interpretations of read recruitment results, and the ability to measure genome recovery rates requires alternative strategies that can contextualize the ecological and evolutionary relationships of organisms recovered in genome collections with environmental populations accessible

through metagenomics.

One approach to gaining insight into microbial life underrepresented in genome catalogues Âăinvolves the use of marker genes. *De novo* assembly, in which individual sequencing reads are stitched together to recover much longer contiguous segments of DNA (contigs), is common to the vast majority of genome recovery efforts. While in most cases contigs only represent fragments of genomes, they still explain a much greater genomic context than unassembled reads and give access to entire open reading frames, including phylogenetically informative marker genes. Employing such phylogenetically informative genes assembled from metagenomes in conjunction with metagenomic read recruitment enables fine-grained analyses of phylogeny and biogeography of individual taxa, as demonstrated by previous studies that used the *rpoC1* gene to characterize the phylogeography of marine bacteria (Kent et al., 2019; Ustick et al., 2023) or RNA and DNA polymerases to identify and guide the genomic recovery of major viral clades (Weinheimer and Aylward, 2020; Gaia et al., 2023).

Among all phylogenetically informative genes, ribosomal proteins represent a special class as they (1) occur as a single-copy gene in genomes across the tree of life, (2) are consistently assembled even for complex or relatively rare populations in metagenomes due to their relatively short length, and (3) contain enough phylogenetic information to delineate distinct branches of life at relatively high levels of resolution (Olm et al., 2020). Recognizing their utility, many studies have leveraged individual ribosomal proteins to analyze community composition Wu et al. (2009); Crits-Christoph et al. (2022) , integrating ribosomal protein phylogenies with metagenomic read recruitment to track individual clades of microbes (Hug et al., 2016; Emerson et al., 2016; Hamilton et al., 2016; Diamond et al., 2019; Matheus Carnevali et al., 2021). Ribosomal proteins are thus ideally suited gene markers for tracking microbial populations underrepresented within genome collections.

Here we present EcoPhylo, a workflow to simultaneously visualize the phylogenetic relationships and biogeographical distribution patterns of sequences that match any given gene

family from genomes and metagenomes, and demonstrate its application to the phylogeography of ribosomal proteins for quantification of genome recovery rates across biomes. Our results show that bringing together multi-domain ribosomal protein phylogenies with distribution patterns of individual clades across environments in a single interface offers a valuable data analysis and visualization strategy to benchmark genome recovery efforts scaling from individual projects to global surveys of large genome collections and metagenomes.

3.3 Results

3.3.1 EcoPhylo enables integrated surveys of gene family phylogeography

EcoPhylo implements a computational workflow to integrate the phylogeny and biogeography of any given gene family and enables its users to track the distribution patterns and evolutionary relationships between homologous genes across environments and/or experimental conditions (Figure 4.1, also see Materials and Methods).

When applied to phylogenetically tractable single-copy core genes, such as ribosomal proteins, in tandem with metagenomes and a genome collection, EcoPhylo identifies populations assembled in metagenomes but absent in the genomic collections (and *vice versa*), highlighting the ecological and evolutionary relevance of organisms detected through metagenomic assemblies but lacking genomic representation (Figure 4.1). This allows for the quantification of genome recovery rates of different methods (e.g., isolate genomes, MAGs, SAGs) across taxa and provides a means to investigate phylogenetic and ecological features of organisms without genomic representation. Importantly, the unbroken link between genes and contigs enables downstream targeted binning efforts when necessary.

Using ribosomal proteins to *de novo* characterize the phylogenetic makeup of microbiomes and benchmark genome recovery rates has numerous advantages. However, these advantages also pose noteworthy challenges. Ribosomal proteins are short protein sequences (~300

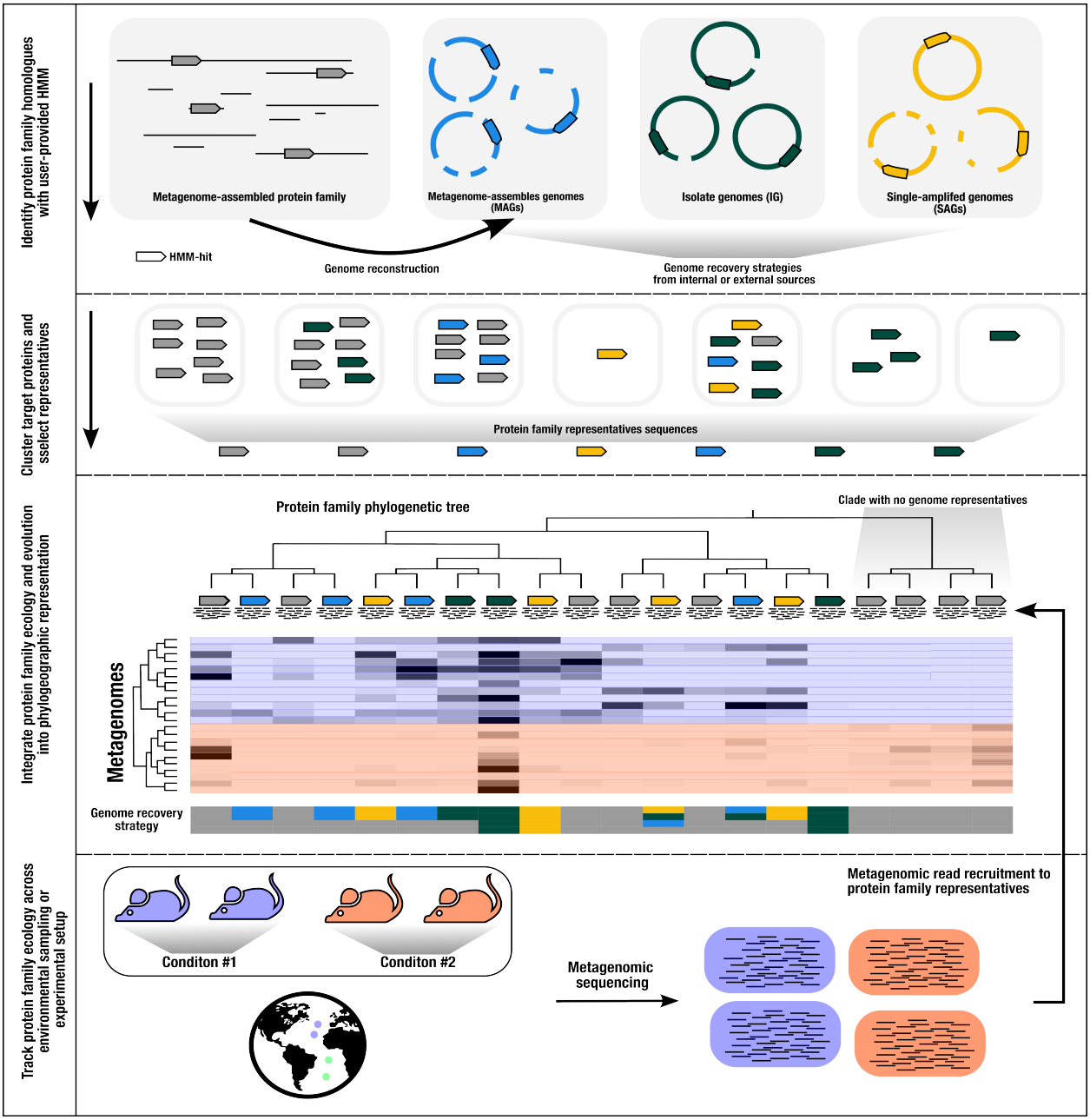

Figure 3.1: **Schematic of the EcoPhylo workflow applied to a Ribosomal protein family.** The proposed workflow integrates biogeography from metagenomic read recruitment and protein phylogenetics to display the phylogeographical distribution of closely related lineages. When including genome sources the workflow highlights which genome recovery strategies are more effective for sampling specific taxa. Although this manuscript focuses on ribosomal proteins, the proposed workflow is generalizable to any protein family.

amino acids), which substantially limits their ability to resolve deep phylogenetic branching patterns. Furthermore, their evolution is subject to strong purifying selection, as a result, the average nucleotide identity (ANI) threshold often used to define 'species' boundaries between whole genomes is 95% (Jain et al., 2018) increases to 99% for ribosomal protein sequences (Olm et al., 2020) . Therefore, ribosomal proteins are more vulnerable than other genes to non-specific read-recruitment from closely related proteins within metagenomes. To identify criteria for reliably resolving taxa, we started our investigation by developing a series of benchmarks to optimize the use of ribosomal proteins in EcoPhylo with appropriate parameters to maximize the ecological and evolutionary signal they can offer while minimizing non-specific read recruitment. These benchmarks, which are detailed in the Supplementary information (1) inspected hidden Markov model (HMM) alignment coverage thresholds to accurately detect ribosomal proteins in genomes and metagenomes; (2) examined the copy number distribution of ribosomal protein HMMs across archaeal and bacterial genomes to only consider single-copy candidates; and (3) explored nucleotide similarity thresholds to cluster ribosomal gene sequences to maximize the taxonomic resolution of representative sequences while maintaining sufficient nucleotide distance between distinct representative sequences to reduce non-specific read recruitment from metagenomes (Supplementary information).

Based on these considerations, we implemented routines and adjusted default EcoPhylo parameters to (1) use a minimum of 80% model coverage for ribosomal protein HMMs for a match; (2) filter for complete open reading frame sequences to remove assembly artifacts; and (3) cluster HMM hits with target coverage to ensure grouping of extended open reading frames and leverage 97% nucleotide similarity as the most appropriate clustering threshold to minimize non-specific read recruitment (Supplementary information). We also compared broad ecological insights recovered from EcoPhylo to state-of-the-art taxonomic profiling tools, confirming that this framework offered qualitatively comparable results (Supplementary information). Altogether, these evaluation and optimization steps yielded EcoPhylo default parameters

to obtain representative ribosomal protein sequences that are suitable for investigations of the phylogeny, biogeography, and genome recovery of populations they describe.

3.3.2 Ribosomal proteins quantify and contextualize genome recovery rates from metagenomes

Thanks to its diverse physiological properties that promote a variety of chemical gradients and surfaces (Bowen et al., 2018), the human oral cavity is home to diverse communities of microbes (Dewhirst et al., 2010). The human oral microbiome is a relatively well-characterized environment with a wealth of isolate genomes accessible through the Human Oral Microbiome Database (HOMD) (Escapa et al., 2018; Chen et al., 2010), and numerous genome-resolved metagenomics surveys that have captured representative genomes of microbial clades that have largely eluded cultivation efforts. Using EcoPhylo we first focused on a genome-resolved metagenomics survey which reconstructed multiple high-quality MAGs from tongue and plaque samples from the human oral cavity (Shaiber et al., 2020). While Shaiber et al. (2020) reported numerous genomes for elusive taxa, such as *Saccharimonadia* (TM7), *Absconditabacteria* (SR1), and *Gracilibacteria* (GN02), the genome-resolved metagenomic workflow failed to reconstruct MAGs that resolved to some of the best-represented organisms in culture collections from the oral cavity, such as members of the genus *Streptococcus*, (Escapa et al., 2018), which was represented by only two MAGs in Shaiber et al. (2020). This discrepancy compelled us to combine isolate genomes from the HOMD together with metagenomes and MAGs from Shaiber et al. (2020), to investigate whether EcoPhylo could reveal the differential recovery of genomes through distinct recovery approaches.

We started our analysis by combining 790 non-redundant MAGs and 14 metagenomic co-assemblies of tongue and plaque metagenomes reported by Shaiber et al. (2020) with 8,615 isolate genomes we obtained from the HOMD (Supplementary Table 1). To characterize these data, we elected to use EcoPhylo with *rpL19* HMM, since it was the most frequent riboso-

mal protein with an average length of 393 nucleotides across all genomes in our collection, occurring in 98.59% of the HOMD genomes and 81.81% of the Shaiber et al. (2020) (Supplementary information). To assess the generalizability of observations made from *rpL19*, we also ran EcoPhylo on the same dataset with *rpS15* and *rpS2*, with the average length of 275 and 781 nucleotides, respectively (Supplementary Table 2, Supplementary information).

The EcoPhylo analysis of the *rpL19* genes found in the genomes and metagenomic assemblies resulted in a phylogenetic tree with 277 non-redundant bacterial representative sequences (Figure 3.2, Supplementary Table 3). Hierarchical clustering of metagenomes based on the detection patterns of these *rpL19* sequences organized metagenomes into tongue and plaque sampling sites *de novo* (Figure 3.2, Supplementary Figure 1), demonstrating that a single ribosomal gene family is able to capture the known ecological differences between these habitats. Many closely related *rpL19* genes that resolved to prevalent oral taxa, such as *Prevotella* and *Steptococcus*, showed within-genus differences in site specificity, a previously observed phenomenon (Eren et al., 2014) that is attributed to divergent accessory genomes (Mark Welch et al., 2019; Utter et al., 2020). Multiple ribosomal protein representative sequences recruited reads from tongue as well as plaque metagenomes, also matching prior observations of cosmopolitan taxa (Figure 3.2, Supplementary information). Overall, the ecological insights revealed by *rpL19* recapitulated known ecology of oral microbes (Mark Welch et al., 2019) and provided a framework to assess genome recovery rates.

EcoPhylo tracks the origins of each sequence in each sequence cluster. Some *rpL19* clusters, representatives of which are shown in the phylogenetic tree in Figure 3.2A, were composed of sequences found only in metagenomic assemblies and not in MAGs or isolate genomes, highlighting clades present in the environment but not in genome collections. Other *rpL19* clusters only contained sequences represented in HOMD isolate genomes; despite their consistent detection in oral samples through metagenomic read recruitment, they were absent in metagenomic assemblies or MAGs, highlighting clades that are less accessible to short-

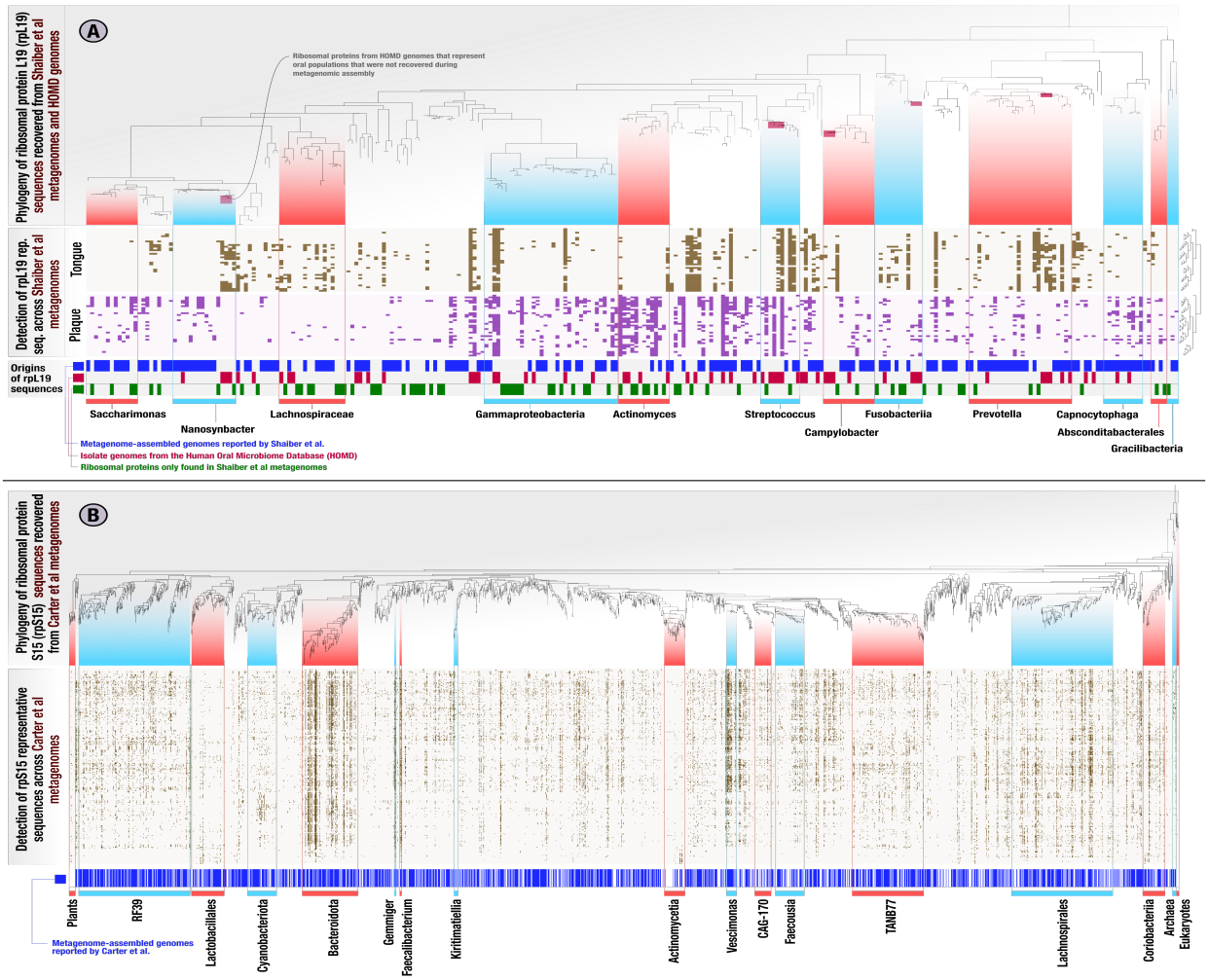

Figure 3.2: **Ribosomal protein phylogeny and detection patterns across metagenomes from the human oral cavity and gut microbiomes.** In the heatmaps in both panels, each column represents a ribosomal protein representative sequence, each row represents a metagenome, and each data point indicates whether a given ribosomal protein was detected in a given metagenome. The columns of the heatmaps are ordered by a tree representing a phylogenetic analysis of all ribosomal protein representative sequences, and the rows are ordered by a hierarchical clustering dendrogram calculated based on the ribosomal protein detection patterns across metagenomes. Panel (A) represents the EcoPhylo analysis of rpL19 sequences across Shaiber et al. (2020) metagenome-assembled genomes (MAGs), Shaiber et al. (2020) oral metagenomes, and HOMD genomes, and includes three additional rows that indicate the origin of a given ribosomal protein: whether it is a metagenome-assembled genome (MAG, blue), HOMD isolate genome (red), or only recovered from metagenomic assemblies with no representation in genomes (green). Smaller red boxes in the phylogenetic tree mark microbial clades absent in the collection of MAGs and assemblies reported by Shaiber et al. (2020) but detected in Shaiber et al. (2020) metagenomes solely due to the inclusion of HOMD isolate genomes.

Figure 3.2 continued: The panel (B) represents the EcoPhylo analysis of rpS15 sequences across the Carter et al. (2023) metagenome-assembled genomes (MAGs) and Carter et al. (2023) gut metagenomes from a Hadza tribe, and includes an additional row that indicates whether a MAG was reported for a given ribosomal protein (blue).

read metagenomic assembly approaches (Figure 3.2). To calculate genome recovery rates for any given taxon, we divided the number of sequence clusters that contained a sequence from a given genome recovery method by the total number of representative sequences EcoPhylo reported for that taxon (Materials and Methods). This analysis revealed that 60.3% of the bacterial populations defined by *rpL19* gene clusters that were detected in metagenomic reads also appeared in MAGs. In other words, the overall bacterial MAG recovery rate in the study by Shaiber et al. (2020) was 60.3% (*rpS15*:Âă 62.8%, *rpS2*:Âă 53.2%) (Figure 3.2A, Supplementary Table 3, Supplementary Table 4). However, this rate of recovery was not uniform across individual taxa. EcoPhylo revealed higher MAG recovery rates for taxa such as *Saccharimonas* at 69.2% (*rpS15*: N/A, *rpS2:* 63.6%), and *Prevotella* at 76.9% (*rpS15*: 82.6%, *rpS2*: 84%). In contrast, the MAG recovery was lower for populations in other clades, including *Gammaproteobacteria* and *Fusobacteriia*, with MAG recovery rates of 47.1% (*rpS15*: 58.1%, *rpS2*: 44.8%) and 41.7% (*rpS15*: 46.2%, *rpS2*: 31.2%), respectively (Figure 3.2A, Supplementary Table 3, Supplementary Table 4). The MAG recovery rate was particularly low for *Streptococcus* at 30% (*rpS15*: 15.4%, *rpS2*: 10%), consistent with the presence of only two MAGs in Shaiber et al. (2020). However, the MAG recovery rate for *Actinomyces* was also very low at 23.1% (*rpS15*: 36.4%, *rpS2*: 13.3%) despite the characterization of nine *Actinomyces* MAGs by Shaiber et al. (2020) reveals a large number of distinct *Actinomyces* populations missed by MAGs even though they were present in the assemblies (Figure 3.2A, Supplementary Table 3, Supplementary Table 4). Overall, this analysis not only confirmed that MAG recovery rates are not uniform across microbial clades, but also showed that quantification of these rates is possible and may yield unexpected insights into the extent of diversity that is not represented in the final set of MAGs for some clades.

The inclusion of genomes from the HOMD increased the number of *rpL19* sequence clusters that contained genomes in this dataset, i.e., the total genome recovery rate, from 60.3% to 73.3% (*rpS15*: 74.8%, *rpS2*: 81.3%), and led to the representation of 35 additional microbial clades for which the metagenomic sequencing and analysis workflow implemented in Shaiber et al. (2020) did not assemble. As with MAGs, the improved detection of taxa among HOMD genomes was not uniform across clades (Figure 3.2A). For example, HOMD genomes offered genomic context for five additional *Streptococcus* populations, increasing the genome recovery rate from 30% with MAGs only, up to 80% when including the HOMD collection. When taking into account both MAGs and isolate genomes, the overall genome recovery rate of Shaiber et al. (2020) from a human oral microbiome dataset determined by EcoPhylo was 73.3%, showing that ribosomal protein phylogeography is an effective means to quantify genome recovery statistics for individual studies. Conversely, EcoPhylo results showed that 26.7% of the individual clades that could be detected through the presence of *rpL19* sequences in assemblies of Shaiber et al. (2020) metagenomes lacked genomic representation in both Shaiber et al. (2020) MAGs and HOMD isolates (Figure 3.2A). Clades that were solely detected through their assembled yet not binned ribosomal proteins increased the detection of populations of *Lachnospiraceae*, *Actinomyces*, *Gammaproteobacteria*, and *Patescibacteria* (Figure 3.2A). As EcoPhylo clusters ribosomal proteins at 97% nucleotide similarity, a conservative threshold that underestimates biodiversity by often grouping genomes with gANI below 95% (Olm et al., 2020).

Next, we applied EcoPhylo to another genome-resolved metagenomics study that recently characterized the gut microbiome of a Hadza hunter-gatherer tribe with a deep sequencing effort by Carter et al. (2023), in which the authors reported nearly 50,000 redundant bacterial and archaeal MAGs from 338 metagenomes with an average of 76 million paired-end reads (Supplementary Table 1). EcoPhylo analysis of this dataset with *rpS15* with an average length of 276 nucleotides, along with *rpS16* and *rpL19*, with the average length of 297 and 370

nucleotides respectively (Supplementary Figure 3.2), revealed a relatively high bacterial MAG recovery rate of 67.7% (*rpS16*: 72.8%, *rpL19*: 69.5%) (Figure 3.2B, Supplementary Table 2). While there were some clades, such as *Actinomycetia*, for which the genome recovery rate was as low as 31.9% (*rpS16*: 32.4%, *rpL19*: 33.3%), the high MAG recovery rate was generally uniform across all major taxa (Supplementary Table 5, Supplementary Table 6, Supplementary Information).

Through these analyses, we are able to demonstrate that the MAGs obtained by Carter et al. (2023) more comprehensively represents the populations captured by their metagenomic assemblies of the human gut compared to the MAGs obtained by Shaiber et al. (2020) given their metagenomic assemblies of the oral cavity (Figure 3.2B, Supplementary Table 5, Supplementary Table 6, Supplementary Information). The ability to make such a statement highlights the utility of EcoPhylo at providing quantitative insights into the efficacy of genome-resolved surveys independent of biomes while offering a phylogenetic and biogeographical context for the populations that were detected in the assemblies.

Overall, EcoPhylo results from the human oral cavity and human gut ecosystems show that our workflow can scale to large metagenomic surveys, combine genomes from multiple sources to compare distinct recovery strategies at the level of individual phylogenetic clades, and recapitulate known ecological patterns.

3.3.3 Genome collections represent a small fraction of microbial diversity in the global surface ocean microbiome

Marine systems support fundamental biogeochemical cycles that maintain the Earth's habitability, and comprehensively documenting the genomes of marine microbes that are intimately connected to these processes has been one of the key aims of microbiology. In addition to decades of cultivation efforts, recent years witnessed a rapid expansion of marine microbial genome catalogs for bacteria and archaea with new MAGs (Delmont and Eren, 2018; Tully

et al., 2018; Paoli et al., 2022) and SAGs (Pachiadaki et al., 2019; Martinez-Perez et al., 2022). Studies that recover genomes from marine systems recognize that the extent to which these collections represent marine environmental populations is limited (Delmont et al., 2019; Paoli et al., 2022). Yet, quantifying the extent of representation at the level of individual environmental clades across genome collections is a challenge. Having established the utility of EcoPhylo to elicit quantitative answers to such questions, we next surveyed a state-of-the-art globally distributed collection of microbial genomes from marine systems (Paoli et al., 2022) in the context of metagenomes generated by the Tara Oceans Project (Salazar et al., 2019; Sunagawa et al., 2015), the Hawaii Ocean Time-series (HOT) (Biller et al., 2018), the Bermuda Atlantic Time-series (BATS) (Biller et al., 2018), BioGEOTRACES expeditions (Biller et al., 2018), and the Malaspina Project (Sanchez et al., 2024) to simultaneously compare genome recovery rates of MAGs, SAGs, and isolate genomes. Of all 1,038 metagenomes, we focused on those that were collected from up to 30m depth and had a size fraction of 0.22 to 3 micrometers (Supplementary Figure 3, Supplementary Table 1), which left us with a total of 237 metagenomes containing a total of 18,832,767,852 short reads (79,463,155 reads per metagenome on average). Our collection of genomes included 7,282 MAGs, 1,474 SAGs, and 1,723 isolate genomes from The Ocean Microbiomics Database (subsetted from samples of < 30m depth when possible) (Paoli et al., 2022). We expanded this collection with an additional 52 isolate genomes that historically have low MAG recovery rates, such as *Pelagibacterales* (SAR11) and *Cyanobacteriota*, and a collection of 41 SAGs obtained from below the Ross Ice Shelf to improve detection of cold-adapted clades Martinez-Perez et al. (2022) (Materials and Methods), yielding a total of 10,479 genomes (Supplementary Table 1). For characterization of these data by EcoPhylo we primarily used the ribosomal gene *rpL14* with an average length of 363 nucleotides, which we detected in 82% of the final list of genomes (Supplementary information), but we also conducted additional analyses using the ribosomal genes *rpS8* and *rpS11*, with an average length of 398 and 415 nucleotides respectively, to confirm our key

observations (Supplementary Figure 1, Supplementary Table 2).

EcoPhylo analysis of *rpL14* genes across 236 global surface ocean metagenomes characterized 8,075 bacterial, 370 archaeal, and 33 eukaryotic clades and computed their distribution patterns across environments (Supplementary Table 7). Hierarchical clustering of metagenomes based on *rpL14* detection patterns split samples into two major groups, whereby one of the groups represented samples collected from polar regions and the other represented samples collected from temperate oceans (Figure 3.3, Supplementary information); a result that is in line with previous observations that documented water temperature as a major driver of microbial diversity in the surface ocean (Sul et al., 2013; Sunagawa et al., 2015). Notably, temperate and polar water samples did not partition when we included metagenomes with lower sequencing depths in our analysis. This was likely caused by increasing noise in detection patterns of various prevalent populations, which compelled us to only consider metagenomes with 50 million or more paired-end reads for our downstream analyses (Supplementary information), which left us with a total of 100 metagenomes (Supplementary Table 1, Supplementary information). Overall, EcoPhylo captured (1) differential distribution patterns among closely related taxa as a function of temperature and latitude, a form of phylogenetic overdispersion likely due to greater competitive exclusion among closely related organisms in the same ecological niche, and (2) showed that the majority of taxa contained both warm- and cold-adapted clades that exclusively occurred either in polar or temperate waters, an expected observation since marine thermal adaptation is not correlated with phylogenetic signal (Thomas et al., 2012) and is likely acquired through independent processes within each major clade (Figure 3.3). Furthermore, the *rpL14* phylogeography captured well-understood biogeographical patterns of prevalent pelagic taxa (Figure 3.3), in agreement with previous studies that showed the dominance of SAR11 subclade Ia.3.V in temperate waters (Delmont et al., 2019) and the exclusivity of cold-adapted SAR11 clades Ia.1 and Ia.3.II to polar regions (Brown et al., 2012; Delmont et al., 2019). It also corroborated the global distribution

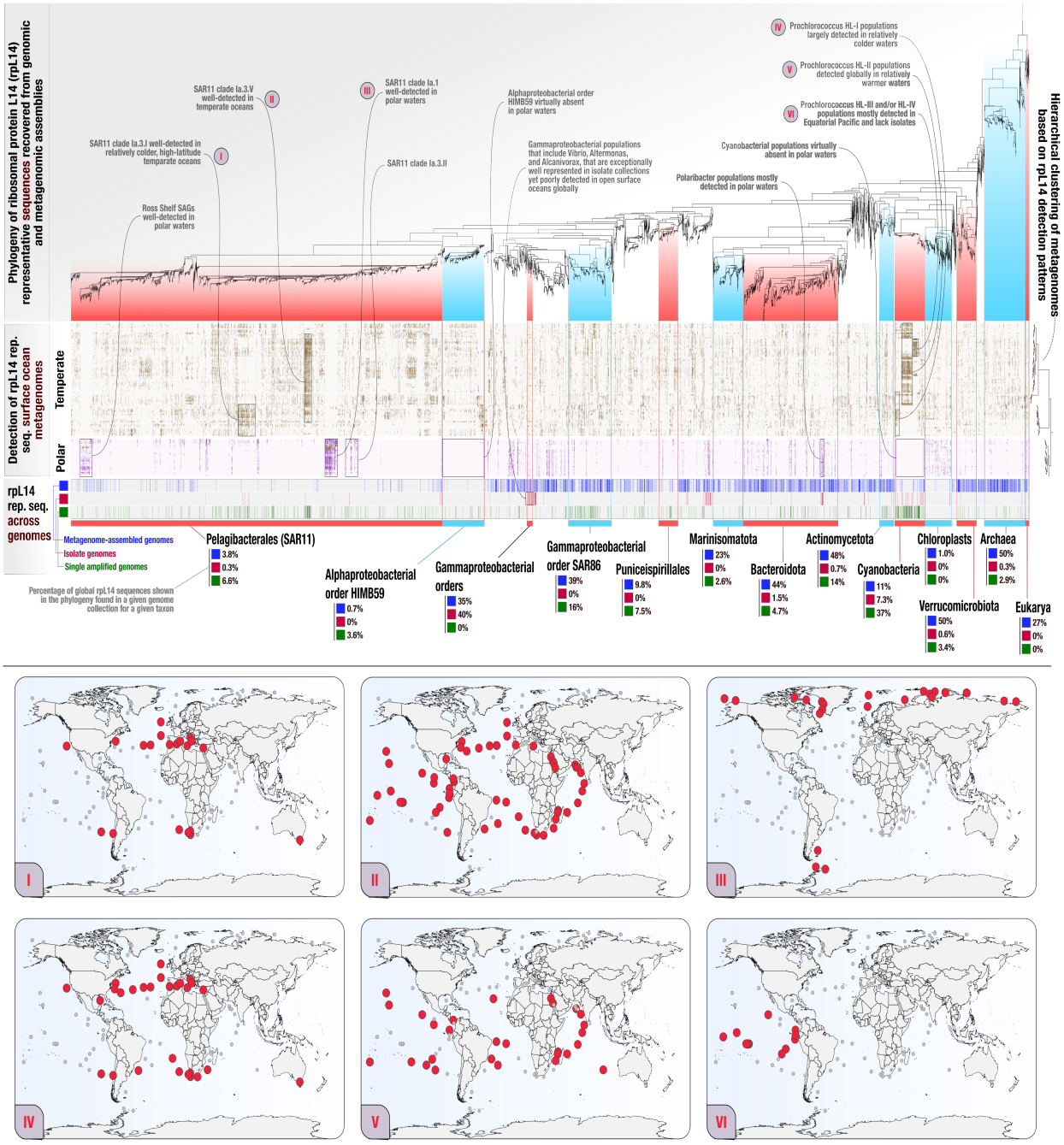

Figure 3.3: **Ribosomal protein L14 phylogeny and detection patterns across metagenomes from the global surface ocean (depth < 30 m).** In the heatmap of panel (A), each column represents a ribosomal protein representative sequence, each row represents a metagenome, and each data point indicates whether a given ribosomal protein was detected in a given metagenome.

Figure 3.3 continued: The heatmap columns are ordered by a tree which represents a phylogenetic analysis of all ribosomal protein representative sequences, and the rows are ordered by a hierarchical clustering dendrogram that is calculated based on ribosomal protein detection patterns across metagenomes. Metagenomes are colored by temperate (gold) or polar (purple) biomes. Each leaf of the phylogenetic tree is decorated below the heatmap with metadata denoting the origin of the RP: metagenome-assembled genome (MAG) (blue), isolate genomes (red), and single amplified genomes (green). In panel (B), each map corresponds to phylogeographical patterns highlighted in panel (A). Colored sampling points correspond to the boxed phylogeographic signals in panel (A).

of *Prochlorococcus* HL-II in temperate waters (Johnson et al., 2006; Biller et al., 2018; Ustick et al., 2023) and the contrasting distribution of this group with *Prochlorococcus* HL-III and *Prochlorococcus* HL-IV, which are mainly found in the Equatorial PacificÂă(Rusch et al., 2010; Huang et al., 2012; Malmstrom et al., 2013; Kent et al., 2016), as well as *Prochlorococcus* HL-I, which is confined to higher latitudes (Johnson et al., 2006; Biller et al., 2015; Delmont and Eren, 2018). The concordance of our results from EcoPhylo with known ecological patterns in marine microbiology underscores the reliability of ribosomal protein phylogeography in characterizing the interplay between microbial ecology and evolution in global surface ocean microbiome (Figure 3.3, Supplementary information).

Using these data, we first compared the overlap between surface ocean microbial populations in the environment and publicly available MAGs generated from this biome by calculating genome recovery rates. The MAG recovery rate for Archaea was relatively high at 49.5% (*rpS8:* N/A, *rpS11:* 50%), however, the MAG recovery rate for Bacteria was only 19.9% (*rpS8*: 22.7%, *rpS11*: 22.2%). In contrast to the MAG recovery rates we observed in individual studies from the human oral and gut microbiome (60.3% and 67.7%, respectively), this much lower recovery rate from multiple sequencing projects reflects the relatively poor efficiency and the contemporary challenges of reconstructing genomes from metagenomes in the ocean biome (Figure 3.3, Supplementary Figure 1, Supplementary Table 7, Supplementary Table 8). Some phyla had relatively high MAG recovery rates, such as 48.2% for *Actinomycetota* (*rpS8*: 50.8%, *rpS11*: 48.9%), 43.9% for *Bacteroidota* (*rpS8*: 47.3%, *rpS11*: 44.3%), and 49.7% for *Verru-*

comicrobiota (*rpS8*: 51.2%, *rpS11*: 44.4%). MAG recovery rates were much lower for other clades, including those containing some of the best-studied autotrophs and heterotrophs of open surface oceans, for example, the MAG recovery rate was only 11% for *Cyanobacteriota* (*rpS8*: 11.7%, *rpS11*: 9.93%). Many clades of *Alphaproteobacteria* had some of the lowest MAG recovery rates, including 12.7% (*rpS8*: 9.21%, *rpS11*: 7.2%) for the uncharacterized order HIMB59 (Supplementary Table 7, Supplementary Table 8). Poor MAG recovery rate was also true for the order *Pelagibacterales*, which remained at 3.76% (*rpS8*: 3.98%, *rpS11*: 4.25%) (Supplementary Table 7, Supplementary Table 8). Compared to taxonomic classification of shotgun metagenomic reads or sequencing of 16S rRNA gene amplicons, prior studies observed much lower relative abundance estimates for populations resolving to *Cyanobacteriota* and *Pelagibacterales* based on MAGs (Pachiadaki et al., 2019; Chang et al., 2024). By elucidating clade-specific discrepancies between different methods of genome recovery, EcoPhylo offers a context for the extent of missing MAGs in prior surveys, which likely is a by product of fragmented metagenomic assemblies due to co-occurring closely related populations with high genomic diversity (Chen et al., 2020a).

De novo characterization of *rpL14* sequences with EcoPhylo uncovers the vast diversity within *Pelagibacterales* compared to the other clades (Figure 3.3). Strikingly, even with the conservative profiling of EcoPhylo that will occasionally pull together ribosomal proteins that belong to genomes from multiple 95% gANI clusters, *Pelagibacterales* made up 41.54% of the non-redundant *rpL14* sequence clusters shown in Figure 3.3, revealing yet another representation of its immense phylogenetic diversity (Morris et al., 2002; Brown et al., 2012; Pachiadaki et al., 2019). While both *Cyanobacteriota* and *Pelagibacterales* suffer from similar rates of poor representation in MAG collections, the missing genomes for environmental populations of *Pelagibacterales* resolved to $\sim$12 times more *rpL14* sequence clusters, which unveils the enormous uncharacterized genomic diversity within this order of many clades that show distinct biogeographical patterns (Figure 3.3), and highlights the importance of ongoing cultivation

efforts to improve its genomic representation (Freel et al., 2024).

3.3.4 Different genome recovery methods come with different clade-specific biases

Finally, we explored the contribution of isolate genomes and SAGs to the genomic representation of surface ocean microbial populations. Isolate genomes had low phylogenetic breadth across the Ribosomal L14 phylogeny and only sampled a few closely related populations, indicating the repeated isolation of similar microbes. In fact, at the phylum level, isolate genomes only effectively sampled *Cyanobacteriota* at a recovery rate of 7.33% (*rpS8*: 10.7%, *rpS11*: 7.80%) despite the fact that we supplemented this clade with extra isolate genomes for this analysis (Supplementary Table 7, Supplementary Table 8). Interestingly, a few closely related orders within class *Gammaprotebacteria* were exceptionally well-covered by bacterial organisms in culture, where 40% of the ribosomal proteins matched to an isolate genome (Figure 3.3). These sister clades represented a relatively small fraction of the overall phylogenetic diversity and were poorly detected across the global surface ocean metagenomes, however, they collectively contained many intensely studied marine model bacterial genera, such as *Vibrio* (Kauffman et al., 2018; Baker-Austin et al., 2017; van Kessel and Camilli, 2024; Septer and Visick, 2024), *Alteromonas* (Pedler et al., 2014; Manck et al., 2022; HenrÃ■quez-Castillo et al., 2022; Lu et al., 2024; Halloran et al., 2025), and *Alcanivorax* (Sabirova et al., 2008; Naether et al., 2013; Prasad et al., 2019, 2023). The clades with some of the highest SAG recovery rates included the order *SAR86* at 16.3% (*rpS8*: 20.1%, *rpS11*: 17.8%) and the phylum *Actinomycetota* at 14.4% (*rpS8*: 16.2%, *rpS11*: 14.5%) (Figure 3.3, Supplementary Table 7, Supplementary Table 8). Furthermore, SAGs augmented the recovery of genomes from prevalent, taxonomically diverse populations with low MAG recovery rates including (1) *Cyanobacteriota* with a three-fold increase compared to MAGs at 37.0% (*rpS8*: 47.2%, *rpS11*: 41.5%), (2) *SAR86* at 16.3% (*rpS8*: 20.1%, *rpS11*: 17.8%), and (3) *Alphaproteobacteria*, in-

cluding *HIMB59* at 5.06% (*rpS8*: 5.61%, *rpS11*: 5.29%) and *SAR11* at 6.35% (*rpS8*: 7.56%, *rpS11*: 7.82%) (Figure 3.3, Supplementary Table 7, Supplementary Table 8). Specifically, SAGs were able to effectively sample the warm-adapted SAR11 clade 1.a.3V as well as an uncharacterized cold-adapted clade of SAR11 likely due to SAG sampling sites that covered both temperate (Pachiadaki et al., 2019) and polar (Martinez-Perez et al., 2022) oceans. Considering that SAR11 has been estimated to be 25% of all plankton (Giovannoni, 2017)and 20-40% of cells counts in the surface ocean (Schattenhofer et al., 2009), SAG methodology, which separates individual bacterial cells from the environment, appears to be optimal for recovering this taxon and avoids the pitfalls of fragmented metagenomic assembly caused by microbiomes with closely related populations (Hosokawa and Nishikawa, 2024). SAGs had greater breadth than MAGs and isolate genomes across the EcoPhylo phylogenies of Ribosomal L14, Ribosomal S11, and Ribosomal S8 (Figure 3.3, Supplementary Figure 1), despite being sampled from only 9 surface ocean sampling sites (Martinez-Perez et al., 2022; Pachiadaki et al., 2019) compared to 237 metagenomes encompassing higher environmental diversity in the global surface ocean. Furthermore, SAGs only represented 6.91% of Bacteria and 2.97% of Archaea populations detected across the dataset while MAGs represented 19.9% of Bacteria and 49.5% of Archaea populations detected across the dataset (Supplementary Table 7, Supplementary Table 8). These results are in line with prior observations that showed pelagic SAGs represent notably more taxonomic richness when compared to MAGs (Pachiadaki et al., 2019).

Similar to the human oral microbiome, we found an uneven phylogenetic distribution of genome recovery rates among genome acquisition strategies in the global surface ocean microbiome. MAGs systematically undersampled globally prevalent clades of *Alphaproteobacteria*, such as SAR11. In contrast SAGs from only a few surface ocean sampling sites (n=9), substantially improved their recovery rates from these clades (Figure 3.3), indicating that while sequencing and assembly of single-cell genomes often lead to severely incomplete genomes

due to amplification biases (Stepanauskas et al., 2017), SAGs show great potential for unbiased genome recovery. The phylogeography of ribosomal proteins adds further evidence that a combination of genome-resolved metagenomics, single amplified genomics, and innovations in microbial isolation strategies are needed to further increase genomic representation of diverse taxa in the global surface ocean.

3.4 Discussion

Our work illuminates the efficiencies of current genome recovery methods and their ability to sample genomes from various microbiomes. By leveraging phylogenetically informative marker genes detected in metagenomic assemblies, such as ribosomal proteins, that are absent from final genome collections, EcoPhylo provides a robust framework for benchmarking genome recovery rates across multiple genome acquisition methods and contextualizing the ecological and evolutionary of genome collections with naturally occurring microbial populations. Our study examined three microbiome projects that used multiple genome recovery strategies (MAGs, SAGs, and isolate genomes) to survey the human oral cavity, global surface ocean, and human gut. Overall, we found that the EcoPhylo workflow can quantitatively measure genome recovery rates and analyze heterogeneous genome collections to assess the efficacy of distinct recovery methods at the level of individual phylogenetic clades. We observed that deep metagenomic sequencing of the human gut microbiome yielded the highest genome recovery rate across these three biomes analyzed. Additionally, we identified that a state-of-the-art genome collection from marine environments represents a small fraction of the total diversity in the open surface ocean through the lens of ribosomal proteins found in assembled metagenomes. By generating insights into multi-domain ribosomal protein phylogeography, EcoPhylo provides a valuable interactive data visualization strategy to evaluate the underlying microbial ecology of metagenomic sequencing projects.

The *de novo* profiling of ribosomal proteins in metagenomic assemblies resembles reference-

based taxonomic profiling of metagenomic short reads to predict relative abundances of taxa, an idea that is implemented in multiple tools that use marker genes, such as Kraken (Wood and Salzberg, 2014), MIDAS (Nayfach et al., 2016), Bracken (Lu et al., 2017), mOTUs (Ruscheweyh et al., 2022), and MetaPhlAn (Manghi et al., 2023), or processed conserved marker gene windows, such as SingleM (Woodcroft et al., 2024). As these tools typically report distinct taxa and their relative abundances, they indeed can help assess genome recovery efforts through direct comparisons of taxon names they identify to the taxonomy of recovered genomes. However, the requirement of a database of reference genomes and/or marker genes, and the absence of a direct link between the genes in assemblies and taxon names reported in tables limit applications with additional downstream opportunities such as targeted genome recovery. In contrast, the flexibility of surveying any marker gene, including ribosomal proteins, across user-provided metagenomic assemblies *de novo* offers an alternative approach that directly connects genes of unrecovered taxa to assemblies and estimates the number of populations detected in metagenomes regardless of their phylogenetic novelty across diverse samples and conditions.

While the phylogeography of ribosomal proteins offers valuable insights into genome recovery, these genes have notable limitations. Rates of evolution as well as the likelihood to be recovered through metagenomic assembly will differ across ribosomal protein families, complicating direct quantitative comparisons between different ribosomal proteins and in some cases will require surveying multiple ribosomal proteins to ensure the generalizability of observations from a single ribosomal protein. Additionally, individual ribosomal protein trees will have less phylogenetic power compared to concatenated ribosomal protein trees or longer marker genes. Although this may lead to suboptimal organism phylogenetics, the efficient organization of ribosomal proteins yields informative insights into the diversity of clades within a sample. Furthermore, when working with incomplete genomes, such as MAGs or SAGs, a single ribosomal gene family will rarely be detected across the entire genome collection and

thus only a subset of genomes will be contextualized per protein. Yet the inherent trade-offs of using incomplete genomes (x âĽě 50% and less than 10% contamination) highlight ongoing challenges in genome recovery, as stricter completeness thresholds would further reduce the number of genomes available for analysis.

The modular design and customizable parameters of EcoPhylo allows users to go beyond ribosomal proteins and leverage other gene families tailored for specific analyses which can improve phylogenetics and the detection of specific taxa. For example, RNApolA and RNApolB have been leveraged for phylogeny-guided binning leading to the discovery of missing branches in viral evolution (Gaia et al., 2023). Furthermore, phylogeography of functional protein families can be leveraged as proxies for microbial metabolism, e.g. phylogeography of ABC transporters can aid in modeling cryptic fluxes of microbial metabolites (Schroer, 2023). The EcoPhylo workflow provides a platform for future microbiome projects to benchmark their genome recovery rates upon release of genome collections. Ribosomal protein phylogeography in tandem with reporting read recruitment percentages to representative genome collections, provides comprehensive insights into genome recovery rates given the biodiversity detected in metagenomes. Future studies can leverage the strategy implemented in EcoPhylo to reanalyze existing metagenomic assemblies to identify missing clades or develop tailored methods to optimize overall genome recovery efforts by taking advantage of the increasing availability of genomes and metagenomes.

3.5 Materials and Methods

3.5.1 The EcoPhylo workflow

EcoPhylo is a computational workflow implemented in the open-source software ecosystem anvi'o (Eren et al., 2015, 2021) using the Python programming language and the workflow management system, Snakemake (Koster and Rahmann, 2012). The primary purpose of

EcoPhylo is to offer an integrated means to study phylogenetic relationships and ecological distribution patterns of sequences that match to any gene family based on user-provided hidden Markov model (HMM) searches from genomic and metagenomic assemblies. A minimal command line instruction to start an EcoPhylo run is ‘anvi-run-workflow -w ecophylo -c config.json‘, where ‘anvi-run-workflow‘ is a program in anvi’o that runs various workflows, and ‘config.json‘ is a JSON formatted configuration file that describes file paths (such as the locations of genomes and/or metagenomes) and other parameters (such as the HMM to be used for a homology search, and sequence identity cutoffs). Comprehensive user documentation for EcoPhylo is available at https://anvio.org/m/ecophylo.

The minimum input for the EcoPhylo is a gene family hidden Markov model (HMM) and a dataset of genomic and/or metagenomic assemblies. EcoPhylo identifies and clusters target genes or translated proteins across assemblies to yield a non-redundant, representative set of open reading frames (ORFs). Next, an amino acid phylogenetic tree is calculated with the translated representative ORFs yielding the evolutionary history captured by homologues from input assemblies. An additional user input to the workflow is a metagenomic sequencing dataset representing ecological sampling or an experimental setup. With this input, the workflow performs metagenomic read recruitment against the representative ORFs to yield ecological insights into the gene family. Finally, the separate data types are integrated into a phylogeographic representation of the gene family (Figure 4.1).

The resulting sequences from the workflow can be organized in the EcoPhylo interactive interface either using an amino acid phylogenetic tree or using hierarchical clustering based on differential read recruitment coverage across metagenomic samples. Additionally, metagenomes can be hierarchically clustered based on the detection of the target gene family. It is recommended to employ hierarchical clustering of metagenomes or sequences in the EcoPhylo interactive interface with the detection read recruitment statistic (rather than coverage values) to minimize the effect of non-specific read recruitment (https://merenlab.org/anvio-

views/).

An application of the EcoPhylo workflow with default settings will (1) identify gene families with the program ‘hmmsearch‘ in (Eddy 2011) using the user-provided HMM model, (2) annotate affiliate hmm-hits with taxonomic names with ‘anvi-run-scg-taxonomy‘ when applicable, (3) remove hmm-hits with less than 80% HMM model alignment coverage and incomplete ORFs with the anvi’o program ‘anvi-script-filter-hmm-hits‘ with parameters ‘–min-model-coverage 0.8‘ and ‘–filter-out-partial-gene-calls‘ to minimize the inclusion of non-target sequences and spurious HMM hits, (4) dereplicate the resulting DNA sequences at 97% gANI and pick cluster representatives using MMseqs2 (Steinegger and Soding, 2017), (5) use the translated representative sequences to calculate a multiple sequence alignment (MSA) using MUSCLE with the ‘-maxiters 2‘ flag (Edgar, 2004), trim the alignment by removing columns of the alignment with trimAL with the ’-gappyout’ flag (Capella-Gutiérrez et al., 2009), (6) remove sequences that have more than 50% gaps using the anvi’o program ‘anvi-script-reformat-fasta‘, (7) calculate a phylogenetic tree using FastTree (Price et al., 2010) with the flag ‘-fastest‘, (8) perform metagenomic read recruitment analysis and profiling of non-translated representative sequences using the anvi’o metagenomic workflow (Shaiber et al., 2020), which by default relies upon Bowtie2 (Langmead and Salzberg, 2012), (9) generate miscellaneous data to annotate the representative sequences including taxonomy with ‘anvi-estimate-scg-taxonomy‘ for ribosomal proteins, cluster size, and sequence length, and finally (10) generate anvi’o artifacts that give integrated access to the phylogenetic tree of all representative sequences and read recruitment results that can be visualized using the anvi’o interactive interface and/or further processed for specific downstream analyses using any popular data analysis environment such as R and/or Python.

The workflow that resulted in the recovery and characterization of ribosomal proteins in our manuscript used the following additional steps: (1) we removed input reference genomes that were not detected in at least one of the input metagenomes above a detection value of

0.9 with their ribosomal protein (we kept all MAGs originating from the samples themselves) to only visualize detected populations, (2) we manually curated the ribosomal protein tree when necessary to remove sequences that appeared to be chimeric and those that formed spurious long branches likely originating from metagenomic assembly artifacts, and/or mitochondrial or plastid genomes (Supplementary information) and recalculated new amino acid phylogenetic trees with curated sequences with ‘FastTree‘ or IQTREE with the parameters ‘-m WAG -B 1000‘ (Minh et al., 2020) and imported the new trees using the program ‘anvi-import-items-order‘, and finally,Âă (3) we generated additional metadata using in-house Python or R scripts and imported additional metadata using the program ‘anvi-import-misc-data‘ to decorate trees or metagenomes.

3.5.2 Benchmarking EcoPhylo workflow with ribosomal proteins using CAMI synthetic metagenomes

We validated the EcoPhylo workflow by benchmarking it against the CAMI synthetic metagenomes (Meyer et al., 2022) to identify nucleotide clustering thresholds of ribosomal gene families to limit non-specific read recruitment while maximizing taxonomic resolution. We applied the EcoPhylo workflow across the three CAMI biome synthetic genomic/metagenomic datasets (Marine, Plant-associated, and Strain-madness). As an initial step, we identified the top five most frequent ribosomal gene families that were detected in single-copy in the associated genomic collections for each synthetic metagenomic dataset. We then conducted a parameter grid search, spanning 95%-100% nucleotide similarity parameter grid search (Meyer et al., 2022). Next, we measured the amount of non-specific read recruitment in each EcoPhylo iteration, i.e. reads with equal mapping scores between their primary and secondary alignments (multi-mapped reads), with the following Samtools command: ‘samtools view $sample | grep XS:i | cut -f12-13 | sed ’s/..:i://g’ | awk ’$1==$2’ | wc -l‘. The percentage of non-specific read recruitment was calculated by dividing the number of multi-mapped reads by the total num-

ber of reads mapped to the representative dataset. With this, we identified that nucleotide clustering thresholds greater than 97% began to show signs of non-specific read recruitment (Supplementary information).

After identifying 97% nucleotide identity as the optimal threshold, we measured EcoPhyloâĂŹs ability to contextualize a genomic collection within metagenomic assemblies by quantifying the amount of genomic ribosomal genes clustering with their associated metagenomic assembly ribosomal gene (Supplementary information). Finally, we benchmarked the Shannon diversity and richness captured by different SCGs within the metagenomes and compared it to other taxonomic profiling tools submitted to CAMI (Meyer et al., 2022). To calculate Shannon diversity and richness values for SCGs processed by EcoPhylo we used the R package vegan (Dixon, 2003) and Phyloseq (McMurdie and Holmes, 2013). To calculate the richness and alpha diversity values for CAMI ground truth and other profiling tools we extracted relative abundance for each genera included in the associated biome files made available from CAMI. Shannon diversity for SCGs in the EcoPhylo were calculated with the anvi'o coverage statistic: Q2Q3 coverage. Datasets were cleaned and visualized with R packages in Tidyverse (Wickham et al., 2019).

3.5.3 Genome collections

All MAG and SAG datasets were filtered for genomes with 50% completion and 10% redundancy using the single-copy core gene collections in anvi'o to meet medium-quality draft status in accordance with the community standards (Bowers et al., 2017). For the human oral cavity analysis, 8,615 human oral isolate genomes were downloaded from HOMD v10.1 (`https://www.homd.org/ftp/genomes/NCBI/V10.1/`) (Escapa et al., 2018) and 790 MAGs were downloaded from Shaiber et al. (2020) via (doi:`https://doi.org/10.6084/m9.figshare.12217805`, doi:`https://doi.org/10.6084/m9.figshare.12217961`).

For the Hadza tribe human gut microbiome analysis we followed the data download guide-

lines shared by Carter et al. (2023) to obtain genomes from doi:10.5281/zenodo.7782708. Carter et al. (2023) formed clusters at 95% gANI by including additional genomes outside of the MAGs they have reconstructed from the Hazda gut metagenomes. To exclusively analyze microbial genomes affiliated with the Hadza metagenomes, we filtered for cluster representatives with cluster members that contained at least one Hadzda adult or infant MAG which produced 2,437 representative Bacterial and Archaeal MAGs.

Finally, the surface ocean genomic collection was based on Paoli et al. (2022) and augmented with SAGs (Martinez-Perez et al., 2022) and isolate genomes for SAR11 and *Prochlorococcus* (Delmont et al., 2018b; Delmont and Eren, 2018). When metadata was available, we only used genomes sampled from x < 30 meters depth to match the surface ocean metagenomic dataset, otherwise, we retained the genomes. The Paoli et al. (2022) MAG collection included manually curated MAGs from co-assemblies, which included samples from depths deeper than 30 meters in the deep chlorophyll maximum (Delmont et al., 2018b). The final input surface ocean genome dataset contained 1,474 SAGs, 1,723 isolate genomes, and 7,282 MAGs.

3.5.4 Metagenome and metagenomic assembly datasets

To explore the phylogeography of ribosomal proteins, we used used 71 tooth and plaque metagenomes from the human oral cavity which were downloaded from the NCBI BioProject PRJNA625082 (Shaiber et al., 2020) along with associated co-assemblies (doi:`https://doi.org/10.6084/m9.figshare.12217799`). Next, to explore deep sequencing in the human gut microbiome we used 388 metagenomes and assemblies from infant and adult members of the Hadza tribe (doi:10.5281/zenodo.7782708) using the FTP links shared in from the file 'Supplemental_Table_S1.csv' and NCBI BioProject PRJEB49206 (Carter et al., 2023). Finally, to explore the global surface ocean microbiome, we used 237 surface ocean metagenomes and associated assemblies (<30 meters depth) from NCBI BioProjects PR-

JEB45951 and PRJEB5245228 (Paoli et al., 2022; Sanchez et al., 2023). All metagenomes and associated assembly accessions can be found at Supplementary Table 1.

3.5.5 Preprocessing of genomic and metagenomic assemblies and metagenomic short reads

Metagenomic and genomic assemblies were preprocessed with the anvi'o contigs workflow with the program 'anvi-run-workflow -w contigs' to predict open-reading frames with Prodigal (V2.6.3) and identify SCGs for taxonomic inference with 'anvi-run-scg-taxonomy' (Hyatt et al., 2010; Shaiber et al., 2020). No contig size filters were implemented during this process to include ribosomal proteins located on small contigs. To limit detection of misassemblies in downstream analyses, only ribosomal proteins with complete open-reading frames (as predicted by Prodigal) were analyzed with EcoPhylo (Hyatt et al., 2010) . Additionally, metagenomic samples were quality controlled with the anvi'o metagenomics workflow with the program 'anvi-run-workflow -w metagenomics' (Shaiber et al., 2020). This workflow uses the tool 'iu-filter-quality-minoche' (Eren et al., 2013), which implements methods described in (Minoche et al., 2011). All Snakemake workflows in this manuscript leveraged Snakemake v7.32.4 (Koster and Rahmann, 2012).

3.5.6 Gene-level taxonomy of ribosomal proteins

To assign gene level taxonomy to ribosomal proteins, the EcoPhylo workflow relies upon the anvi'o tools 'anvi-run-scg-taxonomy' and 'anvi-estimate-scg-taxonomy', which leverage the genomes and their taxonomy made available by the GTDB (Parks et al., 2022) to identify taxonomic affiliations of genes that match to any of the ribosomal proteins L1, L13, L14, L16, L17, L19, L2, L20, L21p, L22, L27A, L3, L4, L5, S11, S15, S16, S2, S6, S7, S8, or S9. During the workflow, EcoPhylo uses 'anvi-run-scg-taxonomy' to search for ribosomal genes

annotated within each anvi'o contigs database against the downloaded marker gene dataset with DIAMOND v0.9.14 (Buchfink et al., 2021). Later in the workflow, EcoPhylo runs 'anvi-estimate-scg-taxonomy –metagenome-mode' on the representative set of ribosomal proteins, which assigns a consensus taxonomy to each sequence. The program 'anvi-estimate-scg-taxonomy' does not provide a taxonomic annotation if the ribosomal protein is less than 90% similar to any of the ribosomal proteins found in GTDB genomes. In some cases, ribosomal proteins without taxonomic annotation can be manually annotated with taxonomy based on the annotated sequences that surround them in the phylogenetic tree, as we described in the section "Taxonomic binning to improve genome recovery estimations".

3.5.7 Selection of ribosomal proteins to contextualize genomic collections in metagenomes

To pick ribosomal gene families to study genome collections, we selected ribosomal genes that were annotated in the majority of genomes in single-copy. We then cross-referenced selected ribosomal genes with their assembly rates in metagenomes and disregarded candidate ribosomal gene families that were under- or over-assembled in the dataset. To do this, we ran the EcoPhylo workflow with the input dataset of genomic and metagenomic assemblies until the rule 'process_hmm_hits', which will filter for high-quality HMM-hits as described above. Finally, we extracted ribosomal protein hits from all assemblies with the anvi'o command 'anvi-script-gen-hmm-hits-matrix-across-genomes' and tabulated/visualized the distribution in R using the Tidyverse (Wickham et al., 2019).

3.5.8 Distribution of HMM alignment coverage and SCG detection across GTDB

To identify optimal ribosomal proteins and HMM hit filtering thresholds, we explored the distribution of SCG detection and HMM alignment coverage across GTDB genomes. The analysis used the first two rules of the EcoPhylo workflow (anvi_run_hmms_hmmsearch and filter_hmm_hits_by_model_coverage) to annotate the RefSeq representative genomes from GTDB release 95 (Parks et al., 2020), with the single-copy core gene HMM collections included in anvi'o. The first rule of the workflow used the program 'hmmsearch' to identify HMM hits, while the second rule was modified to include all HMM hit model coverage values by setting the parameter 'anvi-script-filter-hmm-hits-table –min-model-coverage 0'. We stopped the workflow after this rule and visualized the raw distribution of model and gene coverage values from 'hmmsearch –domtblout' output file leading us to to identify an 80% HMM hit model coverage as an optimal filtering threshold to identify ribosomal proteins. Next, we restarted the workflow but re-modified the second rule parameter 'anvi-script-filter-hmm-hits-table –min-model-coverage 0.8' to filter for HMMs hits with at least 80% model alignment coverage. Finally, we extracted all ribosomal gene families from the genome dataset with anvi'o program 'anvi-script-gen-hmm-hits-matrix-across-genomes' and visualized the genome detection and SCG copy number across the dataset in in R using the Tidyverse (Wickham et al., 2019).

3.5.9 Detection of whole genomes in metagenomic data

In some cases, ribosomal proteins clustering at 97% brought together large groups of highly similar isolate genomes. To identify the specific genome that is detected in the metagenomic datasets, we re-clustered the target EcoPhylo protein at 98% to resolve sequence clusters and thus increase the number of representative sequences. We then used the whole genomes associated with the new, larger set of representative proteins to explore their distribution in

metagenomes by performing the anvi'o metagenomic workflow (Shaiber et al., 2020). Our threshold for detection of a whole-genome in metagenomic data was 50% (percent of genome covered by at least one read from metagenomic read recruitment), which was found to be efficient for human oral cavity microbes (Utter et al., 2020).

3.5.10 Genome recovery rate estimations

Genome recovery rates were estimated to measure which individual or combination of genome types (MAGs, SAGs, isolate genomes) most effectively sampled clades in the ribosomal protein phylogenetic trees calculated during the EcoPhylo workflow. To calculate genome recovery rates for any given taxon, we divided the number of sequence clusters that contained a sequence from a given genome recovery method to the total number of representative sequences EcoPhylo reported for that taxon. Taxonomic assignments of sequence cluster representatives were determined with 'anvi-estimate-scg-taxonomy'.

3.5.11 Taxonomic binning to improve genome recovery estimations

A subset of ribosomal proteins lacked taxon assignments from 'anvi-estimate-scg-taxonomy' due to their sequence similarity being x < 90% to GTDB genomes (See methods section: Gene-level taxonomy of ribosomal proteins). Using the 'anvi-interactive' interface, we examined the placement of these proteins in the EcoPhylo ribosomal protein phylogenetic tree and manually assigned taxon names based on the taxonomic affiliations of neighboring sequences. Unannotated sequences were assigned taxonomy only when phylogenetic clustering demonstrated clear consistency among neighboring sequences. These refined taxonomic annotations were used to improve estimations of genome recovery in the main figures (*rpL19* and *rpS15* in Figure 3.2 and *rpL14* in Figure 3.3).

3.6 Data and code availability

The URL https://merenlab.org/data/ecophylo-ribosomal-proteins/ serves all code and data needed to reproduce our study. Additionally, all anvi'o artifacts that give interactive access to EcoPhylo interfaces are publicly available at doi:10.6084/m9.figshare.28207481. Publicly available genomes and metagenomes we used in our study are listed in the Supplementary Tables, which are available via doi:10.6084/m9.figshare.28200050, along with the Supplementary Information text.

3.7 Acknowledgements

We thank all authors who made the raw metagenomic reads, metagenomic assemblies, and genomes from their manuscripts publically available and conveniently accessible for secondary analyses. We also thank the members of the Meren Lab (`https://merenlab.org/people/`), the Light Lab (`https://www.lightlab.uchicago.edu/people/`), and the Blekhman Lab (`http://blekhmanlab.org/members.html`) for helpful discussions and whiteboard sessions. We are also thankful to Pedram Esfahani, Kimberly Grasch, and the rest of the University of Chicago Center for Research and Computing Center for their patience and support. MSS acknowledges support from NIH Genetics and Regulation Training Grant (T32 GM07197). AME acknowledges support from the Center for Chemical Currencies of a Microbial Planet (C-CoMP) (NSF Award OCE-2019589, C-CoMP publication #070), and Simons Foundation (grant #687269).

3.8 Author contributions

MSS and AME conceptualized the study. MSS curated data and performed formal analyses. MSS, IAV, MLK, MS, SEM, and AME developed software tools. MSS, FT, and AME interpreted findings. LM, TOD, and SHL helped with interpretation of results. MSS and AME wrote the

original draft of the study. SHL and AME managed the project and acquired funding. All authors commented on and made suggestions, and approved the final manuscript.

3.9 Supplementary Information

3.9.1 Supplemental Figures

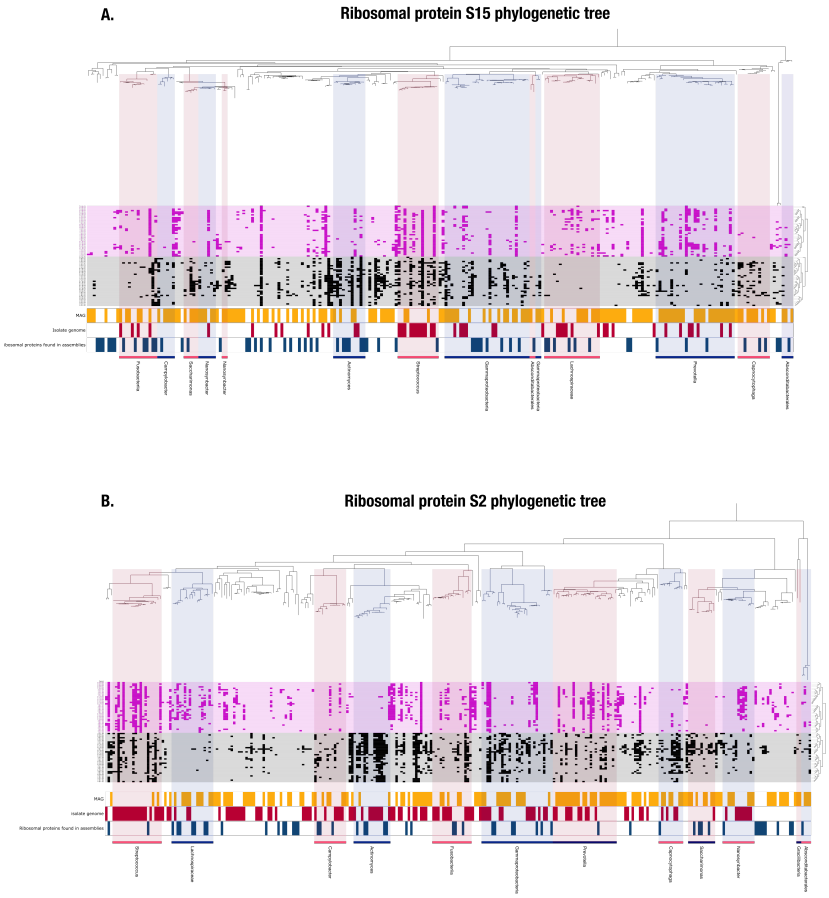

Figure 3.4: **Ribosomal protein phylogeny and detection patterns across metagenomes from the human oral cavity.** In the heatmaps in both panels, each column represents a ribosomal protein representative sequence, each row represents a metagenome, and each data point indicates whether a given ribosomal protein was detected in a given metagenome. The columns of heatmaps are ordered by a tree representing a phylogenetic analysis of all ribosomal protein representative sequences, and the rows are ordered by a hierarchical clustering dendrogram that is calculated based on the ribosomal protein detection patterns across metagenomes. Panel (A) represents the EcoPhylo analysis of rpS15 and panel (B) is rpS2 sequences. Both analyses include Shaiber et al. (2020) metagenome-assembled genomes (MAGs), Shaiber et al. (2020) oral metagenomes, and HOMD genomes, and include three additional rows that indicate the origin of a given ribosomal protein, whether it is a metagenome-assembled genome (MAG, gold), HOMD isolate genome (red), or only recovered from metagenomic assemblies with no representation in genomes (blue).

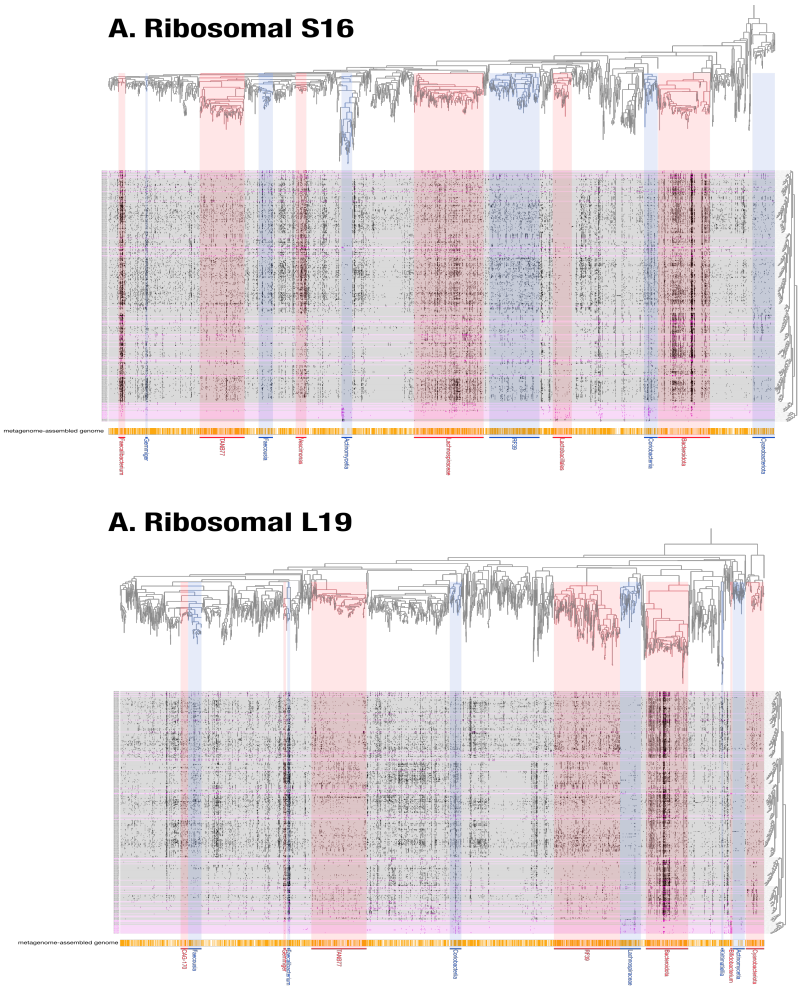

Figure 3.5: **Ribosomal protein phylogeny and detection patterns across Carter et al. (2023) metagenomes.** In the heatmaps in both panels, each column represents a ribosomal protein representative sequence, each row represents a metagenome, and each data point indicates whether a given ribosomal protein was detected in a given metagenome. The columns of heatmaps are ordered by a tree representing a phylogenetic analysis of all ribosomal protein representative sequences, and the rows are ordered by a hierarchical clustering dendrogram that is calculated based on the ribosomal protein detection patterns across metagenomes. Panel (A) represents the EcoPhylo analysis of rpS16 and panel (B) is rpL19 sequences. Below both analyses, each leaf of the phylogenetic tree is decorated denoting if the detected populations contains a metagenome-asembled genome (yellow).

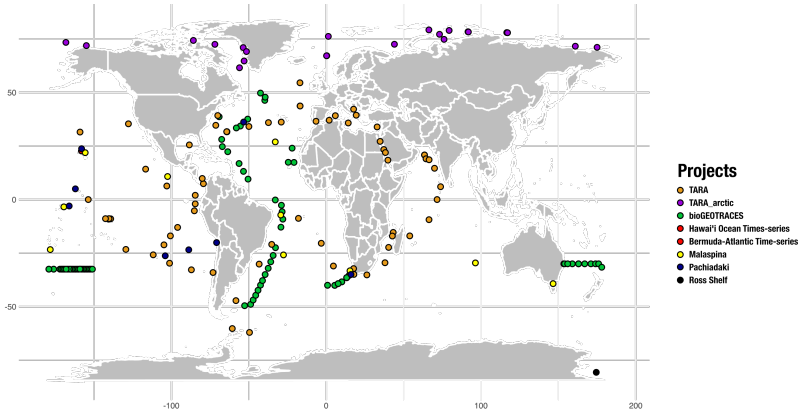

Figure 3.6: **Sample map of surface ocean metagenomes and single-amplified genome sampling sites**

3.9.2 Supplementary Table Legends

This section's supplementary tables are accessible via doi: 10.6084/m9.figshare.28200050

Table 3.1: List of genomes and metagenomes analyzed in this manuscript

Table 3.2: Metadata from all EcoPhylo runs, including human oral cavity (*rpL19*, *rpS15*, and *rpS2*), human gut (*rpS15*, *rpS16*, and *rpL19*), and surface ocean (*rpL14*, rpS8, and *rpS11*

Table 3.3: Genome recovery rates calculated with taxonomic binning results for EcoPhylo analysis of *rpL19* across Shaiber et al. (2020) human oral microbiome MAGs and HOMD

Table 3.4: **Supplementary Table 4: Genome recovery rates for EcoPhylo analysis of Shaiber et al., 2020 oral microbiome metagenomes using *rpL19*, *rpS15*, and *rpS2***. This table presents the genome recovery rates derived from the EcoPhylo workflow applied to oral microbiome metagenomes from Shaiber et al., 2020. The analysis utilized ribosomal proteins *rpL19*, *rpS15*, and *rpS2*, with results directly obtained from the 'anvi-scg-taxonomy' tool.

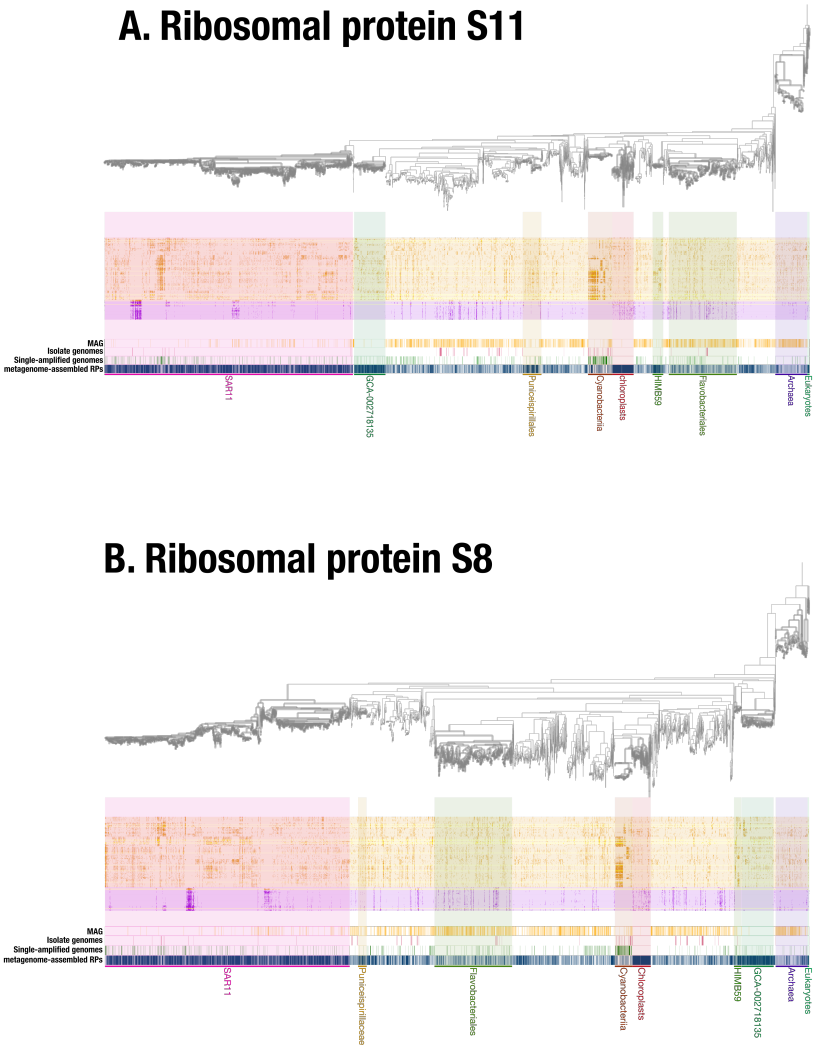

Figure 3.7: **Ribosomal protein phylogeny and detection patterns across global surface ocean metagenomes (depth < 30 m).** In the heatmaps in both panels, each column represents a ribosomal protein representative sequence, each row represents a metagenome, and each data point indicates whether a given ribosomal protein was detected in a given metagenome. The columns of the heatmaps are ordered by a tree representing a phylogenetic analysis of all ribosomal protein representative sequences, and the rows are ordered by a hierarchical clustering dendrogram calculated based on the ribosomal protein detection patterns across metagenomes. Panel (A) represents the EcoPhylo analysis of rpS11 sequences, and Panel (B) represents rpS8 sequences. Below both analyses, each leaf of the phylogenetic tree is marked to denote whether the detected population contains a MAG (yellow), isolate genome (red), SAG (green), or is only recovered from metagenomic assemblies with no representation in genomes (blue).

Table 3.5: Genome recovery rates calculated with taxonomic binning results for EcoPhylo analysis of *rpS15* across Carter et al. (2023) human gut microbiome MAGs from the Hadza tribe adults and infants

Table 3.6: Genome recovery rates for EcoPhylo analysis of Carter et al. (2023) Oral microbiome metagenomes with *rpL19*, *rpS15*, and *rpS2* with direct output of 'anvi-scg-taxonomy'

Table 3.7: Genome recovery rates calculated with taxonomic binning results for EcoPhylo analysis of rpL14 across surface ocean metagenomes assembled from Paoli et al. (2022) and genome collection including MAGs, SAGs, and isolate genomes

Table 3.8: **Supplementary Table 8: Genome recovery rates for EcoPhylo analysis of surface ocean metagenomes.** This table presents the genome recovery rates derived from the EcoPhylo workflow applied to surface ocean metagenomes assembled by Paoli et al. (2022). The analysis includes genome collections comprising MAGs, SAGs, and isolate genomes, and utilizes ribosomal proteins *rpL14*, *rpS8*, and *rpS11*. Results were obtained directly from the 'anvi-scg-taxonomy' tool.

3.9.3 Supplementary Information Text

Single-copy core genes effectively detect microbial populations and contextualize genomic collections in metagenomes

To evaluate the ability of EcoPhylo to process ribosomal proteins to *de novo* survey microbial diversity in metagenomes and effectively contextualize genome collections within metagenomes, we benchmarked the workflow from multiple angles. We first investigated single-copy core gene (SCG) HMM searches against reference genomes as a control to identify thresholds for removing spurious HMM hits in metagenomic assemblies. To do this, we searched the SCG Pfam HMM collections Bacteria_71 and Archaea_76 (modified from (Lee, 2019) in anvi'o (Eren et al., 2021) across the Genome Taxonomy Database (GTDB) Archaea and Bacteria RefSeq genomes (release95) and surveyed HMM alignment coverages (Parks et al., 2020)(Figure 3.8). We found that most SCG HMMs had high-quality model coverages (x > 95%) in both Bacteria and Archaea. Moving forward, we selected 80% HMM alignment coverage as a threshold for identifying ribosomal proteins with divergent lengths while filtering out spurious

hits in metagenomic assemblies, e.g., chimeras and misassemblies (Salzberg et al., 2012; Sczyrba et al., 2017; Mikheenko et al., 2016). After applying the HMM alignment coverage to the GTDB SCG search results, we identified a subset of SCG HMMs that detected nearly 100% of GTDB genomes in single-copy, however, some models poorly detected genomes in single-copy (Figure 3.9). For example, Ribosomal L3 (PF00297) detected zero Bacteria genomes after 80% HMM alignment coverage filtering. Interestingly, Ribosomal L6 was consistently detected twice per genome in Bacteria and Archaea. Upon further investigation, we found that the Ribosomal L6 HMM model (PF00347) was detected twice per ORF due to a domain duplication event. A possible explanation for ribosomal protein HMM hits with low alignment coverage can be a lack of complete phylogenetic diversity representation of the protein family or an overrepresentation of taxa with shorter lengths. Overall, these findings identified ribosomal protein Pfam models (*rpL1*, *rpL14 rpL32e*, *rpS19*, *rpS9*, *rps12_s23*, *rpL16*, *rpS17*, *rpS7*, *rpS8*, *rpL22*, *rpL29*, *rpL4*, *rpS11*, *rpS13*) and filtering thresholds (80% HMM alignment coverage) that accurately detect genomes in single-copy to effectively detect microbial populations in metagenomic assemblies (Figure 3.9).

After optimizing the workflow to identify ribosomal proteins confidently, our next step was to cluster ribosomal proteins extracted from assemblies into a representative dataset for studying microbial ecology across environments using metagenomic read recruitment. This required identifying an optimal ribosomal protein gene clustering threshold to limit non-specific read recruitment and maximize taxonomic resolution. To do this, we leveraged CAMI, a widely used dataset of synthetic metagenomes, modeled from a dataset input genomes and abundance values, representing multiple microbiomes (e.g., marine and plant-associated biomes) for evaluating metagenomic methods (Meyer et al., 2022). First, we selected ribosomal proteins that detected the majority of the input genome dataset in single-copy CAMI datasets (Figure SI3, Materials and Methods). Next, we ran the EcoPhylo workflow on CAMI datasets with ribosomal protein gene clustering thresholds from 95-100% nucleotide sequence similarity (Figure 3.11,

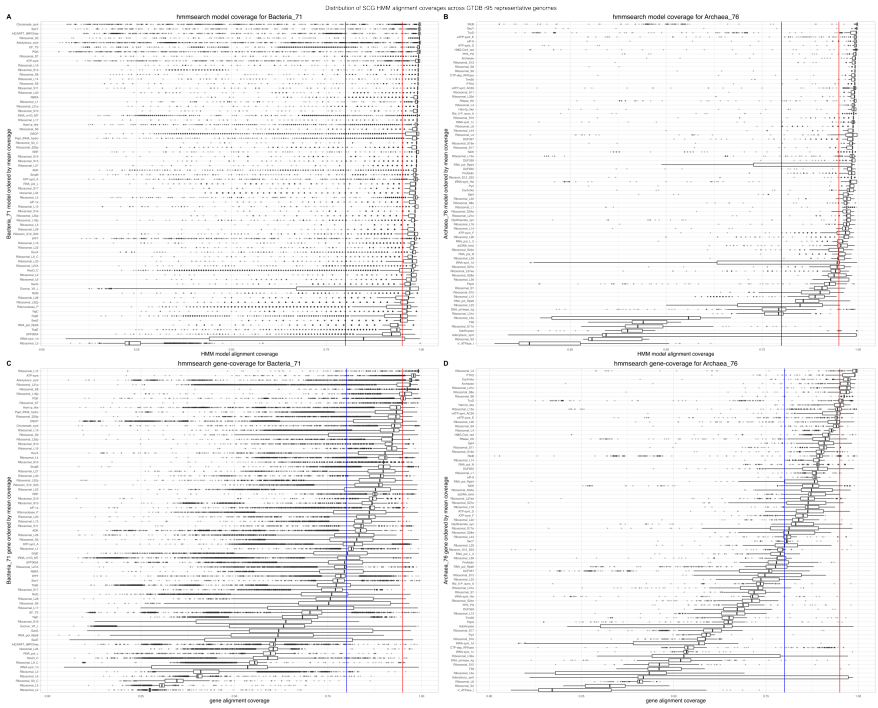

Figure 3.8: **Distribution of HMM alignment coverage of SCG HMM collections searched against GTDB RefSeq Archaea and Bacteria representative genomes.** (A) and (B) HMM model coverage. The x-axes are boxplots representing the distribution of HMM model coverages against query open reading frames from GTDB Archaea and Bacteria representative genomes (r95). The blue vertical line indicates 80% alignment coverage, and the red verticle line represents 95%. The Y-axis represents the list of HMM models. (C) and (D) Target gene coverage

Materials and Methods). This analysis suggested that 97% nucleotide similarity was optimal to limit non-specific read recruitment while maximizing taxonomic resolution.

Next, we benchmarked how clustering ribosomal protein genes at 97% nucleotide similarity would impact EcoPhylo workflow's ability to contextualize a genome collection within a metagenomic dataset. In other words, measure the threshold's ability to group ribosomal proteins from a genome collection with metagenome assembly-derived ribosomal proteins, thereby highlighting microbial diversity lacking genome representation. To explore this, we analyzed the EcoPhylo analyses with the 97% nucleotide identify parameter and observed that all genomic SCGs clustered with their metagenomics counterparts in all three modeled datasets: Marine, plant-associated, and strain-madness. Interestingly, there was a small percentage of

ribosomal protein clusters containing only metagenome-derived ribosomal proteins: 0.55% of the *rpL5* iteration of the Plant Associated dataset of clusters and 0.63% of the *rpS19* iteration of the Marine dataset. This was surprising because that would imply that the *in silico* CAMI metagenomes yielded distinct ribosomal proteins from the input metagenomes. Upon further investigation, this was driven by truncated ORFs called by Prodigal (Hyatt et al. 2010) in the metagenomic assemblies, which formed their own clusters due to limited homology to the genome-derived SCGs. To enforce ribosomal protein cluster formation being driven by complete gene homology, we removed all incomplete open reading frames from downstream analysis using 'anvi-script-filter-hmm-hits-table –filter-out-partial-gene-calls', which can be toggled TRUE/FALSE in the EcoPhylow workflow config file. Furthermore, to ensure effective clustering of open reading frames, we encourage users to change the mmseqs clustering parameter to '–cov-mode 1' (coverage of target), which will improve clustering of both fragmented and extended open reading frames with complete ORF homologs (Steinegger and Soding, 2017). Overall, EcoPhylo effectively contextualized genomic collections within metagenomic assemblies by clustering genomic SCGs with their metagenomic counterparts, highlighting its ability to link genome collections with related populations and identify distinct populations absent from the genome collection.

So far, we have refined the application of the EcoPhylo workflow to identify and yield a representative dataset of ribosomal proteins optimized for measuring biogeography with metagenomic read recruitment and properly contextualizing genomic collection with relation populations in metagenomic assemblies. We next sought to compare standard reference-based taxonomic profiling to the *de novo* taxonomic composition revealed by surveying SCGs across metagenomic assemblies (Olm et al., 2020) . To do this, we used CAMI to compare the alpha diversity captured by ribosomal proteins to the CAMI ground truth and other reference-based taxonomic profiling tools to confirm comparable results (Figure 3.12, Materials and Methods). We first observed that ribosomal proteins had low variance across richness and Shan-

non diversity, highlighting that different ribosomal proteins provide comparable results. Next, we found that in most cases, ribosomal proteins underestimated the ground-truth number of genomes across samples yet overestimated Shannon diversity. This is likely due to 97% ribosomal protein clustering yielding taxonomic units that group genomes together genomes with x < 95% gANI. Furthermore, since ribosomal proteins consistently detected fewer genomes, the increased Shannon index values can be explained by inflated evenness. This revealed that the relative abundance of individual ribosomal proteins was unreliable in this experiment. Moving forward, we used anvi'o detection values (percent of sequence covered by at least 1x reads) as a discrete read recruitment statistic to determine the presence or absence of ribosomal proteins across metagenomes.

Despite EcoPhylo-processed ribosomal proteins yielding a conservative view of the biodiversity detected in metagenomes (presence/absence and taxonomic units grouping genomes with x < 95% gANI), the workflow could detect more taxa than other tools in a subset of the rhizosphere CAMI datasets (Figure 3.12). This highlighted that reference-free taxonomic profiling with ribosomal proteins can be potentially be more sensitive in metagenomes when reference databases have a low representation of the underlying microbiome taxonomic diversity. Overall, our benchmarking demonstrated that EcoPhylo and ribosomal proteins can effectively contextualize genome collections within metagenomic assemblies and enable *de novo* taxonomic profiling of metagenomes, offering a robust approach for exploring microbial diversity in complex datasets.

3.9.4 Selection of Ribosomal proteins to contextualize genomic collections in the human oral cavity, Hadza gut microbiomes, and surface ocean

To compare the oral microbiome genomic collection against populations detected in the metagenomic assemblies, we first identified the most frequent single-copy ribosomal proteins in the genome collection containing MAGs and isolate genomes. The top three ribosomal proteins

included *rpL19*, *rpS15*, and *rpS2*. The overlapping genomes these ribosomal proteins covered were 99.82% of the genomic collection, including 97.8% of the MAGs (Figure 3.13, Supplementary Table SI1, Supplementary Table SI2, Materials and Methods). The genome collection coverage of *rpL19*, *rpS15*, and *rpS2* was documented for the HOMD genomes (Supplementary Table SI1) and MAGs (Supplementary Table SI2). After the initial analysis of *rpL19*, multiple input reference genomes from HOMD were not detected in the metagenomes (90% detection of the target *rpL19* in at least one metagenome). This was expected because the HOMD database contains isolate genomes from numerous other studies. We thus removed these genomes from subsequent analyses, lowering the total number of HOMD genomes in the genomic collection from 8,615 to 1,680.

To explore the MAG recovery rate of the Hadza metagenomes, we analyzed the 2,437 representative genomes where each 95% gANI cluster had at least one Hadza MAG (genome dereplication included MAGs from other populations) (Carter et al., 2023). We next used the most frequent single-copy ribosomal proteins in the MAG collection: *rpS15*, *rpS16*, *rpL19* (Figure 3.14). The combination of the three ribosomal genes contextualized 99.34% of genomes reconstructed from the Hadza metagenomes (Figure 3.14) and the MAG collection coverage was documented (Supplementary Table SI3).

The most frequent single-copy ribosomal proteins in the surface ocean genome collection (MAGs, SAGs, and Isolates) included ribosomal proteins *rpS11*, *rpS8*, and *rpL14*. All three ribosomal proteins together were detected x > 93.77% of the genomic collection (Figure 3.15; Materials and Methods). The genome collection coverage of *rpS11*, *rpS8*, and *rpL14* across MAGs, SAGs, and IGs was documented (Table SI5, Materials and Methods). The initial analysis of *rpL14* revealed isolate genomes from the MAR databases (MarDB v.4) (Klemetsen et al., 2018) that were not detected in the surface ocean metagenomes (90% detection of the target *rpL19* in at least one metagenome). This was most likely due to their isolation from non-pelagic environments such as coastal environments and animal microbiomes (Paoli et al.,

2022) . We thus filtered out genomes that were not detected (Materials and Methods) and reperformed the analysis.

3.9.5 Investigation into cosmopolitan taxa in the Oral Microbiome

EcoPhylo analysis of *rpL19* of the Oral Microbiome (Shaiber et al., 2020) showed four microbial populations that appeared to be cosmopolitan taxa - highly detected in 50% both plaque and tongue biomes - from the genera *Streptococcus*, *Haemophilus* (n = 2), and *Porphyromonas* (Figure 3.16A). We further explored these four populations and found they had large ribosomal protein cluster sizes, the majority of which were multiple closely related isolate genomes collapsed by 97% *rpL19* clustering. In fact, 508 HOMD *Streptococcus rpL19* clustered into one population and appeared in x > 70% of both tongue and plaque samples, highlighting the lack of genome dereplication in the isolate genome dataset (Figure 3.16A). Additionally, a subset of the *rpL19* read recruitment plots from the four cosmopolitan populations showed signs of non-specific read recruitment. This indicated that potentially many similar genomes in HOMD had collapsed into one population and the representative *rpL19* from the EcoPhylo workflow did not reflect the population in the samples.

To explore this, we re-clustered *rpL19* representative sequences and their cluster members with a higher sequence similarity threshold (98% nucleotide identity) to further resolve clusters. This would theoretically increase the number of references to identify a more accurate reference genome for the underlying population captured by the metagenome (Figure 3.16C). Clustering *rpL19* at 98% increased cluster representatives from 4 to 11 representatives, highlighting how clustering ribosomal proteins at 97% collapse closely related genomes. We then selected whole genomes based on the new 11 *rpL19* cluster representatives and performed whole-genome metagenomic read recruitment across the (Shaiber et al. 2020) tongue and plague metagenomes to further explore the ecology and prevalence of these populations across both biomesÂă (Figure 3.16C, Materials and Methods). We found that all of

the original cosmopolitan populations contained at least one genome that was cosmopolitan (identified in both Plaque and tongue samples) which included previously documented oral microbiome generalists including *Haemophilus parainfluenzae, Porphyromonas pasteri*, *Streptococcus pneumoniae*, andÂă *Streptococcus oralis* (Mark Welch et al., 2019; SzafraÅĎski et al., 2021; McLean et al., 2022; Eren et al., 2014; Wilbert et al., 2020) (Figure 3.16C, Table SI4). Interestingly, metapangenomics of *H. parainfluenzae* has identified multiple strains with oral site specificities due to genomic adaptations (Utter et al., 2020).

Interestingly, two *Haemophilus influenzae* genomes, which stemmed from the *Haemophilus* cosmopolitan ribosomal protein, were in fact, biome-specific and predominately detected in the tongue. This indicates that the representative sequence of the original cluster may have recruited reads correctly from one biome but not the other. Leveraging a single ribosomal protein to profile the phylogeography metagenomic read recruitment requires clustering of the gene family to remove redundancy and limit non-specific read recruitment. Clustering brings shortcomings, including large numbers of genomes collapsing into a single cluster, representing a microbial population that can span multiple niches. In fact, one *Haemophilus pittmaniae* genome, which stemmed from the other *Haemophilus* cosmopolitan ribosomal protein, was not detected in any of the metagenomes. Additionally, the non-specific read recruitment signal identified in the so-called cosmopolitan taxa can be explained by the representative sequence not matching the underlying diversity of the population in the sample. Furthermore, the non-specific read recruitment signal highlights the importance of using detection as a read recruitment statistic to avoid the pitfalls of recruiting reads to conserved marker genes. Future directions could explore picking a representative ribosomal protein sequence based on detecting the most prevalent genome in the sample or perform genome dereplication prior to performing the EcoPhylo workflow to avoid this scenario.

3.9.6 Deep sequencing of the human gut yields the highest genome recovery rates across microbiomes

While nucleotide sequencing is expected to become more affordable, a question arises: will deeper sequencing make genome-resolved metagenomics a sufficient genome recovery method in microbial ecology? To investigate this, we analyzed a state-of-the-art metagenomics project with ultra-deep sequencing, which characterized the gut microbiomes of adults and infants from the Hadza hunter-gatherer tribe, generating nearly 50,000 Bacterial and Archaeal MAGs (Carter et al., 2023).

Genome-resolved metagenomics has been an effective genome recovery strategy in the human gut, likely due to efficient metagenomic assembly and binning facilitated by its relatively closed ecosystem leading to lower microbial diversity and clonal expansion of microbial populations. With the gut microbiome's ideal conditions for MAG recovery, the Carter et al. (2023) Hadza gut microbiome survey provided an opportunity to assess whether deeper sequencing alone can overcome the challenges of evenly capturing genomes from microbial communities and reduce the need for multiple recovery methods in certain microbiome ecological contexts. To explore this, we leveraged their representative genomes from genome dereplication clusters that contained at least one Hadza MAG (2,437 representative Bacterial and Archaeal MAGs) (Materials and Methods). Leveraging *rpS15*, we were able to contextualize the majority of the genomic collection (92.2%), including archaeal populations, allowing us to impactfully explore MAG recovery rate (Figure SI7). Additionally, we confirmed these findings with *rpS16* and *rpS19* (Supplementary Figure 2). EcoPhylo analysis of *rpS15* across the Hadza gut microbiome metagenomic dataset detected 2,336 Bacteria and 8 Archaea microbial populations. Hierarchical clustering of metagenomes generally separated samples into adults and infants (black and pink samples, respectively), with infants characterized by lower alpha diversity and detection of *Lactobacillales* and *Bifidobacterium* populations (Figure 3.17). Additionally, the *rpS15* HMM detected multiple plant and amoeba taxa reported by (Carter et al.

2023) further highlighting the multi-domain features of ribosomal proteins in microbial ecology. Furthermore, EcoPhyloâĂŹs ability to process a metagenomic project of this size highlighted the workflowâĂŹs scalability.

Excitingly, the total Bacteria MAG recovery rate was 67.7% (*rpS16*: 72.8%, *rpL19*: 69.5%), the highest genome recovery rate among all the metagenomics projects analyzed in this manuscript (Oral: *rpL19* 60.3%, Ocean: *rpL14* 19.9%) demonstrating that deep sequencing can improve MAG recovery rates (Supplementary Table 5, Supplementary Table 6). Furthermore, multiple taxa had high MAG recovery rates (across all profiled ribosomal proteins: *rpS15*, *rpL19*, *rpS15*) including the phyla *Bacteroidota* 79% (*rpS16*: 57.5%, *rpL19*: 71.6%) and *Cyanobacteriota* 75.8% (*rpS16*: 71.1%, *rpL19*: 70.4%), and lower order taxa such as *Coriobacteriia* 66% (*rpS16*: 71.1%, *rpL19*: 77.8%), RF39 80% (*rpS16*: 84.1%, *rpL19*: 80.2%), and *Lactobacillales* 70.4% (*rpS16*: 77.9%, *rpL19*: 81.8%) (Supplementary Table 5, Supplementary Table 6, Materials and Methods). These results were consistent with Carter et al. (2023) , which reported that approximately 80% of metagenomic reads mapped to the Hadza representative genome collection. Overall, the high MAG recovery rates across the majority of taxa add further evidence that deep metagenomics sequencing is an effective genome sampling method in the human gut.

While ultra deep sequencing allowed for higher MAG recovery rate in general, the Class *Actinomycetia* had one of the lowest MAG recovery rates at 31.9% (*rpS16*: 32.4%, *rpL19*: 33.3%) (Supplementary Table 5, Supplementary Table 6, Figure SI10). Prevalent *Bifidobacteriaceae* MAGs were efficiently recovered but families such as *Pseudonocardiaceae*, *Geodermatophilaceae*, and *Pseudonocardiaceae* had no genomes recovered. Interestingly, three samples from three separate individual metagenomes were responsible for the expansion *Actinomycetia*. Upon further investigation, the unbinned *Actinomycetia* taxa have been documented as common DNA extraction kit contamination microbes (Salter et al., 2014; Weyrich et al., 2019). These results show that deep sequencing can yield high MAG recovery across

bacteria and archaea in the human gut. However, it can also assemble ribosomal genes from low abundance kit contamination, highlighting that caution should be taken with this level of sequencing.

Although deep sequencing in the human gut microbiome demonstrated high genome recovery rates across the majority of taxa, EcoPhylo revealed a subset of populations that were prevalent across the dataset yet not part of the MAG dataset. Exploring *rpS15* co-occurrence across the metagenomic dataset uncovered 17 *rpS15* clusters with high prevalence but no genomic representation (detected in at least 30 samples and frequently co-occurred with other taxa across the dataset) (Figure SI11). To confirm these were indeed unbinned populations, we BLASTed the 17 *rpS15* cluster representatives against the redundant Hadza MAG collection to see if they belonged to redundant genomes within genome clusters. 13 *rpS15* sequences had 100% matches to redundant MAGs leaving four true unbinned microbial populations (*Oscillospiraceae*, *Lachnospiraceae*, and *Gastranaerophilaceae*) with high detection across the metagenomic dataset (Figure SI10, blue triangles). These results highlight that non-representative genomes display prevalent ecology within this metagenomic dataset. While genome dereplication aims to create a dataset of representative genomes to highlight phylogenomic novelty, it does not consider ecological relevance across a dataset (Evans and Denef, 2020). Future applications of EcoPhylo could be used to expand representative genome collections, highlighting prevalent microbial populations even when they share 95% gANI with existing genomes in the collection. Furthermore, the four unbinned populations can be used for targeted genome recovery, leveraging a combination of genome recovery strategies in future sampling campaigns.

3.9.7 Sequencing depth in surface ocean metagenomes

After our original EcoPhylo analysis of *rpL14* over the surface ocean dataset, we clustered metagenomes based on detecting *rpL14* as a proxy for community composition (Figure 3.19b).

We first observed that many metagenomes did not cluster based on expected biogeographical epipelagic patterns and showed signs of batch effect correlating with the sequencing project (Figure 3.19b). We thus compared the sequencing depth between samples to the number of SCGs recovered from associated metagenomic assembles and found that samples with lower sequencing depth correlated with fewer recovered SCG (proxies for microbial populations) (Figure 3.20). By filtering studies based on a minimum sequencing depth of 50 million reads, we could recapitulate *a priori* biogeographical patterns driven by latitude and temperature (Figure 3, Figure 3.19). Moving forward, EcoPhylo analyses of this dataset used the assemblies from all projects to extract ribosomal proteins but only visualized the read recruitment results from 100 metagenomes (x > 50 million reads) which only including the projects TARA Oceans and Malaspina. The final hierarchical clustering of metagenomes based on the detection of a single ribosomal protein (*rpL14*) revealed the documented epipelagic microbiome community structure driven by temperature (Figure 3) (Sul et al., 2013; Sunagawa et al., 2015).

3.9.8 Manually curating the surface ocean Ribosomal protein phylogenetic trees

Although multiple automated steps were implemented in the EcoPhylo workflow to remove assembly artifacts from the analysis, ocean metagenomic data required manual tree curation to remove artificially long branches from spurious HMM hits from the metagenomic assemblies and the mitochondrial signal (Figure 3.19, Materials and Methods). Removing mitochondrial phylogenetic signal was specifically a problem for ocean metagenomic datasets because sample filtering will enrich mitochondria into the archaea and bacteria size fraction.

3.9.9 Supplementary Information Tables

This section's supplementary tables are accessible via doi: 10.6084/m9.figshare.28200050

Table 3.9: *rpL19*, *rpS15*, and *rpS2* frequency across the Human Oral Microbiome Database (HOMD)

Table 3.10: *rpL19*, *rpS15*, and *rpS2* frequency across Shaiber et al. (2020) MAGs

Table 3.11: *rpS15, rpS16, and rpS19*frequency across Hadza representative genomes

Table 3.12: Cosmopolitan populations in Shaiber et al., 2020

Table 3.13: *rpL14*, *rpS11*, and *rpS8* frequency across surface ocean genomes

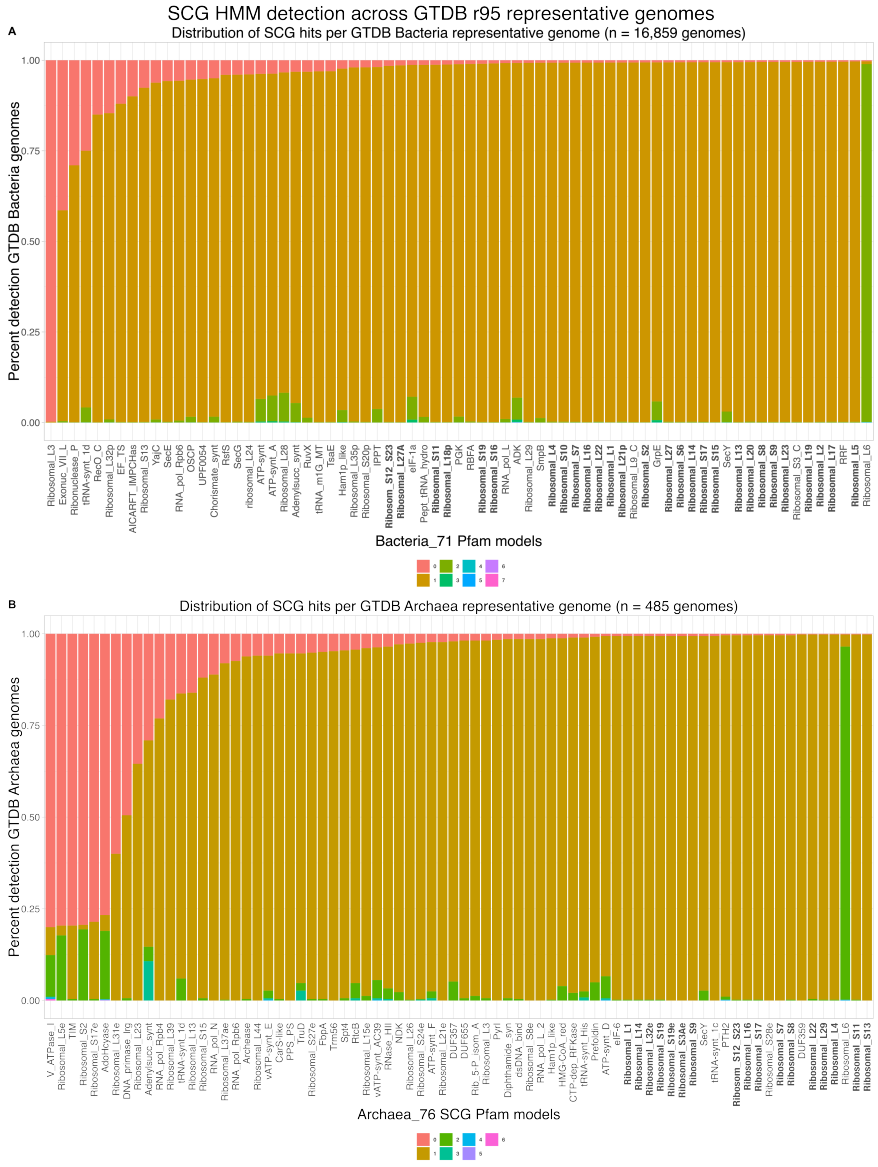

Figure 3.9: **SCG HMM copy number across GTDB r95 RefSeq representative genomes after 80% HMM alignment coverage filtering.** The (A) panel shows bacteria Y-axis represents the percent detection of genomes by each SCG HMM with color-stacked bars representing the distribution of copy-number detected in genomes. Bolded SCGs detect the majority of genomes in single-copy. The (B) shows the same analysis as (A) with archaea representative genomes.

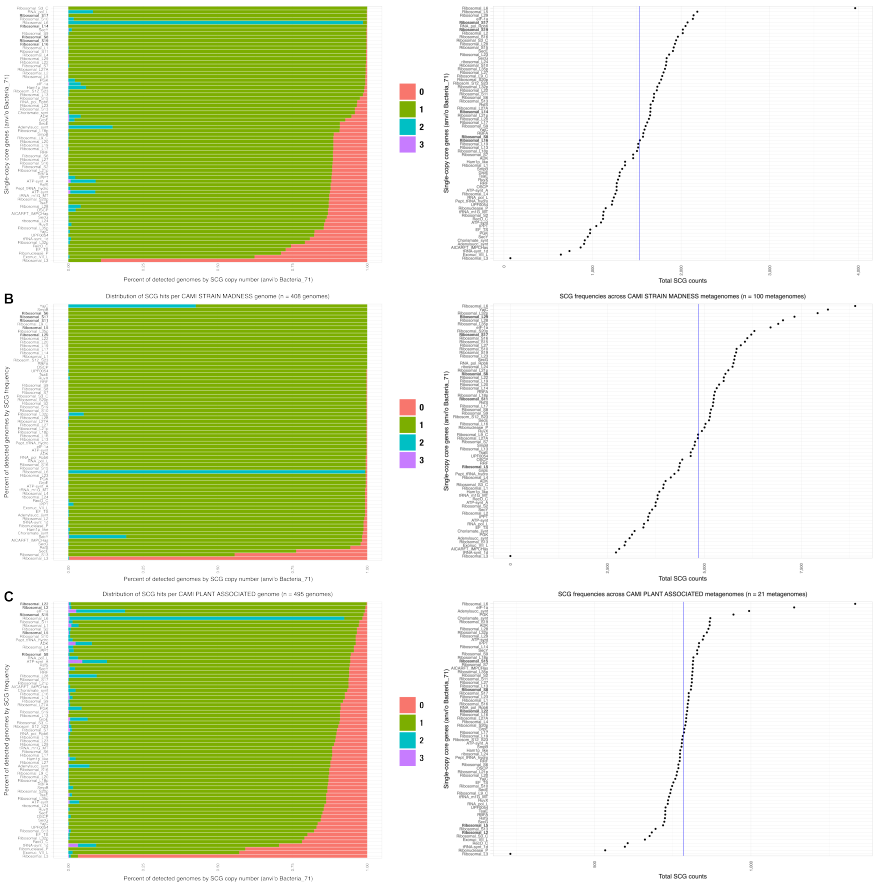

Figure 3.10: **SCG Detection distribution across CAMI genome and synthetic metagenomic assemblies.** The left horizontal bar charts show the copy number detection of each SCG HMM in the genomic dataset. The right plots show the frequency of SCG recovery from the metagenomic assembly. Bolded SCGs were used in subsequent supplementary analyses. (A) CAMI marine dataset (B) CAMI strain madness dataset (C) CAMI plant associated dataset.

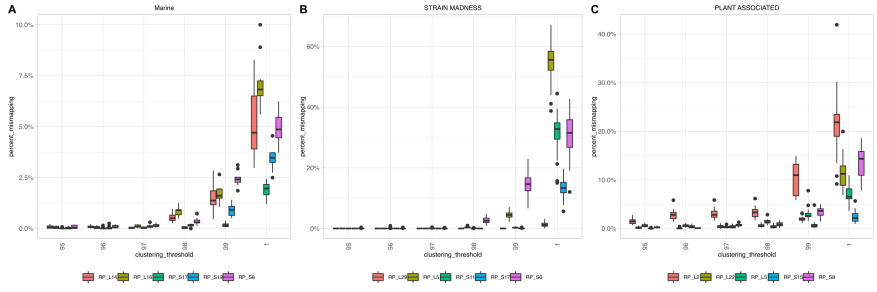

Figure 3.11: **Distribution of multi-mapping reads vs. clustering threshold using CAMI datasets.** The EcoPhylo workflow processed the CAMI metagenomic datasets (Marine, plant-associated, strain madness) with the top 5 SCGs (Figure 3.10) and different ribosomal gene clustering thresholds (95-100%). The y-axis represents the proportion of multi-mapped reads, which is the number of multi-mapped reads divided by the total reads mapped.

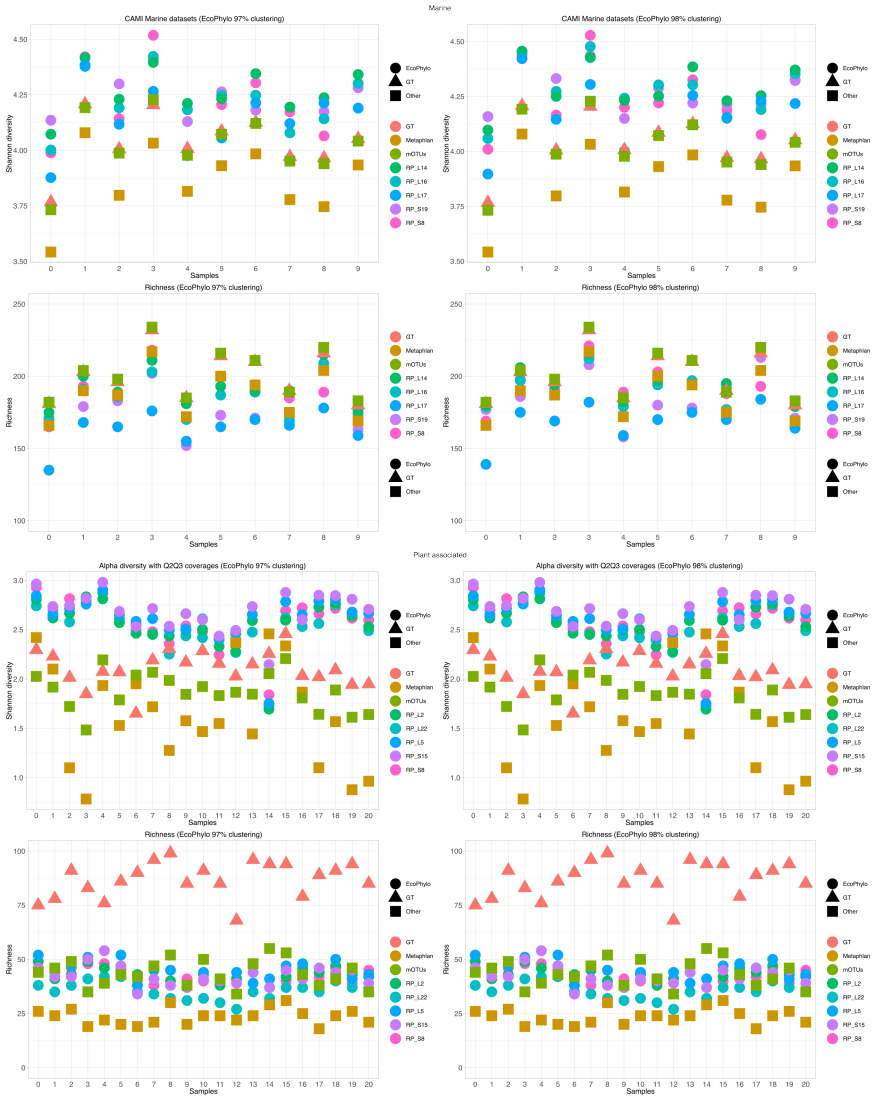

Figure 3.12: **Shannon alpha diversity and Richness of EcoPhylo vs CAMI ground truth and submitted tools with 97% and 98% nucleotide similarity clustering threshold for Marine and Plant associated datasets.** Shannon alpha diversity measurements and richness (number of genomes detected) were calculated for each CAMI synthetic sample. The X-axis represents each CAMI synthetic metagenome.

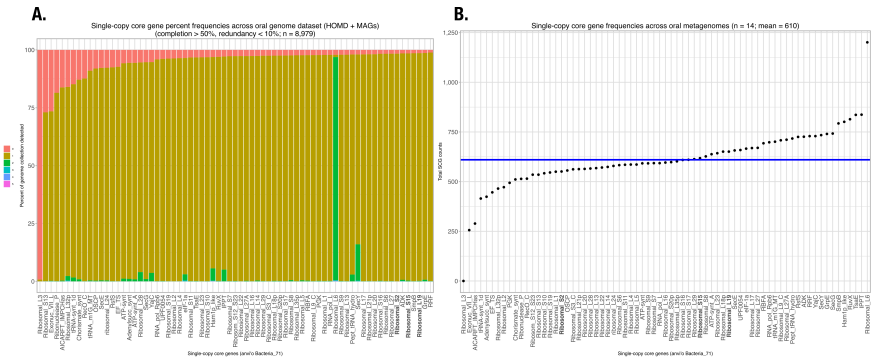

Figure 3.13: **SCG Detection distribution across Shaiber et al., (2020) oral cavity genome dataset (MAG + HOMD) and metagenomic assembly dataset.** The (A) panel shows the SCG HMM copy number for the genome dataset that were filtered for medium-quality draft status in accordance with the community standards (8,615 HOMD isolate genomes and 364 MAGs filtered for medium quality). The color-stacked bars represent the relative distribution of copy-number detected in genomes. Panel (B) shows the SCG recovery number 14 co-assemblies. In (A) and (B), the bolded ribosomal proteins were selected for downstream analyses and the blue line represents mean count of SCGs across all metagenomic assemblies.

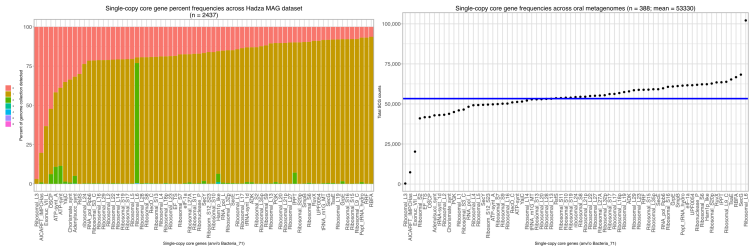

Figure 3.14: **SCG Detection distribution across Carter et al., (2023) Hadza MAG and metagenomic assembly dataset.** The (A) panel shows the SCG HMM copy number for the genome dataset and the color-stacked bars represent the relative distribution of copy-number detected in genomes. Panel (B) shows the SCG recovery number 388 co-assemblies. In (A) and (B), the bolded ribosomal proteins were selected for downstream analyses and the blue line represents mean count of SCGs across all metagenomic assemblies.

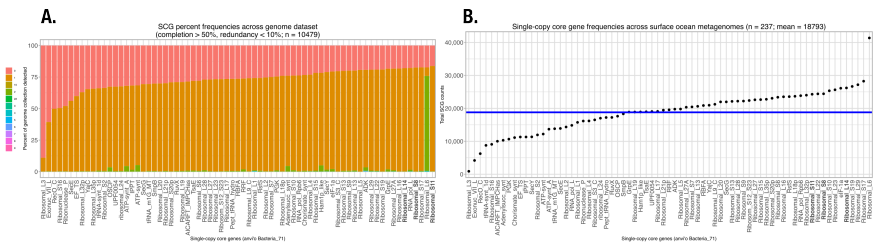

Figure 3.15: **SCGs detected across surface ocean genomic and metagenomic assembly datasets.** The (A) panel shows the SCG HMM copy number for the genome dataset including isolate genomes, MAGs, and SAGs were filtered for medium-quality draft status in accordance with the community standards (Bowers et al. 2017) and the color-stacked bars represent the relative distribution of copy-number detected in genomes. Panel Panel (B) shows the SCG recovery number across 237 surface ocean metagenomic assemblies. The blue line represents mean count of SCGs (n = 18,793) across all metagenomic assemblies.

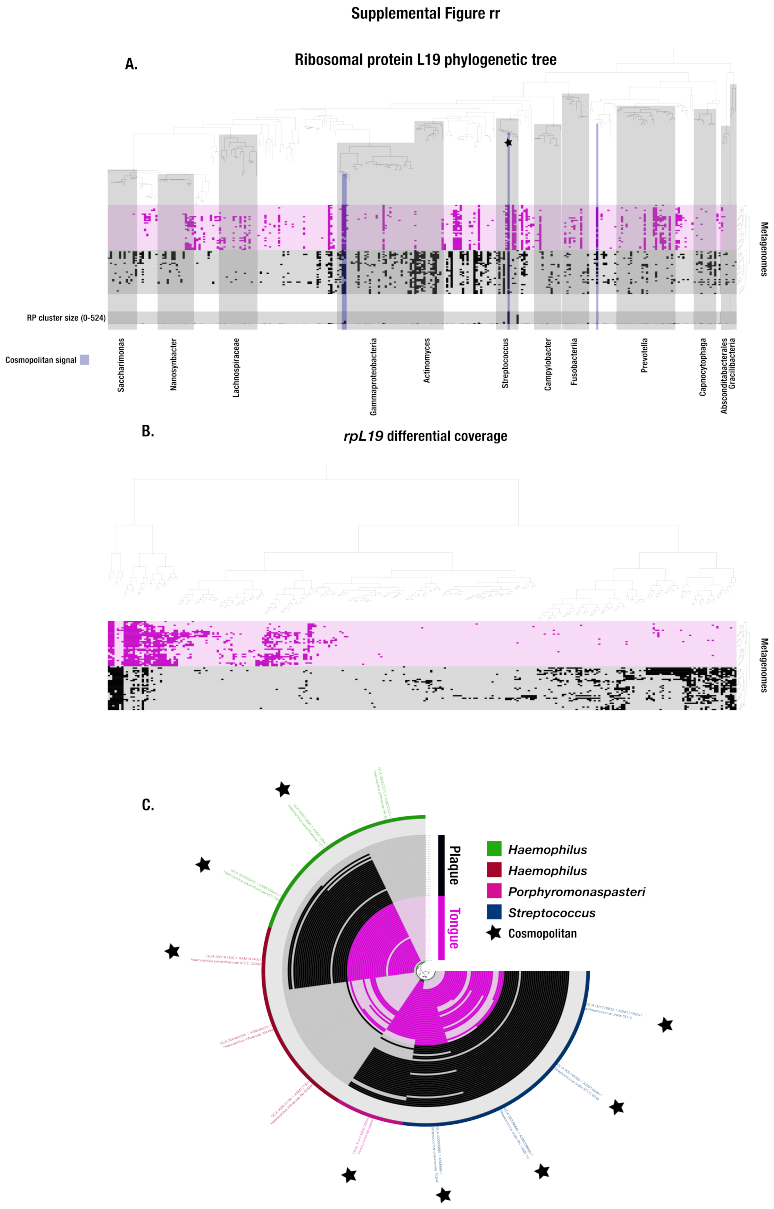

Figure 3.16: **Ribosomal L19 phylogeny and detection patterns across metagenomes across Shaiber et al., 2020 human oral cavity metagenomes.** In panel (A) the presence/absence heat map columns are ordered by a Ribosomal L19 protein phylogenetic tree. The rows (metagenomes: (black = plaque; pink = tongue)) are ordered by a hierarchical clustering dendrogram calculated based on the ribosomal protein detection patterns across metagenomes. Each Ribosomal L19 phylogenetic tree leaf is decorated below the heatmap with rpL19 cluster size (0 - 524 sequences). Cosmopolitan populations (detected in at least 50% of both tongue and plaque metagenomes) are highlighted in purple, and the star represents a Streptococcus population with a cluster size of 524.

Figure 3.16 continued: In Panel (B), the data from panel (A) is ordered by co-occurrence of each rpL19 across the metagenomes. Panel (C) shows the metagenomic read recruitment results of 11 genomes associated with cosmopolitan populations in (A) across tongue and plaque metagenomes identified by reclustering rpL19 sequences at 98% nucleotide similarity. Concentric circles represent metagenomes from (A) and are colored if the genome was detected (anvi'o detection = 50%). A star denotes if the genomes were confirmed to be cosmopolitan (detected in both tongue and plague metagenomes). Genome detection statistics are summarized in Table SI4.

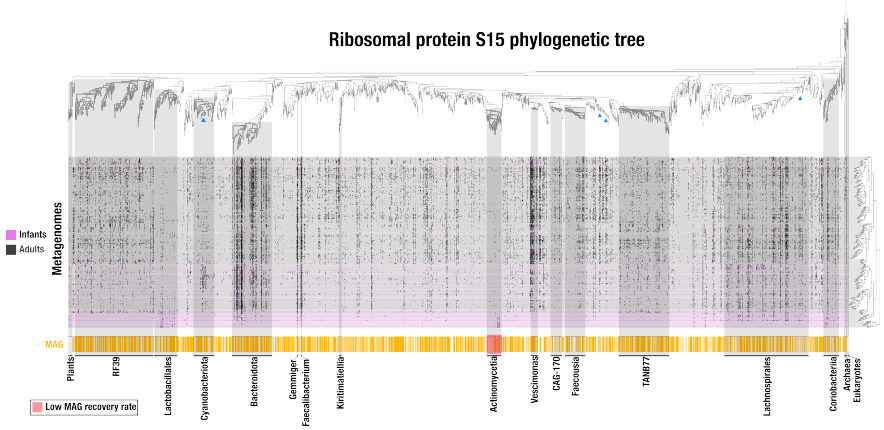

Figure 3.17: **Ribosomal S15 phylogeny and detection patterns across 388 gut microbiomes from Carter et al. (2023).** Heatmap represents rpS15 phylogeography across the Carter et al. (2023) metagenomic dataset. Each column represents a rpS15 DNA sequence, each row represents a metagenome for adult (black) or infant (pink) tribe members, and each data point is the presence/absence of rpS15 across metagenomes. Rows (metagenomes) are clustered by detection of rpS15 across metagenomes via metagenomic read recruitment and columns are organized by a Ribosomal S15 protein tree. Each leaf of the phylogenetic tree is decorated below the heatmap, denoting if the detected populations contains a MAG (yellow). Taxa with low MAG recovery rates are shaded in salmon.

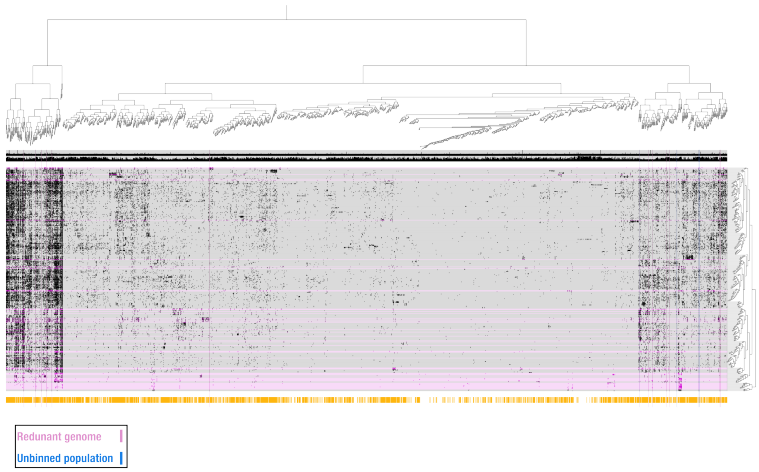

Figure 3.18: **Presence/absence heatmap of rpS15 Carter et al., 2024 Hadza gut metagenomes.** This figure is derived for Figure 2B, where the columns are reordered by a dendrogram derived from hierarchical clustering of rpS15 sequences based on co-occurrence across the metagenomic dataset. Each row of the detection matrix represents a metagenome (Black = Adult; Pink = Infant). Sequences found in genome dereplication clusters or unbinned are marked with pink or blue lines, respectively.

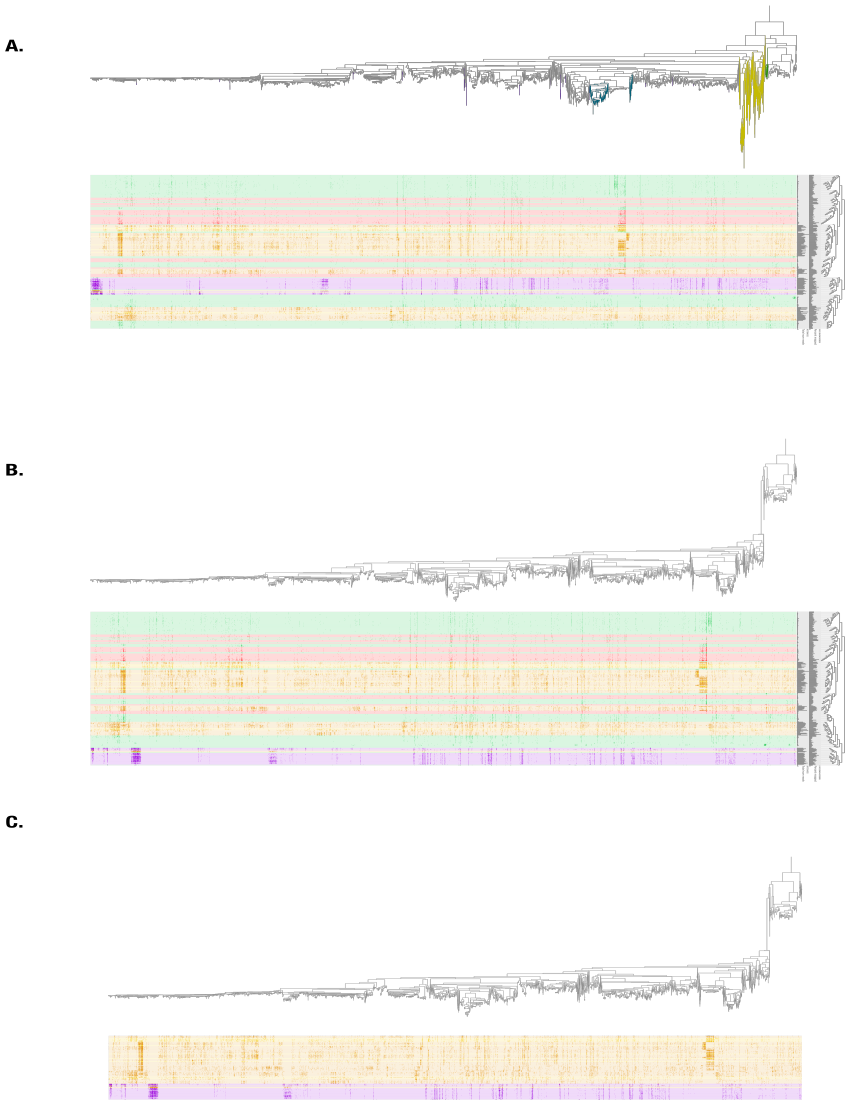

Figure 3.19: **Global surface ocean Ribosomal L14 protein data curation.** (A) Original phylogenetic tree output from EcoPhylo workflow. We manually curated the ribosomal protein phylogenetic tree by removing the mitochondria signal (yellow) and spurious branches (purple) derived from the metagenomic assemblies. Each row of the heatmap is a metagenome which are clustered by microbial community composition distance. Additionally, each row is decorated with bar graphs on the right that denote the number of reads that mapped rom the metagenome that mapped to the rpL14 sequences (0-507,356,552) and the percent mapped (0-100%). (B) Manually curated tree with mitochondria signal and spurious branches removed and all metagenomic samples. (C) Metagenomic samples with low sequencing depth were removed from the analysis because they did not reflect accurate biogeographic signals with and added low amounts of microbial diversity.

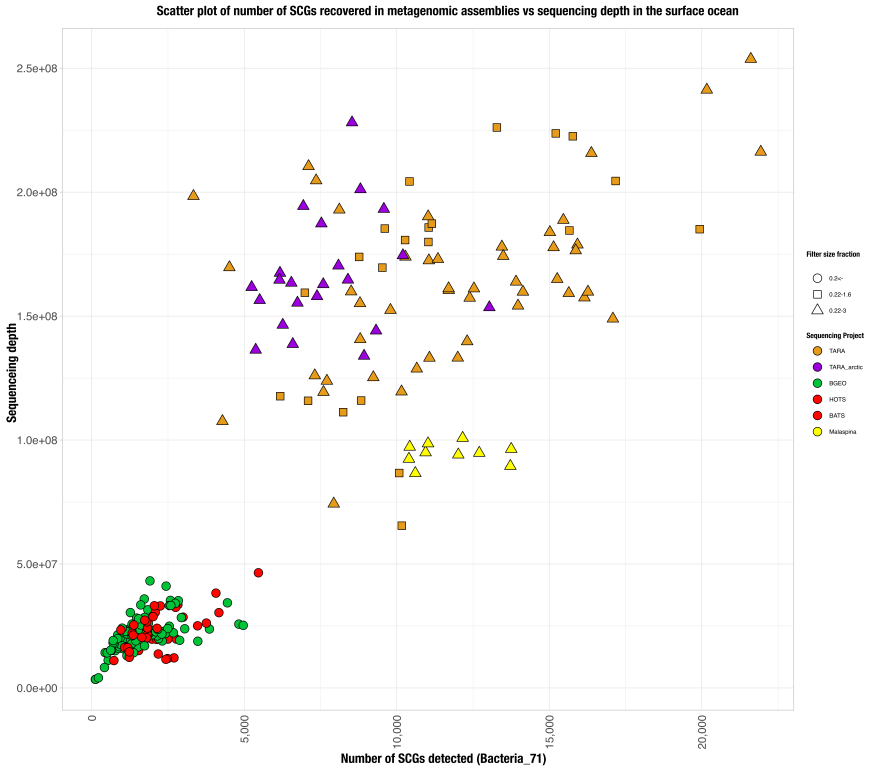

Figure 3.20: **Scatter plot of sequencing depth vs number of SCGs detected in assembled metagenomes from 237 surface ocean metagenomes.** Shapes designate the filter size and colors refer to the sequencing project.

CHAPTER 4

EXPANDED APPLICATIONS OF GENE FAMILY PHYLOGEOGRAPHY IN MICROBIOLOGY

4.1 Preface

In the previous chapter, I demonstrated that tracking the biogeography and phylogenetics of ribosomal proteins can yield *de novo* insights into microbial ecology and be used to assess genome recovery rates across different methods, including isolate genomes, MAGs, and SAGs. I also described the implementation of the EcoPhylo workflow to scale ribosomal protein phylogeography from individual metagenomic sequencing projects to entire biome-level genome and metagenome collections e.g. in the global surface ocean. To expand the applications of this framework, I designed the EcoPhylo workflow in a modular fashion to explore any protein family phylogenies.

Examining the phylogenetics of a protein family can yield insights into protein evolution, particularly when expanding your search for homologs outside of genome collections into environmental metagenomic assemblies. For example, environmental metagenomes have been surveyed for rhodopsin proteins from the phylum Saccharibacteria to explore protein family diversity and the exchange of retinol between these episymbionts and their hosts (Jaffe et al., 2022). EcoPhylo empowers analysis like these because (1) its scalability allows for the processing of massive genome and metagenomic datasets and (2) the workflow can use any protein family HMM to identify homologs. In the first section of this chapter, I demonstrate how I used EcoPhylo to efficiently uncover evolutionary insights into a protein family of respiratory reductases and their coevolution with substrates by extracting hundreds of homologs from reference genomes of the human gut microbiome and analyzing their phylogenetic clade structure. (Little et al., 2024). Next, I highlight a unique application of EcoPhylo to efficiently detect genomes in metagenomic data (Veseli et al., 2024a). Finally, I discuss future applica-

tions of EcoPhylo to profile other protein families explored in Schroer (2023) to expand novel functional annotations of model organisms to the environment.

4.2 Investigation into evolutionary trajectories of flavin reductases reveals distinct active sites with related electron acceptors

This section is derived from the following publication:

Little, A. S., Younker, I. T., **Schechter, M. S.**, Bernardino, P. N., MÃľheust, R., Stemczynski, J., Scorza, K., Mullowney, M. W., Sharan, D., Waligurski, E., Smith, R., Ramanswamy, R., Leiter, W., Moran, D., McMillin, M., Odenwald, M. A., Iavarone, A. T., Sidebottom, A. M., Sundararajan, A., ... Light, S. H. (2024). Dietary- and host-derived metabolites are used by diverse gut bacteria for anaerobic respiration. Nature Microbiology, 9(1), 55âĂŞ69. https://doi.org/10.1038/s41564-023-01560-2

4.2.1 Abstract

Respiratory reductases enable microorganisms to use molecules present in anaerobic ecosystems as energy-generating respiratory electron acceptors. Here, we identify three taxonomically distinct families of human gut bacteria (*Burkholderiaceae*, *Eggerthellaceae*, and *Erysipelotrichaceae*) that encode large arsenals of tens to hundreds of respiratory-like reductases per genome. Screening species from each family (*Sutterella wadsworthensis*, *Eggerthella lenta*, and *Holdemania filiformis*), we discover 22 metabolites used as respiratory electron acceptors in a species-specific manner. Identified reactions transform multiple classes of dietary- and host-derived metabolites, including bioactive molecules resveratrol and itaconate. Products of identified respiratory metabolisms highlight poorly characterized compounds, such as the itaconate-derived 2-methylsuccinate. Reductase substrate profiling defines enzyme–substrate pairs and reveals a complex picture of reductase evolution, provid-

ing evidence that reductases with specificities for related cinnamate substrates independently emerged at least four times. This study establishes an exceptionally versatile form of anaerobic respiration that directly links microbial energy metabolism to the gut metabolome.

4.2.2 Introduction

Heterotrophic cellular respiration is defined by the oxidation of an electron donor and the transfer of electrons through an electron transport chain to a terminal electron acceptor (Moodie and Ingledew, 1990). This is leveraged to form an ion gradient that powers oxidative adenosine triphosphate (ATP) synthesis through ATP synthase. While oxygen is the classic respiratory electron acceptor, microorganisms residing in oxygen-poor environments possess respiratory metabolisms that use alternative electron acceptors. In fact, these alternative electron acceptors, such as carbon, nitrogen, and sulfur chemical species, play roles in global biogeochemical processes.

Fermentative metabolisms are the primary source of energy conservation for the microbiome in the human gut. However, some classic anaerobic respiration pathways have also been identified in the human gut. For example, acetogens and methanogens leverage carbon dioxide as a terminal electron acceptor (Smith et al., 2019). Interestingly, inflammation states in the human gut stimulate the production of terminal electron acceptors from the microbiome (Hydrogen peroxide and tetrathionate) and the immune system, which produces nitrate. This allows both commensal and pathogens to use respiration as a source of energy conservation (Winter et al., 2010).

The majority of research into anaerobic cellular respiration focuses on classic inorganic soluble and insoluble (Aguilar and Nealson, 1994) electron acceptors. However, recent evidence has shown that taxa such as *Eggerthella lenta* the neurotransmitter dopamine as a terminal electron acceptor (Maini Rekdal et al., 2020). This recent find opens doors to new questions in the metabolism landscape of the human gut microbiome and how anaerobic respiration plays

a part.

In this study, Little et al. (2024) used a genome-mining approach to explore the usage of respiratory electron acceptors in the human gut microbiome. As a result, we identified three taxonomically distinct families of gut microbes that encode unusually high numbers of respiratory-like reductases, in some cases, over 100 paralogues. To further explore this, I performed a pangenome analysis on publicly available isolate genomes to compare respiratory reductase gene content between genomes, highlighting how multiple genomes from the same taxa contain vast arsenals of these paralogues.

Using reductase substrate profiling through a combination of growth assays and RNAseq Little et al. (2024) identified enzyme-substrate *Sutterella wadsworthensis*, *Eggerthella lenta*, and *Holdemania filiformis* which included multiple cinnamates, flavonoids, four-carbon dicarboxylates, sulfoxides with electron-accepting groups including carbonâĂŞoxygen bonds, hydroxyls, sulfoxides and alkenes. To further explore the evolution of reductases and their different active-site architectures in the article, I leveraged EcoPhylo to explore the clade structure of a reductase phylogeny as it relates to paired substrates and perform protein structure alignments to identity differences in active-site residue content. We demonstrate that distinct evolutionary pathways have produced enzymes with similar functions through different mechanisms. These findings highlight a unique mode of bacterial respiration characterized by the flexible use of organic electron acceptors and shed light on how energy metabolism influences the composition of the gut metabolome.

4.2.3 Results

E. lenta, *S. wadsworthensis*, and *H. filiformis* exhibit dynamic paralogue expansion of respiratory reductases

To evaluate reductase evolutionary dynamics, I performed a 'reductase pangenome' analysis, which assessed the variability in the content of paralogous flavin reductases across representative species *E. lenta*, *S. wadsworthensis*, and *H. filiformis*. For all three species, we observed that individual strains differed in their reductase content, containing a 'core' set of reductases, which were present in all strains, and an 'accessory' set of reductases present in a subset of strains (Figure 4.1). The *E. lenta* reductase pangenome included the largest number of genomes (Supplementary Table 4.1) and revealed the most intricate picture of reductase evolution (23 core and 86 accessory reductases). Prevalence of accessory reductases also varied, ranging from present in a single *E. lenta* strain to absent in a single strain. These results provide evidence of the continued dynamism of reductase evolution and suggest the selective advantage conferred by some reductases is strain-dependent.

Flavin reductases exhibit evolutionary complexity

We next explored the relationship between reductase evolution and substrate specificity. We reasoned that new reductase activities could be acquired either by (1) horizontal gene transfer or (2) gene duplication followed by functional divergence, and that these two scenarios would lead to distinct relationships between reductase sequence and substrate specificity. To investigate this, we constructed a phylogenetic tree using flavin reductases from *E. lenta*, *S. wadsworthensis*, and *H. filiformis* (Figure 4.2).

Several branches in the resulting tree include reductases from multiple species. For example, urocanate reductases from the three species are monophyletic and thus presumably share a more recent common evolutionary history (Figure 4.2). However, a majority of branches on

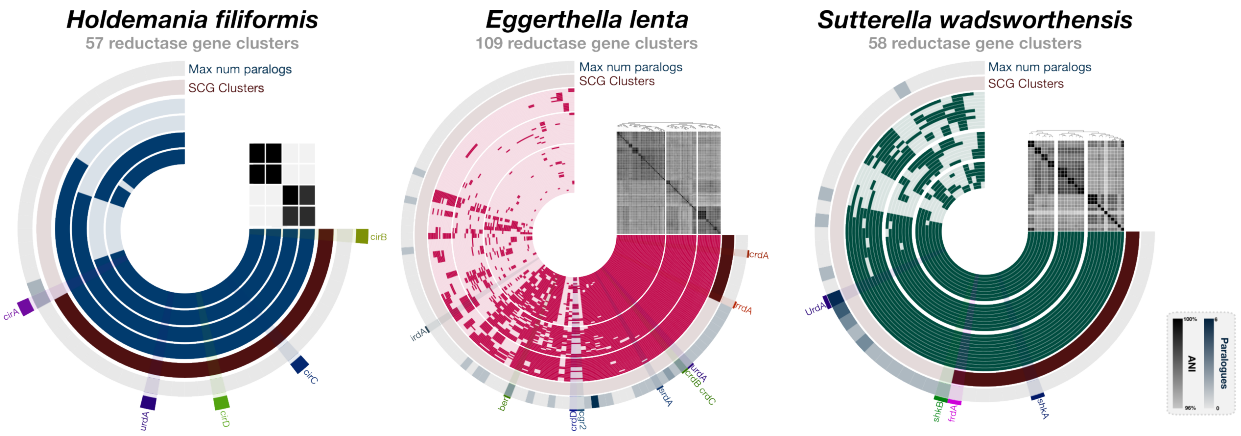

Figure 4.1: **Flavin reductase pangenomes of*E. lenta*, *S. wadsworthensis*, and *H. filiformis*.** For each pangenome, the inner concentric layers represent unique genomes while the radial elements represent gene cluster presence (darker color) or absence (lighter color) across the genomes. The outermost concentric circle, 'Max number paralogues,' indicates the maximum number of paralogues (defined as reductases with 60% sequence identity) one genome contributes to the gene cluster. The second outermost circle, 'SCG clusters,' indicates single-copy core reductase that is, gene clusters for which every genome contributed exactly one gene. Genomes (inner concentric layers) are clustered by the presence/absence of reductase gene clusters. All vs all genome average nucleotide identity is depicted in the heat map above the genome concentric layers.

the tree are highly sub-branched and exclusively contain reductases from a single genus. These patterns suggest that both horizontal gene transfer and gene duplication may have played substantial roles in reductase evolution within bacterial lineages.

The distribution of reductases with different activities in the tree suggests a convoluted evolutionary history, with cinnamate reductases providing a striking example of the complex relationship between amino acid sequence and substrate specificity. We observed that, despite catalyzing highly similar reactions, the eight *E. lenta* and *H. filiformis* reductases induced in the presence of different cinnamate substrates separated into four phylogenetically distinct reductase clades (Figure 4.2). Identified cinnamate reductases typically share >30% amino acid sequence identity with other reductases within their clade, including reductases with distinct substrate specificities. For example, *H. filiformis* cinnamate reductase CirD and urocanate reductase UrdA are both from 'reductase clade 1' and share 32% sequence identity (Figure 4.2).

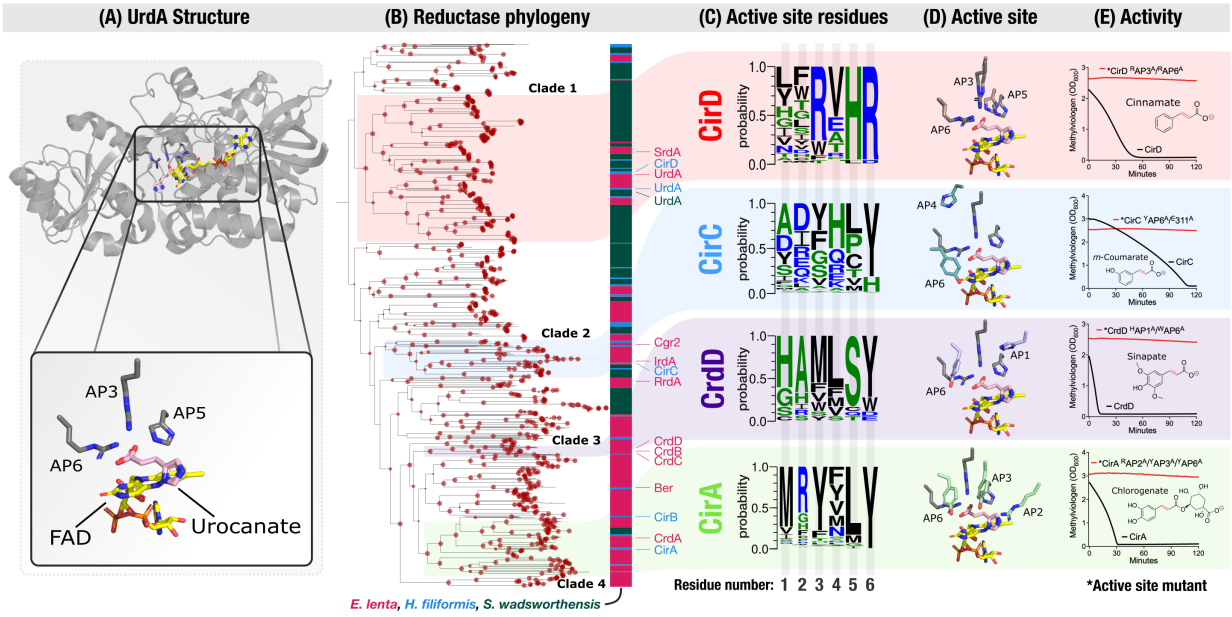

Figure 4.2: **Independent evolutionary trajectories and distinct active sites distinguish flavin reductases with related electron acceptors.** UrdA structure–previously published crystal structure of the *S. oneidensis* UrdA in complex with urocanate and flavin adenine dinucleotide (FAD) (PDB code 6T87). Reductase phylogeny–phylogenetic tree of flavin reductases from *E. lenta*, *S. wadsworthensis*, and *H. filiformis* genomes. Bootstrap support values are indicated by the size of the red dots at nodes of the tree and range from 70 to 100. Active-site residues–representation of the sequence identity of active-site amino acids in reductase clades 1–4 scaled to frequency within the multiple sequence alignment. Positions within the multiple sequence alignment have been renumbered to active-site alignment position (AP). Active site–AlphaFold models of CirD, CirC, CrdD, and CirA cinnamate reductases superimposed on the UrdA crystal structure. Activity–reductase activity of CirD, CirC, CrdD, and CirA and active-site point mutants. Active-site mutations and alignment positions they correspond to: CrdD H313A (AP1) and W510A (AP6); CirA R417A (AP2), Y469A (AP3), and Y634A (AP6); CirC E311A and Y511A (AP6); CirD R542A (AP3) and R716A (AP6). The y-axis shows the amount of reduced methyl viologen in the presence of the indicated electron acceptor.

By contrast, the cinnamate reductases from different reductase clades share <26% sequence identity, despite using similar substrates (Figure 4.2). The observations suggest that flavin reductases with similar cinnamate substrate specificities independently evolved at least four times.

Dissimilar reductase active sites catalyze related reactions

To clarify how different evolutionary trajectories may have independently generated reductases with similar cinnamate substrate specificities, we turned to previous mechanistic studies of enzymes from the flavin reductase superfamily. Fumarate and urocanate reductases were previously shown to contain two conserved active-site arginines including one that forms a critical salt bridge with the substrate carboxyl group and a second proposed to facilitate catalytic proton transfer to reduce the substrate enoate. As a similar substrate enoate group is reduced by cinnamate reductases, we compared the substrate-bound urocanate reductase crystal structure with AlphaFold (Jumper et al., 2021) structural models of cinnamate reductases from the four reductase clades (Figure 4.2). We found that active-site arginines were conserved in reductase clade 1, including in the UrdA urocanate reductases and the CirD cinnamate reductase (Figure 4.2, alignment positions 4 and 6). By contrast, arginines were not conserved at the same positions in reductase clades 2–4. Instead, tyrosine was conserved at alignment position 6 and other active-site amino acids exhibited variable, clade-specific patterns of conservation (Figure 4.2).

To test whether the distinct patterns of active-site conservation in the four reductase clades might reflect mechanistic distinctions, Little et al. (2024) generated point mutants that targeted conserved amino acids in representative cinnamate reductases (CirA, CirC, CirD and CrdD) from reductase clades 1–4. In each case, we found that conserved clade-specific active-site amino acids were essential for activity (Figure 4.2). These results thus show that distinct active-site architectures functionally distinguish the cinnamate reductase clades and suggest

that parallel evolutionary processes produced flavin reductases with similar substrate specificities but substantial functional distinctions.

4.2.4 Discussion

In this study, we identify gut microbiome members characterized by a high count of respiratory-like reductases in their genomes. Our findings indicate that these bacteria engage in respiratory metabolism and highlight organic respiratory electron acceptors that show variability in usage depending on the strain. Although many aspects of these functions are yet to be investigated, our results imply that a unique form of respiration, marked by the flexible utilization of various organic respiratory electron acceptors, could be significant in the gut environment.

Interestingly, the bacteria identified in our studies possess reductase arsenals (often >50 reductases per genome) that stand apart from previously characterized (non-host associated) bacterial respiratory specialists. For example, the fresh-water-inhabiting bacterium *Shewanella oneidensis* is widely cited as a model organism with exceptionally broad respiratory capabilities, but possesses 'only' 22 molybdopterin and flavin reductases (Ikeda et al., 2021; Heidelberg et al., 2002).

The majority of known respiratory electron acceptors utilized by non-host-associated respiratory specialists, such as S. oneidensis, are small inorganic compounds. In contrast, the electron acceptors identified in this study are predominantly organic metabolites. This suggests that the extensive expansion of reductase enzymes in gut bacteria may reflect the need for enzymatic diversity to process the chemically complex and abundant metabolite electron acceptors found in the gut environment from the human diet. These findings point to ecosystem-specific ecological factors driving the evolution of distinct respiratory strategies: (1) a broad reductase repertoire enabling the respiration of diverse organic metabolites, versus (2) a more limited reductase arsenal specialized for the respiration of fewer inorganic compounds.

In this article, I demonstrate that EcoPhylo can be a powerful workflow to explore the phy-

logenetics of a protein family. The combination of phylogenetics and computational structural biology was able to guide targeted experimentation in the lab, confirming respiratory reductase active site residues. Future studies can leverage this computational biology - lab integration strategy to perform more targeted biochemical experiments to confirm protein function. Given that most reductases encoded by gut bacteria remain functionally uncharacterized, the identified metabolic pathways likely represent only a fraction of the interactions between respiratory reductases and the gut metabolome. Future directions into respiratory electron acceptors in the human gut could significantly enhance our understanding of the functional capabilities and metabolic outputs of the gut microbiome leading to novel therapeutics.

4.2.5 Material and Methods

Reductase pangenomics

Pangenomes for *E. lenta*, *S. wadsworthensis* and *H. filiformis* were calculated using anvi'o 7.1 (with parameters anvi-pan-genome –mcl-inflation 10) (Delmont and Eren, 2018). Briefly, the program (1) performed an all-versus-all National Center for Biotechnology Information (NCBI) BLAST (Altschul et al., 1990) to create a sequence similarity network, (2) used the Markov Cluster algorithm to resolve gene clusters and (3) was visualized with 'anvi-display-pan' (Altschul et al., 1990). Subsequently, the species pangenomes were subsetted for gene clusters annotated with the Pfam PF00890 using 'anvi-split' (Eren et al., 2021).

Reductase phylogenetics

Genomes for *E. lenta*, *S. wadsworthensis* and *H. filiformis* were downloaded using ncbi-genome-download (https://github.com/kblin/ncbi-genome-download; see Supplementary Table 4.1 for genome accessions). We then used anvi'o v7.1 to convert genome FASTA files into contig databases

(https://anvio.org/m/contigs-db) using the contig workflow (https://doi.org/10.1186/s13059-020-02195-w), during which Prodigal v2.6.3 identified open reading frames (contigs workflow) (Hyatt et al., 2010; Eren et al., 2021). Next, we used the EcoPhylo workflow implemented in anvi'o (https://anvio.org/m/ecophylo) in âĂŸtree-modeâĂŹ to recover reductase genes in genomes using the Pfam model PF00890 and to explain their phylogeny. Briefly, the EcoPhylo workflow (1) used the program hmmsearch in HMMER v3.3.2 (Eddy, 2011) to identify reductases, (2) removed hits that had less than 80% model coverage to minimize the likelihood of false positives due to partial hits, (3) dereplicated resulting sequences using MMseqs2 13.45111 (Steinegger and Soding, 2017) to avoid redundancy, (4) calculated a multiple sequence alignment with MUSCLE v3.8.1551 (Edgar, 2004) and trimmed the alignment by removing columns of the alignment with trimal v1.4.rev15 (with the parameters âĂŸ–gappyoutâĂŹ)(Capella-Gutiérrez et al., 2009), (5) removed sequences that have more than 50% gaps using the anvi'o program anvi-script-reformat-fasta, (6) calculated a phylogenetic tree with the final alignment with IQ-TREE 2.2.0-beta COVID-edition (Minh et al., 2020) (with the parameters 'nt AUTO -m WAG -B 1000') that resulted in a NEWICK formatted tree file and finally (7) visualized the tree in the anvi'o interactive interface.

Reductase active-site conservation analyses

AlphaFold models of CirA, CirC, CirD and CrdD were downloaded from Uniprot (Jumper et al., 2021). Active-site amino acids were identified by independently superimposing N- and C-terminal domains of AlphaFold models to the substrate-bound urocanate reductase crystal structure (PDB code 6T87) using PyMOL v2.5.1 (http://www.pymol.org/pymol) 'align' (Venskutonytė et al., 2021). Next, sequences in the monophyletic clade surrounding the experimentally validated sequences were subsetted from the reductase multiple sequence alignment using the program anvi-script-reformat-fasta and alignment positions were sliced in Jalview (v2.11.2.5) (Waterhouse et al., 2009). Finally, the extent of conservancy among the active-

site-associated residues was visualized with WebLogo 3 https://weblogo.threeplusone.com/create.cgi.

4.2.6 Supplementary tables

This section's supplementary tables are accessible at doi: 10.1038/s41564-023-01560-2

Table 4.1: E. lenta, S. wadworthensis, and H. filiformis genomes used for reductase phylogenetics (Table S14).

4.3 Application of EcoPhylo to quickly survey genomes in metagenomes

This section is derived from the following publication:

Veseli, I., Chen, Y. T., **Schechter, M. S.**, Vanni, C., Fogarty, E. C., Watson, A. R., Jabri, B., Blekhman, R., Willis, A. D., Yu, M. K., Fernández-Guerra, A., FÃijssel, J., & Eren, A. M. (2023). Microbes with higher metabolic independence are enriched in human gut microbiomes under stress. eLife, 12, RP89862. https://doi.org/10.7554/eLife.89862

Reduced microbial diversity in the gut is linked to numerous inflammatory diseases. However, the ecological drivers of this decline in microbial richness remain unclear, complicating our understanding of the microbiota's role in disease. To investigate this, Veseli et al. (2024a) explored the hypothesis that environmental stresses from host inflammation in various GI disease states select for microbes with high metabolic independence (HMI) (Watson et al., 2023). This refers to microbes with larger genomes that harbor an increased number of biosynthetic metabolic modules, enabling them to thrive with minimal reliance on cross-feeding in less diverse microbiomes.

Veseli et al. (2024b) systematically analyzed human gut metagenomes and found that microbes with higher HMI indexes were more prevalent in gut metagenomes of patients with Inflammatory Bowel Disease (IBD). Testing this hypothesis required a large collection of gut microbial genomes, but no publicly available gut-specific genome collections were sufficiently large. To address this, Veseli et al. (2024a) leveraged EcoPhylo to subset 19,226 genomes from the Genome Taxonomy Database (GTDB) (Parks et al., 2018), focusing on three major gut-associated phyla (Bacteroidetes, Firmicutes, and Proteobacteria (Woting and Blaut, 2016; Turnbaugh et al., 2008)) by profiling their presence in 150 healthy gut metagenomes from the Human Microbiome Project (HMP) (asd, 2012). Standard approaches for identifying if a genome is detected in a metagenomic dataset would require performing metagenomic read recruitment across the whole genome. However, scaling this up to thousands of genomes

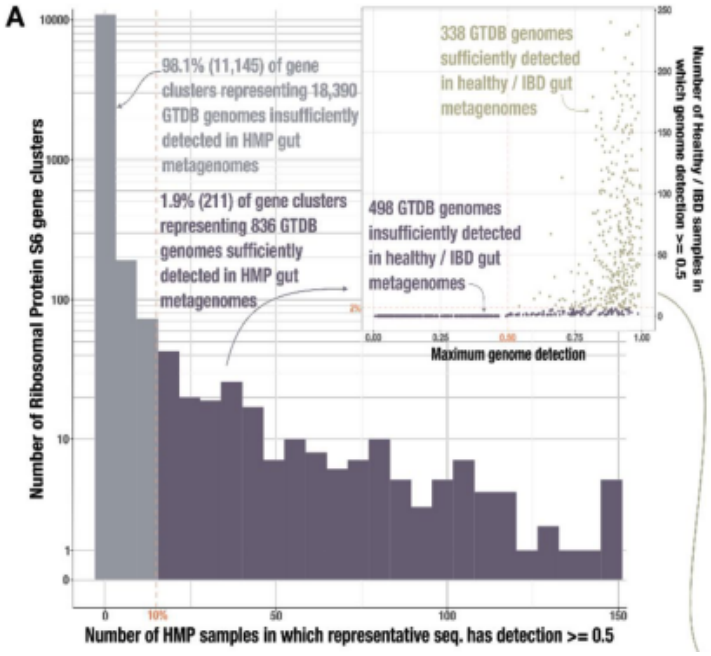

Figure 4.3: **Identification of HMI genomes and their distribution across gut samples.** A) Histogram of Ribosomal Protein S6 gene clusters (94% ANI) for which at least 50% of the representative gene sequence is covered by at least 1 read (>= 50% 'detection') in fecal metagenomes from the Human Microbiome Project (HMP) (Human Microbiome Project Consortium 2012). The dashed line indicates our threshold for reaching at least 50% detection in at least 10% of the HMP samples; gray bars indicate the 11,145 gene clusters that do not meet this threshold while purple bars indicate the 836 clusters that do. The subplot shows data for the 836 genomes whose Ribosomal Protein S6 sequences belonged to one of the passing (purple) gene clusters. The y-axis indicates the number of healthy/IBD gut metagenomes from our set of 330 in which the full genome sequence has at least 50% detection, and the x-axis indicates the genome's maximum detection across all 330 samples. The dashed line indicates our threshold for reaching at least 50% genome detection in at least 2% of samples; the 338 genomes that pass this threshold are tan and those that do not are purple.

would be computationally expensive. By contrast, EcoPhylo can use individual ribosomal proteins to act as a proxy for the presence of a whole genome in a metagenome, thereby dramatically improving computational efficiency. Veseli et al. (2024a) thus used EcoPhylo and Ribosomal Protein S6 to profile the presence 19,226 input genomes in HMP. This resulted in a final subsetted list of 338 reference genomes (Figure 4.3). The subsetted genomes detected with EcoPhylo in HMP included 258 (76.3%) Firmicutes, 60 (17.8%) Bacteroidetes, and 20 (5.9%) Proteobacteria. Most of these genomes resolved to families common to the colonic microbiota, such as Lachnospiraceae (30.0%), Ruminococcaceae / Oscillospiraceae (23.1%), and Bacteroidaceae (10.1%) (Arumugam et al. 2011), while 5.9% belonged to poorly-studied families.

4.3.1 Material and Methods

Identification of gut microbial genomes from the GTDB

We took 19,226 representative genomes from the GTDB species clusters belonging to the phyla Firmicutes, Bacteroidetes, and Proteobacteria, which are most common in the human gut microbiome (Woting and Blaut 2016). To evaluate which of these genomes might represent gut microbes in a computationally-tractable manner, we ran the anviâĂŹo âĂŸEcoPhyloâĂŹ workflow (https://anvio.org/m/ecophylo) to contextualize these populations within 150 healthy gut metagenomes from the Human Microbiome Project (HMP) (asd, 2012). Briefly, the EcoPhylo workflow (1) recovers sequences of a gene family of interest from each genome and metagenomic sample in the analysis, (2) clusters resulting sequences and picks representative sequences using mmseqs2 (Steinegger and Soding, 2017), and (3) uses the representative sequences to rapidly summarize the distribution of each population cluster across the metagenomic samples through metagenomic read recruitment analyses. Here, we used the Ribosomal Protein S6 as our gene of interest, since it was the most frequently-assembled

single-copy-core gene in our set of GTDB genomes. We clustered the Ribosomal Protein S6 sequences from GTDB genomes at 94% nucleotide identity.

To identify genomes that were likely to represent gut microbes, we selected genomes whose ribosomal protein S6 belonged to a gene cluster where at least 50% of the representative sequence was covered (i.e. detection >= 0.5x) in more than 10% of samples (i.e. n > 15). There are 100 distinct individuals represented in the 150 HMP gut metagenomes âĂŞ 56 of which were sampled just once and 46 of which were sampled at 2 or 3 time points âĂŞ so this threshold is equivalent to detecting the genome in 5% - 15% of individuals. From this selection we obtained a set of 836 genomes; however, these were not exclusively gut microbes, as some non-gut populations have similar ribosomal protein S6 sequences to gut microbes and can therefore pass this selection step. To eliminate these, we mapped our set of 330 healthy and IBD metagenomes to the 836 genomes using the anvi'o metagenomics workflow, and extracted genomes whose entire sequence was at least 50% covered (i.e. detection >= 0.5x) in over 2% (n > 6) of these samples. Our final set of 338 genomes was used in downstream analysis.

Data availability

Accession numbers for publicly available data are listed in our Supplementary Tables at doi:10.6084/m9.figshare.22679080. Our Supplementary Files are also available at doi:10.6084/m9.figshare.22679080. Contigs databases of our assemblies for the 408 deeply-sequenced metagenomes can be accessed at doi:10.5281/zenodo.7872967, and databases for our assemblies of the (Palleja et al., 2018) metagenomes can be accessed at doi:10.5281/zenodo.7897987. Contigs databases of the 338 GTDB gut reference genomes are available at doi:10.5281/zenodo.7883421.

4.4 Future directions of the EcoPhylo

In this chapter of my thesis, I demonstrated the versatility of EcoPhylo in addressing various biological questions, including the investigation of a reductase protein family. A notable example is the work by Little et al. (2024), which combined genome mining across the tree of life with experimental validation to investigate the evolution of a specific protein family. This strategy can be implemented with protein family and other protein expression systems. A potential future project could investigate other protein families across a large collection of genomes e.g. examining the phylogenetic distribution of mucin degradation enzymes and testing their ability to degrade mucins in different animal species.

Additionally, Veseli et al. (2024a) cleverly applied EcoPhylo to significantly reduce the computational time required for detecting genomes within metagenomes. This strategy could be applied to any project that seeks to quickly identify a subset of genomes from metagenomic surveys. For example, one potential application would be to leverage an isolate bank of commensal microbes to determine which strains are most prevalent across various human populations. This could help inform decisions regarding the development of future probiotics.

Another example of a future application of the EcoPhylo workflow is a study *in preparation* that creatively implements this approach to ABC transporter protein complexes.

This section is derived from the following publications *in prep*:

Schroer, W. F., **Schechter, M. S.**, Saunders, J. K., Eren, A. M., Moran, M. A. (2024) High-confidence global mapping of bacterial substrate uptake potential *in prep*. `http://proxy.uchicago.edu/login?url=https://www.proquest.com/dissertations-theses/metabolite-transport-role-marine-microbial/docview/2917419506/se-2.`

In the same light as ribosomal proteins, other gene families can act as proxies for biological phenomena. For example, measuring the abundance patterns of *nifH* in environmental genomes offers a dual advantage: it allows for tracking microbial diversity while identifying

organisms with the genetic potential for nitrogen fixation. Expanding on this concept, the phylogeography of other protein families could offer insights into the biogeography of microbial functions. Further, tracking microbial proteins involved in substrate transport could enable indirect monitoring of those substrates using environmental metagenomic surveys.

In Schroer (2023), I used EcoPhylo to track the ATPase subunit of a taurine ABC transporter to identify regions in the ocean with high taurine flux, a metabolite known for its cryptic flux. Cryptic flux refers to metabolites that are difficult to measure directly due to low concentrations or rapid consumption by surrounding microbes (Durham et al., 2014). For instance, 2,3-dihydroxypropane-1-sulfonate (DHPS) in the North Pacific is rapidly consumed, making it challenging to detect with conventional analytical methods (Durham et al., 2014). A potential workaround is to measure the presence of a protein responsible for transporting the metabolite into cells in that environment. The success of this application of EcoPhylo in Schroer (2023) was enabled by the wet lab experiments in Schroer et al. (2023), which functionally annotated ABC transporters in the model pelagic heterotroph *Ruegeria pomeroyi*. By combining functional annotations in model organisms with environmental metagenomic surveys, this workflow provides a powerful approach to linking proteins to their environmental roles in global biogeochemical cycles.

CHAPTER 5

CONCLUSIONS

5.1 Summary of contributions

Microbiome research is critical because of its impact on human health, ecosystems, and global processes. As the cost of DNA sequencing has plummeted (x < $1000 to sequence the human genome), it is now possible to sequence incredible amounts of environmental genomes and metagenomes, enabling researchers to generate unprecedented volumes of 'omics data. To handle this influx of data, innovative strategies are required to analyze, interpret, and store these datasets effectively. In Chapter 2, I detail my contributions to the open-source 'omics platform anvi'o to empower microbiologists in processing and interpreting large-scale data. These contributions range from standalone bioinformatics tools like 'anvi-export-locus' to scalable computational workflows capable of processing terabytes of data, such as the anvi'o EcoPhylo workflow. Beyond developing software, I prioritized improving accessibility to bioinformatics by providing comprehensive documentation and tutorials. These efforts reflect the ethos of my PhD work: making computational biology more accessible to the broader scientific community.

The rapid growth of genome collections in recent years has further highlighted the need for new methods to address gaps in microbial ecology research. In Chapter 3, I argue that studying microbial genomic data beyond final genome collections is essential. This thesis demonstrates that the phylogeography of gene families can offer valuable insights into microbial ecology, extending our understanding beyond individual genomes to metagenomic assemblies.

In Chapter 3, I showcase the application of the anvi'o EcoPhylo workflow to ribosomal proteins to track genome recovery across microbiomes from the human oral cavity, gut, and surface ocean. This analysis revealed discontinuities between genome collections and micro-

bial populations in the environment. My research provides a framework for microbiologists to make informed decisions about the most effective methods for recovering specific taxa in particular environments. Future microbiome sampling projects can use the EcoPhylo workflow to assess the phylogenetic and functional diversity of microbial communities, identify gaps in genome collections, and prioritize the recovery of genomes from underrepresented taxa. By tailoring recovery strategies to specific ecological functions or conserved markers, microbiologists can optimize sampling, assembly, and annotation workflows to enhance genome recovery. This targeted approach accelerates the discovery of novel microbial lineages and improves our understanding of their ecological roles. Furthermore, microbiologists can quantify prior genome recovery rates to justify selecting a specific genome recovery method in future grant applications.

In Chapter 4, I showcase the modular nature of the EcoPhylo workflow by extending its application to alternative gene families beyond ribosomal proteins. These instances highlight the workflow's adaptability in revealing new insights into microbial ecology and protein evolution, further illustrating its potential for widespread utilization in the field. Importantly, I demonstrate how EcoPhylo facilitates the examination of evolutionary patterns in flavin reductase protein families and effectively identifies genomes within metagenomic datasets.

This thesis collectively showcases how innovations in bioinformatics data analysis tools and workflows can reveal even greater insights from publicly available microbial 'omics data.

5.2 Concluding remarks and perspectives

My Ph.D. training experience has equipped me with a valuable skill set in developing scientific software and workflows alongside my research. The skills I've honed in scientific programming, data analysis (particularly with terabytes of data), data visualization, engineering, and creative thinking have deeply influenced my approach to science. Through this journey, I have gained a deep appreciation for the complexity and effort required in both software development and

the pursuit of reproducible open science. There were times when I wondered how much quicker I could have completed my projects had I not developed bioinformatics software as a byproduct of my work. However, upon reflection, I realized that the value added to the scientific community lies not in completing projects faster but in the lasting impact of science that is reproducible, scalable, and accessible. By providing software, documentation, and tutorials, I believe I've contributed to making science more accessible in addition to reporting my scientific findings – I hope to instill this principle in my future scientific endeavors. Beyond the technical training of this Ph.D., I have built resilience and productivity skills that will last me a lifetime. Through extensive independent thinking, writing, and programming (what's the difference?), I've built the confidence to tackle any challenge. Even challenges as complex as microbial ecology.

At the start of my Ph.D., I argued in a blog post that microbiology was transitioning from a genome-centric phase (Metagenomics 2.0) to a high-resolution phase (Metagenomics 3.0). This new phase would emphasize innovation in data analysis techniques to explore the growing genomic collections in unprecedented detail (Schechter, 2020). To contribute to this transition, I developed the EcoPhylo workflow to better analyze, benchmark, and leverage the wealth of genomic data generated during Metagenomics 2.0. However, the ability to conduct such analyses on publicly available data hinges on the commitment of laboratories to make their data FAIR (Findable, Accessible, Interoperable, and Reusable) (Wilkinson et al., 2016). I strongly advocate for the continued adoption of open data and open science practices, especially as large language models increasingly leverage these datasets for advancements in artificial intelligence.

In Chapter 3 of my thesis, the EcoPhylo workflow provided valuable insights into the ecology and evolution of microbial communities in the global surface ocean. Compared to its application in human gut and oral cavity microbiome sequencing projects, the workflow was particularly effective at uncovering ecological patterns in the marine microbiome. This raises

an important question: why was the EcoPhylo workflow more successful in analyzing marine microbiomes?

One explanation may lie in the size and diversity of the dataset, which included genomes and metagenomes collected from across an entire biome, the global surface ocean. Applying EcoPhylo to this expansive and varied dataset allowed me to identify broad ecological trends, such as the ecology of SAR11 subtypes and patterns of cold adaptation across multiple clades. These results suggest that using similarly large-scale datasets for human gut and oral cavity microbiomes from diverse populations worldwide could yield equally profound insights into those habitats.

If a particular goal of microbiome research is to understand the fundamental principles of microbial ecology, it is essential to focus on environments that are both easily accessible for large-scale sampling and rich in diverse habitats that can be enriched with metadata. While the global ocean provides immense ecological insights, its sampling remains expensive and logistically challenging. Similarly, while the human gut microbiome is the most extensively studied, most samples are restricted to the endpoint of the large intestine, offering only a partial perspective of the broader microbial ecosystem of the GI tract. Innovations in sampling the biogeography of the gut microbiome could improve this effort (Mimee et al., 2018). In contrast, the human oral cavity represents a highly accessible target for large-scale microbiome studies. It encompasses multiple distinct habitats (e.g., teeth, tongue, cheeks) within a single sampling site and allows for the collection of rich metadata, including saliva composition, host genetics, immune factors, and diet. A global-scale human oral microbiome sequencing project could generate many new genomes and serve as a foundational study to unravel microbial ecological principles such as colonization, ecosystem stability, host-microbe interactions, and genomic adaptation.

This is an exciting time for microbiome research, as high-throughput genome recovery is becoming the norm. While I hope new projects continue to generate high-quality genomes and

explore the vast seas of microbial ecology, I also believe it is equally important to innovate data analysis techniques to extract even more insights from existing datasets. I envision the anvi'o EcoPhylo workflow playing a crucial role in benchmarking these expanding genome collections and enhancing our understanding of microbial ecology across diverse environments.

REFERENCES

(2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207âĂŞ214.

Aguilar, C. and Nealson, K. H. (1994). Manganese reduction in oneida lake, new york: Estimates of spatial and temporal manganese flux. *Can. J. Fish. Aquat. Sci.*, 51(1):185–196.

Al-Shayeb, B., Skopintsev, P., Soczek, K. M., Stahl, E. C., Li, Z., Groover, E., Smock, D., Eggers, A. R., Pausch, P., Cress, B. F., Huang, C. J., Staskawicz, B., Savage, D. F., Jacobsen, S. E., Banfield, J. F., and Doudna, J. A. (2022). Diverse virus-encoded CRISPR-Cas systems include streamlined genome editors. *Cell*, 185(24):4574–4586.e16.

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., and Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.*, 31(6):533–538.

Almeida, A., Mitchell, A. L., Boland, M., Forster, S. C., Gloor, G. B., Tarkowska, A., Lawley, T. D., and Finn, R. D. (2019). A new genomic blueprint of the human gut microbiota. *Nature*, 568(7753):499–504.

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K. S., Sakharova, E., Parks, D. H., Hugenholtz, P., Segata, N., Kyrpides, N. C., and Finn, R. D. (2020). A unified catalog of 204, 938 reference genomes from the human gut microbiome. *Nature Biotechnology*, 39(1):105–114.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Backer, R., Rokem, J. S., Ilangumaran, G., Lamont, J., Praslickova, D., Ricci, E., Subramanian, S., and Smith, D. L. (2018). Plant growth-promoting rhizobacteria: Context, mechanisms of action, and roadmap to commercialization of biostimulants for sustainable agriculture. *Front. Plant Sci.*, 9:1473.

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Li, Y., Xia, Y., Xie, H., Zhong, H., Khan, M. T., Zhang, J., Li, J., Xiao, L., Al-Aama, J., Zhang, D., Lee, Y. S., Kotowska, D., Colding, C., Tremaroli, V., Yin, Y., Bergman, S., Xu, X., Madsen, L., Kristiansen, K., Dahlgren, J., and Wang, J. (2015). Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe*, 17(5):690–703.

Baker-Austin, C., Trinanes, J., Gonzalez-Escalona, N., and Martinez-Urtaza, J. (2017). Non-cholera vibrios: The microbial barometer of climate change. *Trends Microbiol.*, 25(1):76–84.

Berendsen, R. L., Pieterse, C. M., and Bakker, P. A. (2012). The rhizosphere microbiome and plant health. *Trends in Plant Science*, 17(8):478âĂŞ486.

Biller, S. J., Berube, P. M., Dooley, K., Williams, M., Satinsky, B. M., Hackl, T., Hogle, S. L., Coe, A., Bergauer, K., Bouman, H. A., Browning, T. J., De Corte, D., Hassler, C., Hulston, D., Jacquot, J. E., Maas, E. W., Reinthaler, T., Sintes, E., Yokokawa, T., and Chisholm, S. W. (2018). Marine microbial metagenomes sampled across space and time. *Sci Data*, 5:180176.

Biller, S. J., Berube, P. M., Lindell, D., and Chisholm, S. W. (2015). Prochlorococcus: the structure and function of collective diversity. *Nat. Rev. Microbiol.*, 13(1):13–27.

Bowen, W. H., Burne, R. A., Wu, H., and Koo, H. (2018). Oral biofilms: Pathogens, matrix, and polymicrobial interactions in microenvironments. *Trends Microbiol.*, 26(3):229–242.

Bowers, R. M., The Genome Standards Consortium, Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., Tringe, S. G., Ivanova, N. N., Copeland, A., Clum, A., Becraft, E. D., Malmstrom, R. R., Birren, B., Podar, M., Bork, P., Weinstock, G. M., Garrity, G. M., Dodsworth, J. A., Yooseph, S., Sutton, G., Glöckner, F. O., Gilbert, J. A., Nelson, W. C., Hallam, S. J., Jungbluth, S. P., Ettema, T. J. G., Tighe, S., Konstantinidis, K. T., Liu, W.-T., Baker, B. J., Rattei, T., Eisen, J. A., Hedlund, B., McMahon, K. D., Fierer, N., Knight, R., Finn, R., Cochrane, G., Karsch-Mizrachi, I., Tyson, G. W., Rinke, C., Lapidus, A., Meyer, F., Yilmaz, P., Parks, D. H., Murat Eren, A., Schriml, L., Banfield, J. F., Hugenholtz, P., and Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.*, 35(8):725–731.

Brazelton, W. J. and Baross, J. A. (2009). Abundant transposases encoded by the metagenome of a hydrothermal chimney biofilm. *The ISME Journal*, 3(12):1420âĂŞ1424.

Brock, T. D. (1999). *Robert Koch : a life in medicine and bacteriology*. ASM Press.

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., Wilkins, M. J., Wrighton, K. C., Williams, K. H., and Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain bacteria. *Nature*, 523(7559):208–211.

Brown, M. V., Lauro, F. M., DeMaere, M. Z., Muir, L., Wilkins, D., Thomas, T., Riddle, M. J., Fuhrman, J. A., Andrews-Pfannkoch, C., Hoffman, J. M., McQuaid, J. B., Allen, A., Rintoul, S. R., and Cavicchioli, R. (2012). Global biogeography of SAR11 marine bacteria. *Mol. Syst. Biol.*, 8(1):595.

Buchfink, B., Reuter, K., and Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using diamond. *Nature Methods*, 18(4):366âĂŞ368.

Callahan, A., Winnenburg, R., and Shah, N. H. (2018). U-index, a dataset and an impact metric for informatics tools and databases. *Sci. Data*, 5:180043.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973.

Carter, M. M., Olm, M. R., Merrill, B. D., Dahan, D., Tripathi, S., Spencer, S. P., Yu, F. B., Jain, S., Neff, N., Jha, A. R., Sonnenburg, E. D., and Sonnenburg, J. L. (2023). Ultra-deep sequencing of hadza hunter-gatherers recovers vanishing gut microbes. *Cell*.

Chang, T., Gavelis, G. S., Brown, J. M., and Stepanauskas, R. (2024). Genomic representativeness and chimerism in large collections of SAGs and MAGs of marine prokaryoplankton. *Microbiome*, 12(1):126.

Chen, J., Jia, Y., Sun, Y., Liu, K., Zhou, C., Liu, C., Li, D., Liu, G., Zhang, C., Yang, T., et al. (2024). Global marine microbial diversity and its potential in bioprospecting. *Nature*, 633(8029):371–379.

Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M., and Banfield, J. F. (2020a). Accurate and complete genomes from metagenomes. *Genome Res.*, 30(3):315–333.

Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M., and Banfield, J. F. (2020b). Accurate and complete genomes from metagenomes. *Genome Res.*, 30(3):315–333.

Chen, T., Yu, W.-H., Izard, J., Baranova, O. V., Lakshmanan, A., and Dewhirst, F. E. (2010). The human oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database (Oxford)*, 2010(0):baq013.

Coleman, M. L., Sullivan, M. B., Martiny, A. C., Steglich, C., Barry, K., DeLong, E. F., and Chisholm, S. W. (2006). Genomic islands and the ecology and evolution of prochlorococcus. *Science*, 311(5768):1768–1770.

Cornman, A., West-Roberts, J., Camargo, A. P., Roux, S., Beracochea, M., Mirdita, M., Ovchinnikov, S., and Hwang, Y. (2024). The OMG dataset: An open MetaGenomic corpus for mixed-modality genomic language modeling. *bioRxiv*.

Coyne, M. J., Tzianabos, A. O., Mallory, B. C., Carey, V. J., Kasper, D. L., and Comstock, L. E. (2001). Polysaccharide biosynthesis locus required for virulence of bacteroides fragilis. *Infect. Immun.*, 69(7):4342–4350.

Crits-Christoph, A., Bhattacharya, N., Olm, M. R., Song, Y. S., and Banfield, J. F. (2020). Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity. *Genome Research*, 31(2):239–250.

Crits-Christoph, A., Diamond, S., Al-Shayeb, B., Valentin-Alvarado, L., and Banfield, J. F. (2022). A widely distributed genus of soil acidobacteria genomically enriched in biosynthetic gene clusters. *ISME Communications*, 2(1):1–8.

Cross, K. L., Campbell, J. H., Balachandran, M., Campbell, A. G., Cooper, C. J., Griffen, A., Heaton, M., Joshi, S., Klingeman, D., Leys, E., Yang, Z., Parks, J. M., and Podar, M. (2019). Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. *Nat. Biotechnol.*, 37(11):1314–1321.

Da Cunha, V., Gaia, M., Gadelle, D., Nasir, A., and Forterre, P. (2017). Lokiarchaea are close relatives of euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet.*, 13(6):e1006810.

Delmont, T. O. and Eren, A. M. (2018). Linking pangenomes and metagenomes: the prochlorococcus metapangenome. *PeerJ*, 6:e4320.

Delmont, T. O., Kiefl, E., Kilinc, O., Esen, Ö. C., Uysal, I., RappÃľ, M. S., Giovannoni, S., and Eren, A. M. (2019). Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *Elife*, 8.

Delmont, T. O., Quince, C., Shaiber, A., Esen, Ö. C., Lee, S. T., RappÃľ, M. S., McLellan, S. L., LÃijcker, S., and Eren, A. M. (2018a). Nitrogen-fixing populations of planctomycetes and proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol*, 3(7):804–813.

Delmont, T. O., Quince, C., Shaiber, A., Esen, Ö. C., Lee, S. T., RappÃľ, M. S., McLellan, S. L., LÃijcker, S., and Eren, A. M. (2018b). Nitrogen-fixing populations of planctomycetes and proteobacteria are abundant in surface ocean metagenomes. *Nat. Microbiol.*, 3(7):804–813.

DeLong, E. F. (1992). Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences*, 89(12):5685–5689.

DeLong, E. F., Wu, K. Y., Prézelin, B. B., and Jovine, R. V. M. (1994). High abundance of archaea in antarctic marine picoplankton. *Nature*, 371(6499):695–697.

Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C. R., Yu, W.-H., Lakshmanan, A., and Wade, W. G. (2010). The human oral microbiome. *J. Bacteriol.*, 192(19):5002–5017.

Diamond, S., Andeer, P. F., Li, Z., Crits-Christoph, A., Burstein, D., Anantharaman, K., Lane, K. R., Thomas, B. C., Pan, C., Northen, T. R., and Banfield, J. F. (2019). Mediterranean grassland soil C-N compound turnover is dependent on rainfall and depth, and is mediated by genomically divergent microorganisms. *Nat Microbiol*, 4(8):1356–1367.

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *J. Veg. Sci.*, 14(6):927–930.

Duhamel, S., Diaz, J. M., Adams, J. C., Djaoudi, K., Steck, V., and Waggoner, E. M. (2021). Phosphorus as an integral component of global marine biogeochemistry. *Nat. Geosci.*, 14(6):359–368.

Durham, B. P., Sharma, S., Luo, H., Smith, C. B., Amin, S. A., Bender, S. J., Dearth, S. P., Van Mooy, B. A. S., Campagna, S. R., Kujawinski, E. B., Armbrust, E. V., and Moran, M. A. (2014). Cryptic carbon and sulfur cycling between surface ocean plankton. *Proceedings of the National Academy of Sciences*, 112(2):453âĂŞ457.

Durrant, M. G., Fanton, A., Tycko, J., Hinks, M., Chandrasekaran, S. S., Perry, N. T., Schaepe, J., Du, P. P., Lotfy, P., Bassik, M. C., Bintu, L., Bhatt, A. S., and Hsu, P. D. (2022). Systematic discovery of recombinases for efficient integration of large dna sequences into the human genome. *Nature Biotechnology*, 41(4):488âĂŞ499.

Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics*, 14(9):755–763.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.*, 7(10):e1002195.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797.

Emerson, J. B., Thomas, B. C., Alvarez, W., and Banfield, J. F. (2016). Metagenomic analysis of a high carbon dioxide subsurface microbial community populated by chemolithoautotrophs and bacteria and archaea from candidate phyla. *Environ. Microbiol.*, 18(6):1686–1703.

Eren, A. M. and Banfield, J. F. (2024). Modern microbiology: Embracing complexity through integration across scales. *Cell*, 187(19):5151–5170.

Eren, A. M., Borisy, G. G., Huse, S. M., and Mark Welch, J. L. (2014). Oligotyping analysis of the human oral microbiome. *Proc. Natl. Acad. Sci. U. S. A.*, 111(28):E2875–84.

Eren, A. M. and Delmont, T. O. (2024). Bioprospecting marine microbial genomes to improve biotechnology. *Nature*, 633(8029):287–288.

Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., and Delmont, T. O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3(e1319):e1319.

Eren, A. M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S. E., Schechter, M. S., Fink, I., Pan, J. N., Yousef, M., Fogarty, E. C., Trigodet, F., Watson, A. R., Esen, Ã. C., Moore, R. M., Clayssen, Q., Lee, M. D., Kivenson, V., Graham, E. D., Merrill, B. D., Karkman, A., Blankenberg, D., Eppley, J. M., SjÃűdin, A., Scott, J. J., Vázquez-Campos, X., McKay, L. J., McDaniel, E. A., Stevens, S. L. R., Anderson, R. E., Fuessel, J., Fernandez-Guerra, A., Maignien, L., Delmont, T. O., and Willis, A. D. (2021). Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol*, 6(1):3–6.

Eren, A. M., Vineis, J. H., Morrison, H. G., and Sogin, M. L. (2013). A filtering method to generate high quality short reads using illumina paired-end technology. *PLoS One*, 8(6):e66643.

Escapa, I. F., Chen, T., Huang, Y., Gajare, P., Dewhirst, F. E., and Lemon, K. P. (2018). New insights into human nostril microbiome from the expanded human oral microbiome database (eHOMD): a resource for the microbiome of the human aerodigestive tract. *mSystems*, 3(6).

Evans, J. T. and Denef, V. J. (2020). To dereplicate or not to dereplicate? *mSphere*, 5(3).

Falkowski, P. (2012). Ocean science: the power of plankton. *Nature*, 483(7387):S17–S20.

Falkowski, P. G., Fenchel, T., and Delong, E. F. (2008). The microbial engines that drive earthâĂŹs biogeochemical cycles. *Science*, 320(5879):1034âĂŞ1039.

Fan, Y. and Pedersen, O. (2020). Gut microbiota in human metabolic health and disease. *Nature Reviews Microbiology*, 19(1):55–71.

Fernández, A. B., Ghai, R., Martin-Cuadrado, A.-B., Sánchez-Porro, C., Rodriguez-Valera, F., and Ventosa, A. (2014). Prokaryotic taxonomic and metabolic diversity of an intermediate salinity hypersaline habitat assessed by metagenomics. *FEMS Microbiol. Ecol.*, 88(3):623–635.

Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong, D. T., Manara, S., Zolfo, M., Beghini, F., Bertorelli, R., De Sanctis, V., Bariletti, I., Canto, R., Clementi, R., Cologna, M., Crifó, T., Cusumano, G., Gottardi, S., Innamorati, C., Masé, C., Postai, D., Savoi, D., Duranti, S., Lugli, G. A., Mancabelli, L., Turroni, F., Ferrario, C., Milani, C., Mangifesta, M., Anzalone, R., Viappiani, A., Yassour, M., Vlamakis, H., Xavier, R., Collado, C. M., Koren, O., Tateo, S., Soffiati, M., Pedrotti, A., Ventura, M., Huttenhower, C., Bork, P., and Segata, N. (2018). Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host Microbe*, 24(1):133–145.e5.

Fierer, N. and Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences*, 103(3):626âĂŞ631.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2016). The pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, 44(D1):D279–85.

Flores, H. A. and O'Neill, S. L. (2018). Controlling vector-borne diseases by releasing modified mosquitoes. *Nat. Rev. Microbiol.*, 16(8):508–518.

Fogarty, E. C., Schechter, M. S., Lolans, K., Sheahan, M. L., Veseli, I., Moore, R. M., Kiefl, E., Moody, T., Rice, P. A., Yu, M. K., Mimee, M., Chang, E. B., Ruscheweyh, H.-J., Sunagawa, S., Mclellan, S. L., Willis, A. D., Comstock, L. E., and Eren, A. M. (2024). A cryptic plasmid is among the most numerous genetic elements in the human gut. *Cell*, 187(5):1206–1222.e16.

Freel, K. C., Tucker, S. J., Freel, E. B., Giovannoni, S. J., Eren, A. M., and Rapé, M. S. (2024). New isolate genomes and global marine metagenomes resolve ecologically relevant units of sar11. *BioRxiv*.

Fuhrman, J. A., McCallum, K., and Davis, A. A. (1992). Novel major archaebacterial group from marine plankton. *Nature*, 356(6365):148âĂŞ149.

Gaia, M., Meng, L., Pelletier, E., Forterre, P., Vanni, C., Fernandez-Guerra, A., Jaillon, O., Wincker, P., Ogata, H., Krupovic, M., and Delmont, T. O. (2023). Mirusviruses link herpesviruses to giant viruses. *Nature*, 616(7958):783–789.

Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, 31(22):3718–3720.

Garber, A. I., Ramírez, G. A., and D'Hondt, S. (2024). Genomic stasis over millions of years in subseafloor sediment. *Environmental Microbiology*, 26(8).

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., and Knight, R. (2018). Current understanding of the human microbiome. *Nat. Med.*, 24(4):392–400.

Giovannoni, S. J. (2017). SAR11 bacteria: The most abundant plankton in the oceans. *Ann. Rev. Mar. Sci.*, 9:231–255.

Grieneisen, L., Dasari, M., Gould, T. J., Bjork, J. R., Grenier, J.-C., Yotova, V., Jansen, D., Gottel, N., Gordon, J. B., Learn, N. H., Gesquiere, L. R., Wango, T. L., Mututua, R. S., Warutere, J. K., Siodi, L., Gilbert, J. A., Barreiro, L. B., Alberts, S. C., Tung, J., Archie, E. A., and Blekhman, R. (2021). Gut microbiome heritability is nearly universal but environmentally contingent. *Science*, 373(6551):181–186.

Grieneisen, L., Muehlbauer, A. L., and Blekhman, R. (2020). Microbial control of host gene regulation and the evolution of host-microbiome interactions in primates. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 375(1808):20190598.

Gruber, N. and Galloway, J. N. (2008). An earth-system perspective of the global nitrogen cycle. *Nature*, 451(7176):293–296.

Gushgari-Doyle, S., Lui, L. M., Nielsen, T. N., Wu, X., Malana, R. G., Hendrickson, A. J., Carion, H., Poole, F. L., Adams, M. W. W., Arkin, A. P., and Chakraborty, R. (2022). Genotype to ecotype in niche environments: adaptation of arthrobacter to carbon availability and environmental conditions. *ISME Communications*, 2(1).

Halloran, K. H., Braakman, R., Coe, A., Swarr, G., Kido Soule, M. C., Chisholm, S. W., and Kujawinski, E. B. (2025). Uptake of prochlorococcus-derived metabolites by alteromonas macleodii MIT1002 shows high levels of substrate specificity. *Microbiology*, (biorxiv;2025.01.10.632383v1).

Hamilton, T. L., Bovee, R. J., Sattin, S. R., Mohr, W., Gilhooly, 3rd, W. P., Lyons, T. W., Pearson, A., and Macalady, J. L. (2016). Carbon and sulfur cycling below the chemocline in a meromictic lake and the identification of a novel taxonomic lineage in the FCB superphylum, candidatus aegiribacteria. *Front. Microbiol.*, 7:598.

Hammer, T. J., Sanders, J. G., and Fierer, N. (2019). Not all animals need a microbiome. *FEMS Microbiol. Lett.*, 366(10).

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry and Biology*, 5(10):R245âĂŞR249.

He, X., McLean, J. S., Edlund, A., Yooseph, S., Hall, A. P., Liu, S.-Y., Dorrestein, P. C., Esquenazi, E., Hunter, R. C., Cheng, G., Nelson, K. E., Lux, R., and Shi, W. (2015). Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proc. Natl. Acad. Sci. U. S. A.*, 112(1):244–249.

Heidelberg, J. F., Paulsen, I. T., Nelson, K. E., Gaidos, E. J., Nelson, W. C., Read, T. D., Eisen, J. A., Seshadri, R., Ward, N., Methe, B., et al. (2002). Genome sequence of the dissimilatory metal ion–reducing bacterium shewanella oneidensis. *Nature biotechnology*, 20(11):1118–1123.

HenrÃ■quez-Castillo, C., Plominsky, A. M., RamÃ■rez-Flandes, S., Bertagnolli, A. D., Stewart, F. J., and Ulloa, O. (2022). Metaomics unveils the contribution of alteromonas bacteria to carbon cycling in marine oxygen minimum zones. *Front. Mar. Sci.*, 9.

Hosokawa, M. and Nishikawa, Y. (2024). Tools for microbial single-cell genomics for obtaining uncultured microbial genomes. *Biophys. Rev.*, 16(1):69–77.

Huang, S., Wilhelm, S. W., Harvey, H. R., Taylor, K., Jiao, N., and Chen, F. (2012). Novel lineages of prochlorococcus and synechococcus in the global oceans. *ISME J.*, 6(2):285–297.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., and Banfield, J. F. (2016). A new view of the tree of life. *Nat Microbiol*, 1:16048.

Hultman, J., Waldrop, M. P., Mackelprang, R., David, M. M., McFarland, J., Blazewicz, S. J., Harden, J., Turetsky, M. R., McGuire, A. D., Shah, M. B., VerBerkmoes, N. C., Lee, L. H., Mavrommatis, K., and Jansson, J. K. (2015). Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature*, 521(7551):208–212.

Human Microbiome Project Consortium (2012). A framework for human microbiome research. *Nature*, 486(7402):215–221.

Hutchins, D. A. and Capone, D. G. (2022). The marine nitrogen cycle: new developments and global change. *Nat. Rev. Microbiol.*, 20(7):401–414.

Hwang, Y., Cornman, A. L., Kellogg, E. H., Ovchinnikov, S., and Girguis, P. R. (2024). Genomic language model predicts protein co-regulation and function. *Nat. Commun.*, 15(1):2880.

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11:119.

Ikeda, S., Takamatsu, Y., Tsuchiya, M., Suga, K., Tanaka, Y., Kouzuma, A., and Watanabe, K. (2021). Shewanella oneidensis mr-1 as a bacterial platform for electro-biotechnology. *Essays in Biochemistry*, 65(2):355–364.

Imachi, H., Nobu, M. K., Nakahara, N., Morono, Y., Ogawara, M., Takaki, Y., Takano, Y., Uematsu, K., Ikuta, T., Ito, M., Matsui, Y., Miyazaki, M., Murata, K., Saito, Y., Sakai, S., Song, C., Tasumi, E., Yamanaka, Y., Yamaguchi, T., Kamagata, Y., Tamaki, H., and Takai, K. (2020). Isolation of an archaeon at the prokaryote–eukaryote interface.

Jaffe, A. L., Konno, M., Kawasaki, Y., Kataoka, C., BÃľjÃă, O., Kandori, H., Inoue, K., and Banfield, J. F. (2022). Saccharibacteria harness light energy using type-1 rhodopsins that may rely on retinal sourced from microbial hosts. *The ISME Journal*, 16(8):2056âĂŞ2059.

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*, 9(1):5114.

Jiang, C.-Y., Dong, L., Zhao, J.-K., Hu, X., Shen, C., Qiao, Y., Zhang, X., Wang, Y., Ismagilov, R. F., Liu, S.-J., and Du, W. (2016). High-throughput single-cell cultivation on microfluidic streak plates. *Appl. Environ. Microbiol.*, 82(7):2210–2218.

Johnson, Z. I., Zinser, E. R., Coe, A., McNulty, N. P., Woodward, E. M. S., and Chisholm, S. W. (2006). Niche partitioning among prochlorococcus ecotypes along ocean-scale environmental gradients. *Science*, 311(5768):1737–1740.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, 8(3):275–282.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Å¡Ãdek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*, 14(6):587–589.

Kang, D. D., Froula, J., Egan, R., and Wang, Z. (2015). Metabat, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3:e1165.

Kauffman, K. M., Hussain, F. A., Yang, J., Arevalo, P., Brown, J. M., Chang, W. K., VanInsberghe, D., Elsherbini, J., Sharma, R. S., Cutler, M. B., Kelly, L., and Polz, M. F. (2018). A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature*, 554(7690):118–122.

Kautsar, S. A., Blin, K., Shaw, S., Weber, T., and Medema, M. H. (2020). Big-fam: the biosynthetic gene cluster families database. *Nucleic Acids Research*, 49(D1):D490âĂŞD497.

Kawano-Sugaya, T., Arikawa, K., Saeki, T., Endoh, T., Kamata, K., Matsuhashi, A., and Hosokawa, M. (2024). A single amplified genome catalog reveals the dynamics of mobilome and resistome in the human microbiome. *Microbiome*, 12(1).

Kent, A. G., Baer, S. E., Mouginot, C., Huang, J. S., Larkin, A. A., Lomas, M. W., and Martiny, A. C. (2019). Parallel phylogeography of *Prochlorococcus* and *Synechococcus*. *ISME J.*, 13(2):430–441.

Kent, A. G., Dupont, C. L., Yooseph, S., and Martiny, A. C. (2016). Global biogeography of prochlorococcus genome diversity in the surface ocean. *ISME J.*, 10(8):1856–1865.

Klemetsen, T., Raknes, I. A., Fu, J., Agafonov, A., Balasundaram, S. V., Tartari, G., Robertsen, E., and Willassen, N. P. (2018). The MAR databases: development and implementation of databases specific for marine metagenomics. *Nucleic Acids Res.*, 46(D1):D692–D699.

Koster, J. and Rahmann, S. (2012). Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522.

Kostic, A. D., Xavier, R. J., and Gevers, D. (2014). The microbiome in inflammatory bowel disease: Current status and the future ahead. *Gastroenterology*, 146(6):1489âĂŞ1499.

Krüger, K., Chafee, M., Ben Francis, T., Glavina del Rio, T., Becher, D., Schweder, T., Amann, R. I., and Teeling, H. (2019). In marine bacteroidetes the bulk of glycan degradation during algae blooms is mediated by few clades using a restricted set of genes. *The ISME Journal*, 13(11):2800âĂŞ2816.

Lane, N. (2015). The unseen world: reflections on leeuwenhoek (1677) 'concerning little animals'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1666):20140344.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4):357–359.

Lee, M. D. (2019). GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics*, 35(20):4162–4164.

Lewis, W. H., Tahon, G., Geesink, P., Sousa, D. Z., and Ettema, T. J. G. (2021). Innovations to culturing the uncultured microbial majority. *Nat. Rev. Microbiol.*, 19(4):225–240.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

Little, A. S., Younker, I. T., Schechter, M. S., Bernardino, P. N., MÃľheust, R., Stemczynski, J., Scorza, K., Mullowney, M. W., Sharan, D., Waligurski, E., Smith, R., Ramanswamy, R., Leiter, W., Moran, D., McMillin, M., Odenwald, M. A., Iavarone, A. T., Sidebottom, A. M., Sundararajan, A., Pamer, E. G., Eren, A. M., and Light, S. H. (2024). Dietary- and host-derived metabolites are used by diverse gut bacteria for anaerobic respiration. *Nat. Microbiol.*, 9(1):55–69.

Lu, J., Breitwieser, F. P., Thielen, P., and Salzberg, S. L. (2017). Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.*, 3(e104):e104.

Lu, Z., Entwistle, E., Kuhl, M. D., Durrant, A. R., Filho, M. M. B., Goswami, A., and Morris, J. J. (2024). Coevolution of marine phytoplankton and alteromonas bacteria in response to pCO2 and co-culture. *ISME J.*, page wrae259.

Lynch, J. M., Brimecombe, M. J., and De Leij, F. A. (2001). *Rhizosphere*. John Wiley & Sons, Ltd, Chichester, UK.

Ma, B., Lu, C., Wang, Y., Yu, J., Zhao, K., Xue, R., Ren, H., Lv, X., Pan, R., Zhang, J., Zhu, Y., and Xu, J. (2023). A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. *Nat. Commun.*, 14(1):7318.

Maini Rekdal, V., Nol Bernadino, P., Luescher, M. U., Kiamehr, S., Le, C., Bisanz, J. E., Turnbaugh, P. J., Bess, E. N., and Balskus, E. P. (2020). A widely distributed metalloenzyme class enables gut microbial metabolism of host-and diet-derived catechols. *Elife*, 9:e50845.

Malmstrom, R. R., Rodrigue, S., Huang, K. H., Kelly, L., Kern, S. E., Thompson, A., Roggensack, S., Berube, P. M., Henn, M. R., and Chisholm, S. W. (2013). Ecology of uncultured prochlorococcus clades revealed through single-cell genomics and biogeographic analysis. *ISME J.*, 7(1):184–198.

Manck, L. E., Park, J., Tully, B. J., Poire, A. M., Bundy, R. M., Dupont, C. L., and Barbeau, K. A. (2022). Petrobactin, a siderophore produced by alteromonas, mediates community iron acquisition in the global ocean. *ISME J.*, 16(2):358–369.

Manghi, P., Blanco-MÃ■guez, A., Manara, S., NabiNejad, A., Cumbo, F., Beghini, F., Armanini, F., Golzato, D., Huang, K. D., Thomas, A. M., Piccinno, G., PunÄ■ochÃąÅŹ, M., Zolfo, M., Lesker, T. R., Bredon, M., Planchais, J., Glodt, J., Valles-Colomer, M., Koren, O., Pasolli, E., Asnicar, F., Strowig, T., Sokol, H., and Segata, N. (2023). MetaPhlAn 4 profiling of unknown species-level genome bins improves the characterization of diet-associated microbiome changes in mice. *Cell Rep.*, 42(5):112464.

Mark Welch, J. L., Dewhirst, F. E., and Borisy, G. G. (2019). Biogeography of the oral microbiome: The site-specialist hypothesis. *Annu. Rev. Microbiol.*, 73:335–358.

Mark Welch, J. L., Rossetti, B. J., Rieken, C. W., Dewhirst, F. E., and Borisy, G. G. (2016). Biogeography of a human oral microbiome at the micron scale. *Proceedings of the National Academy of Sciences*, 113(6).

Martinez-Perez, C., Greening, C., Bay, S. K., Lappan, R. J., Zhao, Z., De Corte, D., Hulbe, C., Ohneiser, C., Stevens, C., Thomson, B., Stepanauskas, R., GonzÃąlez, J. M., Logares, R., Herndl, G. J., Morales, S. E., and Baltar, F. (2022). Phylogenetically and functionally diverse microorganisms reside under the ross ice shelf. *Nat. Commun.*, 13(1):117.

Matheus Carnevali, P. B., Lavy, A., Thomas, A. D., Crits-Christoph, A., Diamond, S., MÃľheust, R., Olm, M. R., Sharrar, A., Lei, S., Dong, W., Falco, N., Bouskill, N., Newcomer, M. E., Nico, P., Wainwright, H., Dwivedi, D., Williams, K. H., Hubbard, S., and Banfield, J. F. (2021). Meanders as a scaling motif for understanding of floodplain soil microbiome and biogeochemical potential at the watershed scale. *Microbiome*, 9(1):121.

McFall-Ngai, M. J. and Ruby, E. G. (1991). Symbiont recognition and subsequent morphogenesis as early events in an animal-bacterial mutualism. *Science*, 254(5037):1491–1494.

McLean, A. R., Torres-Morales, J., Dewhirst, F. E., Borisy, G. G., and Mark Welch, J. L. (2022). Site-tropism of streptococci in the oral microbiome. *Mol. Oral Microbiol.*, 37(6):229–243.

McMahon, K. (2015). 'metagenomics 2.0'. *Environmental Microbiology Reports*, 7(1):38âĂŞ39.

McMurdie, P. J. and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, 8(4):e61217.

Meyer, F., Fritz, A., Deng, Z.-L., Koslicki, D., Lesker, T. R., Gurevich, A., Robertson, G., Alser, M., Antipov, D., Beghini, F., Bertrand, D., Brito, J. J., Brown, C. T., Buchmann, J., BuluÃğ, A., Chen, B., Chikhi, R., Clausen, P. T. L. C., Cristian, A., Dabrowski, P. W., Darling, A. E., Egan, R., Eskin, E., Georganas, E., Goltsman, E., Gray, M. A., Hansen, L. H., Hofmeyr, S., Huang, P., Irber, L., Jia, H., JÃÿrgensen, T. S., Kieser, S. D., Klemetsen, T., Kola, A., Kolmogorov, M., Korobeynikov, A., Kwan, J., LaPierre, N., Lemaitre, C., Li, C., Limasset, A., Malcher-Miranda, F., Mangul, S., Marcelino, V. R., Marchet, C., Marijon, P., Meleshko, D., Mende, D. R., Milanese, A., Nagarajan, N., Nissen, J., Nurk, S., Oliker, L., Paoli, L., Peterlongo, P., Piro, V. C., Porter, J. S., Rasmussen, S., Rees, E. R., Reinert, K., Renard, B., Robertsen, E. M., Rosen, G. L., Ruscheweyh, H.-J., Sarwal, V., Segata, N., Seiler, E., Shi, L., Sun, F., Sunagawa, S., SÃÿrensen, S. J., Thomas, A., Tong, C., Trajkovski, M., Tremblay, J., Uritskiy, G., Vicedomini, R., Wang, Z., Wang, Z., Wang, Z., Warren, A., Willassen, N. P., Yelick, K., You, R., Zeller, G., Zhao, Z., Zhu, S., Zhu, J., Garrido-Oter, R., Gastmeier, P., Hacquard, S., HÃďuÃ§ler, S., Khaledi, A., Maechler, F., Mesny, F., Radutoiu, S., Schulze-Lefert, P., Smit, N., Strowig, T., Bremges, A., Sczyrba, A., and McHardy, A. C. (2022). Critical assessment of metagenome interpretation: the second round of challenges. *Nat. Methods*, 19(4):429–440.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics*, 32(7):1088–1090.

Mimee, M., Nadeau, P., Hayward, A., Carim, S., Flanagan, S., Jerger, L., Collins, J., McDonnell, S., Swartwout, R., Citorik, R. J., Bulović, V., Langer, R., Traverso, G., Chandrakasan, A. P., and Lu, T. K. (2018). An ingestible bacterial-electronic system to monitor gastrointestinal health. *Science*, 360(6391):915–918.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.*, 37(5):1530–1534.

Minoche, A. E., Dohm, J. C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on illumina HiSeq and genome analyzer systems. *Genome Biol.*, 12(11):R112.

Mistou, M.-Y., Sutcliffe, I. C., and van Sorge, N. M. (2016). Bacterial glycobiology: rhamnose-containing cell wall polysaccharides in gram-positive bacteria. *FEMS Microbiology Reviews*, 40(4):464–479.

Mizrahi, I., Wallace, R. J., and Morais, S. (2021). The rumen microbiome: balancing food security and environmental impacts. *Nat. Rev. Microbiol.*, 19(9):553–566.

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., and KÃűster, J. (2021). Sustainable data analysis with snakemake. *F1000Res.*, 10(33):33.

Moodie, A. D. and Ingledew, W. J. (1990). Microbial anaerobic respiration. *Advances in microbial physiology*, 31:225–269.

Morris, R. M., Rappé, M. S., Connon, S. A., Vergin, K. L., Siebold, W. A., Carlson, C. A., and Giovannoni, S. J. (2002). SAR11 clade dominates ocean surface bacterioplankton communities. *Nature*, 420(6917):806–810.

Naether, D. J., Slawtschew, S., Stasik, S., Engel, M., Olzog, M., Wick, L. Y., Timmis, K. N., and Heipieper, H. J. (2013). Adaptation of the hydrocarbonoclastic bacterium alcanivorax borkumensis SK2 to alkanes and toxic organic compounds: a physiological and transcriptomic approach. *Appl. Environ. Microbiol.*, 79(14):4282–4293.

Nagalingam, N. A. and Lynch, S. V. (2012). Role of the microbiota in inflammatory bowel diseases. *Inflammatory Bowel Diseases*, 18(5):968–984.

Nayfach, S., Rodriguez-Mueller, B., Garud, N., and Pollard, K. S. (2016). An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.*, 26(11):1612–1625.

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., Huntemann, M., et al. (2021). A genomic catalog of earthâĂŹs microbiomes. *Nature biotechnology*, 39(4):499–509.

Nealson, K. H., Platt, T., and Hastings, J. W. (1970). Cellular control of the synthesis and activity of the bacterial luminescent System1. *J. Bacteriol.*, 104(1):313–322.

Nguyen, E., Poli, M., Durrant, M. G., Thomas, A. W., Kang, B., Sullivan, J., Ng, M. Y., Lewis, A., Patel, A., Lou, A., Ermon, S., Baccus, S. A., Hernandez-Boussard, T., Re, C., Hsu, P. D.,

and Hie, B. L. (2024). Sequence modeling and design from molecular to genome scale with evo.

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, 32(1):268–274.

Nowicki, M., DeVries, T., and Siegel, D. A. (2022). Quantifying the carbon export and sequestration pathways of the ocean's biological carbon pump. *Global Biogeochemical Cycles*, 36(3).

Obregon-Tito, A. J., Tito, R. Y., Metcalf, J., Sankaranarayanan, K., Clemente, J. C., Ursell, L. K., Xu, Z. Z., Van Treuren, W., Knight, R., Gaffney, P. M., Spicer, P., Lawson, P., Marin-Reyes, L., Trujillo-Villarroel, O., Foster, M., Guija-Poma, E., Troncoso-Corzo, L., Warinner, C., Ozga, A. T., and Lewis, C. M. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes.

Olm, M. R., Crits-Christoph, A., Diamond, S., Lavy, A., Matheus Carnevali, P. B., and Banfield, J. F. (2020). Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. *mSystems*, 5(1).

O'Malley, M. A. (2017). From endosymbiosis to holobionts: Evaluating a conceptual legacy. *J. Theor. Biol.*, 434:34–41.

Pachiadaki, M. G., Brown, J. M., Brown, J., Bezuidt, O., Berube, P. M., Biller, S. J., Poulton, N. J., Burkart, M. D., La Clair, J. J., Chisholm, S. W., and Stepanauskas, R. (2019). Charting the complexity of the marine microbiome through single-cell genomics. *Cell*, 179(7):1623–1635.e11.

Palleja, A., Mikkelsen, K. H., Forslund, S. K., Kashani, A., Allin, K. H., Nielsen, T., Hansen, T. H., Liang, S., Feng, Q., Zhang, C., Pyl, P. T., Coelho, L. P., Yang, H., Wang, J., Typas, A., Nielsen, M. F., Nielsen, H. B., Bork, P., Wang, J., Vilsbøll, T., Hansen, T., Knop, F. K., Arumugam, M., and Pedersen, O. (2018). Recovery of gut microbiota of healthy adults following antibiotic exposure. *Nat. Microbiol.*, 3(11):1255–1265.

Pan, S., Zhu, C., Zhao, X.-M., and Coelho, L. P. (2022). A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different environments. *Nature Communications*, 13(1).

Paoli, L., Ruscheweyh, H.-J., Forneris, C. C., Hubrich, F., Kautsar, S., Bhushan, A., Lotti, A., Clayssen, Q., Salazar, G., Milanese, A., Carlstrom, C. I., Papadopoulou, C., Gehrig, D., Karasikov, M., Mustafa, H., Larralde, M., Carroll, L. M., Sánchez, P., Zayed, A. A., Cronin, D. R., Acinas, S. G., Bork, P., Bowler, C., Delmont, T. O., Gasol, J. M., Gossert, A. D., Kahles, A., Sullivan, M. B., Wincker, P., Zeller, G., Robinson, S. L., Piel, J., and Sunagawa, S. (2022). Biosynthetic potential of the global ocean microbiome. *Nature*, 607(7917):111–118.

Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for bacteria and archaea. *Nat. Biotechnol.*, 38(9):1079–1086.

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., and Hugenholtz, P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.*, 50(D1):D785–D794.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, 36(10):996–1004.

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P., and Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol*, 2(11):1533–1542.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M. C., Rice, B. L., DuLong, C., Morgan, X. C., Golden, C. D., Quince, C., Huttenhower, C., and Segata, N. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, 176(3):649–662.e20.

Patrick, S., Blakely, G. W., Houston, S., Moore, J., Abratt, V. R., Bertalan, M., Cerdeño-Tárraga, A. M., Quail, M. A., Corton, N., Corton, C., Bignell, A., Barron, A., Clark, L., Bentley, S. D., and Parkhill, J. (2010). Twenty-eight divergent polysaccharide loci specifying within- and amongst-strain capsule diversity in three strains of bacteroides fragilis. *Microbiology*, 156(Pt 11):3255–3269.

Pedler, B. E., Aluwihare, L. I., and Azam, F. (2014). Single bacterial strain capable of significant contribution to carbon cycling in the surface ocean. *Proc. Natl. Acad. Sci. U. S. A.*, 111(20):7202–7207.

Pieterse, C. M. J., Zamioudis, C., Berendsen, R. L., Weller, D. M., Van Wees, S. C. M., and Bakker, P. A. H. M. (2014). Induced systemic resistance by beneficial microbes. *Annu. Rev. Phytopathol.*, 52(1):347–375.

Porter, K. G. and Feig, Y. S. (1980). The use of dapi for identifying and counting aquatic microflora1. *Limnology and Oceanography*, 25(5):943–948.

Prasad, M., Obana, N., Lin, S.-Z., Zhao, S., Sakai, K., Blanch-Mercader, C., Prost, J., Nomura, N., Rupprecht, J.-F., Fattaccioli, J., and Utada, A. S. (2023). Alcanivorax borkumensis biofilms enhance oil degradation by interfacial tubulation. *Science*, 381(6659):748–753.

Prasad, M., Obana, N., Sakai, K., Nagakubo, T., Miyazaki, S., Toyofuku, M., Fattaccioli, J., Nomura, N., and Utada, A. S. (2019). Point mutations lead to increased levels of c-di-GMP and phenotypic changes to the colony biofilm morphology in alcanivorax borkumensis SK2. *Microbes Environ.*, 34(1):104–107.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, 5(3):e9490.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E., Renault, P., Pons, N., Batto, J.-M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li, S., Yang, H., Wang, J., Ehrlich, S. D., Nielsen, R., Pedersen, O., Kristiansen, K., and Wang, J. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60.

Qu, A., Brulc, J. M., Wilson, M. K., Law, B. F., Theoret, J. R., Joens, L. A., Konkel, M. E., Angly, F., Dinsdale, E. A., Edwards, R. A., Nelson, K. E., and White, B. A. (2008). Comparative metagenomics reveals host specific metavirulomes and horizontal gene transfer elements in the chicken cecum microbiome. *PLoS One*, 3(8):e2945.

Reveillaud, J., Bordenstein, S. R., Cruaud, C., Shaiber, A., Esen, Ö. C., Weill, M., Makoundou, P., Lolans, K., Watson, A. R., Rakotoarivony, I., Bordenstein, S. R., and Eren, A. M. (2019). The wolbachia mobilome in culex pipiens includes a putative plasmid. *Nat. Commun.*, 10(1):1051.

Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M. L., Burdett, T., Burgin, J., Caballero-Pérez, J., Cochrane, G., Colwell, L. J., et al. (2023). Mgnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Research*, 51(D1):D753–D759.

Rusch, D. B., Martiny, A. C., Dupont, C. L., Halpern, A. L., and Venter, J. C. (2010). Characterization of prochlorococcus clades from iron-depleted oceanic regions. *Proc. Natl. Acad. Sci. U. S. A.*, 107(37):16184–16189.

Ruscheweyh, H.-J., Milanese, A., Paoli, L., Karcher, N., Clayssen, Q., Keller, M. I., Wirbel, J., Bork, P., Mende, D. R., Zeller, G., and Sunagawa, S. (2022). Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. *Microbiome*, 10(1):212.

Sabirova, J. S., Chernikova, T. N., Timmis, K. N., and Golyshin, P. N. (2008). Niche-specificity factors of a marine oil-degrading bacterium alcanivorax borkumensis SK2. *FEMS Microbiol. Lett.*, 285(1):89–96.

Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H.-J., Cuenca, M., Field, C. M., Coelho, L. P., Cruaud, C., Engelen, S., Gregory, A. C., Labadie, K., Marec, C., Pelletier, E., Royo-Llonch, M., Roux, S., SÃąnchez, P., Uehara, H., Zayed, A. A., Zeller, G., Carmichael, M., Dimier, C., Ferland, J., Kandels, S., Picheral, M., Pisarev, S., Poulain, J., Tara Oceans Coordinators, Acinas, S. G., Babin, M., Bork, P., Bowler, C., de Vargas, C., Guidi, L., Hingamp, P., Iudicone, D., Karp-Boss, L., Karsenti, E., Ogata, H., Pesant, S., Speich, S., Sullivan, M. B., Wincker, P., and Sunagawa, S. (2019). Gene expression

changes and community turnover differentially shape the global ocean metatranscriptome. *Cell*, 179(5):1068–1083.e21.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., and Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.*, 12:87.

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T. J., Schatz, M. C., Delcher, A. L., Roberts, M., MarÃğais, G., Pop, M., and Yorke, J. A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.*, 22(3):557–567.

Sanchez, P., Coutinho, F. H., SebastiÃąn, M., Pernice, M. C., RodrÃ■guez-MartÃ■nez, R., Salazar, G., Cornejo-Castillo, F. M., Pesant, S., LÃşpez-Alforja, X., LÃşpez-GarcÃ■a, E. M., AgustÃ■, S., Gojobori, T., Logares, R., Sala, M. M., VaquÃľ, D., Massana, R., Duarte, C. M., Acinas, S. G., and Gasol, J. M. (2024). Marine picoplankton metagenomes and MAGs from eleven vertical profiles obtained by the malaspina expedition. *Sci Data*, 11(1):154.

Sanchez, P., SebastiÃąn, M., Pernice, M., RodrÃ■guez-MartÃ■nez, R., Pesant, S., AgustÃ■, S., Gojobori, T., Logares, R., Sala, M. M., VaquÃľ, D., Massana, R., Duarte, C. M., Acinas, S. G., and Gasol, J. M. (2023). Marine picoplankton metagenomes from eleven vertical profiles obtained by the malaspina expedition in the tropical and subtropical oceans. *bioRxiv*, page 2023.02.06.526790.

Schattenhofer, M., Fuchs, B. M., Amann, R., Zubkov, M. V., Tarran, G. A., and Pernthaler, J. (2009). Latitudinal distribution of prokaryotic picoplankton populations in the atlantic ocean. *Environmental Microbiology*, 11(8):2078–2093.

Schechter, M. S. (2020). The history of metagenomics: An incomplete summary. Accessed: 2025-01-07.

Schmidt, T. S. B., Fullam, A., Ferretti, P., Orakov, A., Maistrenko, O. M., Ruscheweyh, H.-J., Letunic, I., Duan, Y., Van Rossum, T., Sunagawa, S., Mende, D. R., Finn, R. D., Kuhn, M., Pedro Coelho, L., and Bork, P. (2023). Spire: a searchable, planetary-scale microbiome resource. *Nucleic Acids Research*, 52(D1):D777–D783.

Schrenk, M. O., Kelley, D. S., Bolton, S. A., and Baross, J. A. (2004). Low archaeal diversity linked to subseafloor geochemical processes at the lost city hydrothermal field, mid–atlantic ridge. *Environmental Microbiology*, 6(10):1086–1095.

Schroer, W. F. (2023). *Metabolite Transport and Its Role in Marine Microbial Interactions*. PhD thesis, University of Georgia.

Schroer, W. F., Kepner, H. E., Uchimiya, M., Mejia, C., Rodriguez, L. T., Reisch, C. R., and Moran, M. A. (2023). Functional annotation and importance of marine bacterial transporters of plankton exometabolites. *ISME Communications*, 3(1).

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., DrÃűge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., JÃÿrgensen, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., DeMaere, M. Z., Chikhi, R., Nagarajan, N., Quince, C., Meyer, F., BalvoÄ■iÅńtÄŮ, M., Hansen, L. H., SÃÿrensen, S. J., Chia, B. K. H., Denis, B., Froula, J. L., Wang, Z., Egan, R., Don Kang, D., Cook, J. J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y.-W., Singer, S. W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M. D., Lingner, T., Lin, H.-H., Liao, Y.-C., Silva, G. G. Z., Cuevas, D. A., Edwards, R. A., Saha, S., Piro, V. C., Renard, B. Y., Pop, M., Klenk, H.-P., GÃűker, M., Kyrpides, N. C., Woyke, T., Vorholt, J. A., Schulze-Lefert, P., Rubin, E. M., Darling, A. E., Rattei, T., and McHardy, A. C. (2017). Critical assessment of metagenome interpretation-a benchmark of metagenomics software. *Nat. Methods*, 14(11):1063–1071.

Septer, A. N. and Visick, K. L. (2024). Lighting the way: how the vibrio fischeri model microbe reveals the complexity of earth's "simplest" life forms. *J. Bacteriol.*, 206(5):e0003524.

Shaiber, A., Willis, A. D., Delmont, T. O., Roux, S., Chen, L.-X., Schmid, A. C., Yousef, M., Watson, A. R., Lolans, K., Esen, Ö. C., Lee, S. T. M., Downey, N., Morrison, H. G., Dewhirst, F. E., Mark Welch, J. L., and Eren, A. M. (2020). Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol.*, 21(1):292.

Shragai, T., Tesla, B., Murdock, C., and Harrington, L. C. (2017). Zika and chikungunya: mosquito-borne viruses in a changing world. *Ann. N. Y. Acad. Sci.*, 1399(1):61–77.

Siegel, D. A., DeVries, T., CetiniÄĞ, I., and Bisson, K. M. (2023). Quantifying the ocean's biological pump and its carbon cycle impacts on global scales. *Ann. Rev. Mar. Sci.*, 15(1):329–356.

Simon, J.-C., Marchesi, J. R., Mougel, C., and Selosse, M.-A. (2019). Host-microbiota interactions: from holobiont theory to analysis. *Microbiome*, 7(1):5.

Simón-Soro, A., Tomás, I., Cabrera-Rubio, R., Catalan, M., Nyvad, B., and Mira, A. (2013). Microbial geography of the oral cavity. *Journal of dental research*, 92(7):616–621.

Smith, N. W., Shorten, P. R., Altermann, E. H., Roy, N. C., and McNabb, W. C. (2019). Hydrogen cross-feeders of the human gastrointestinal tract. *Gut microbes*, 10(3):270–288.

Soubrier, J., Steel, M., Lee, M. S. Y., Der Sarkissian, C., Guindon, S., Ho, S. Y. W., and Cooper, A. (2012). The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Mol. Biol. Evol.*, 29(11):3345–3358.

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L., and Ettema, T. J. G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, 521(7551):173–179.

Staley, J. T. and Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.*, 39:321–346.

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H., and DeLong, E. F. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of Bacteriology*, 178(3):591–599.

Steinegger, M. and Soding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*

Stepanauskas, R., Fergusson, E. A., Brown, J., Poulton, N. J., Tupper, B., LabontÃľ, J. M., Becraft, E. D., Brown, J. M., Pachiadaki, M. G., Povilaitis, T., Thompson, B. P., Mascena, C. J., Bellows, W. K., and Lubys, A. (2017). Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nat. Commun.*, 8(1):84.

Sul, W. J., Oliver, T. A., Ducklow, H. W., Amaral-Zettler, L. A., and Sogin, M. L. (2013). Marine bacteria exhibit a bipolar distribution. *Proc. Natl. Acad. Sci. U. S. A.*, 110(6):2342–2347.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B. T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans coordinators, Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M. B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S. G., and Bork, P. (2015). Ocean plankton. structure and function of the global ocean microbiome. *Science*, 348(6237):1261359.

Surana, N. K. and Kasper, D. L. (2012). The *yin yang* of bacterial polysaccharides: lessons learned from *B. fragilis* PSA. *Immunol. Rev.*, 245(1):13–26.

Suttle, C. A. (2005). Viruses in the sea. *Nature*, 437(7057):356âĂŞ361.

SzafraÅĎski, S. P., Slots, J., and Stiesch, M. (2021). The human oral phageome. *Periodontol. 2000*, 86(1):79–96.

Tagliabue, A., Bowie, A. R., Boyd, P. W., Buck, K. N., Johnson, K. S., and Saito, M. A. (2017). The integral role of iron in ocean biogeochemistry. *Nature*, 543(7643):51–59.

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, 28(1):33–36.

Thomas, M. K., Kremer, C. T., Klausmeier, C. A., and Litchman, E. (2012). A global pattern of thermal adaptation in marine phytoplankton. *Science*, 338(6110):1085–1088.

Tinta, T., KogovÅąek, T., Klun, K., Malej, A., Herndl, G. J., and Turk, V. (2019). Jellyfish-associated microbiome in the marine environment: Exploring its biotechnological potential. *Mar. Drugs*, 17(2):94.

Trivedi, P., Trivedi, C., Grinyer, J., Anderson, I. C., and Singh, B. K. (2016). Harnessing host-vector microbiome for sustainable plant disease management of phloem-limited bacteria. *Front. Plant Sci.*, 7:1423.

Tully, B. J., Graham, E. D., and Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data*, 5:170203.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., Egholm, M., Henrissat, B., Heath, A. C., Knight, R., and Gordon, J. I. (2008). A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480âĂŞ484.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature*, 449(7164):804–810.

Tyrrell, T. (1999). The relative influences of nitrogen and phosphorus on oceanic primary production. *Nature*, 400(6744):525–531.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43.

Ustick, L. J., Larkin, A. A., and Martiny, A. C. (2023). Global scale phylogeography of functional traits and microdiversity in prochlorococcus. *The ISME Journal*, 17(10):1671–1679.

Utter, D. R., Borisy, G. G., Eren, A. M., Cavanaugh, C. M., and Mark Welch, J. L. (2020). Metapangenomics of the oral microbiome provides insights into habitat adaptation and cultivar diversity. *Genome Biol.*, 21(1):293.

van Kessel, J. C. and Camilli, A. (2024). Vibrio cholerae: a fundamental model system for bacterial genetics and pathogenesis research. *J. Bacteriol.*, 206(11):e0024824.

Venskutonytė, R., Koh, A., Stenström, O., Khan, M. T., Lundqvist, A., Akke, M., Bäckhed, F., and Lindkvist-Petersson, K. (2021). Structural characterization of the microbial enzyme urocanate reductase mediating imidazole propionate production. *Nature communications*, 12(1):1347.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H., and Smith, H. O. (2004). Environmental genome shotgun sequencing of the sargasso sea. *Science*, 304(5667):66–74.

Veseli, I., Chen, Y. T., Schechter, M. S., Vanni, C., Fogarty, E. C., Watson, A. R., Jabri, B., Blekhman, R., Willis, A. D., Yu, M. K., Fernández-Guerra, A., Füssel, J., and Eren, A. M. (2024a). Microbes with higher metabolic independence are enriched in human gut microbiomes under stress.

Veseli, I., DeMers, M. A., Cooper, Z. S., Schechter, M. S., Miller, S., Weber, L., Smith, C. B., Rodriguez, L. T., Schroer, W. F., McIlvin, M. R., Lopez, P. Z., Saito, M., Dyhrman, S., Eren, A. M., Moran, M. A., and Braakman, R. (2024b). Digital microbe: a genome-informed data integration framework for team science on emerging model organisms. *Sci. Data*, 11(1):967.

Vineis, J. H., Ringus, D. L., Morrison, H. G., Delmont, T. O., Dalal, S., Raffals, L. H., Antonopoulos, D. A., Rubin, D. T., Eren, A. M., Chang, E. B., and Sogin, M. L. (2016). Patient-Specific *Bacteroides* Genome Variants in Pouchitis. *MBio*, 7(6):e01713–16, /mbio/7/6/e01713–16.atom.

Wardman, J. F., Bains, R. K., Rahfeld, P., and Withers, S. G. (2022). Carbohydrate-active enzymes (cazymes) in the gut microbiome. *Nature Reviews Microbiology*, 20(9):542âĂŞ556.

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009). Jalview version 2âĂŤa multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191.

Watson, A. R., Füssel, J., Veseli, I., DeLongchamp, J. Z., Silva, M., Trigodet, F., Lolans, K., Shaiber, A., Fogarty, E., Runde, J. M., Quince, C., Yu, M. K., Söylev, A., Morrison, H. G., Lee, S. T. M., Kao, D., Rubin, D. T., Jabri, B., Louie, T., and Eren, A. M. (2023). Metabolic independence drives gut microbial colonization and resilience in health and disease. *Genome Biology*, 24(1).

Watterson, W. J., Tanyeri, M., Watson, A. R., Cham, C. M., Shan, Y., Chang, E. B., Eren, A. M., and Tay, S. (2020). Droplet-based high-throughput cultivation for accurate screening of antibiotic resistant gut microbes. *Elife*, 9.

Weber, M., Teeling, H., Huang, S., Waldmann, J., Kassabgy, M., Fuchs, B. M., Klindworth, A., Klockow, C., Wichels, A., Gerdts, G., Amann, R., and Glöckner, F. O. (2011). Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics. *ISME J.*, 5(5):918–928.

Weinheimer, A. R. and Aylward, F. O. (2020). A distinct lineage of caudovirales that encodes a deeply branching multi-subunit RNA polymerase. *Nat. Commun.*, 11(1):4506.

Weyrich, L. S., Farrer, A. G., Eisenhofer, R., Arriola, L. A., Young, J., Selway, C. A., Handsley-Davis, M., Adler, C. J., Breen, J., and Cooper, A. (2019). Laboratory contamination over time during low-biomass sample analysis. *Mol. Ecol. Resour.*, 19(4):982–996.

Whitfield, C., Wear, S. S., and Sande, C. (2020). Assembly of bacterial capsular polysaccharides and exopolysaccharides. *Annu. Rev. Microbiol.*, 74(1):521–543.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., FranÃğois, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., MÃijller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *J. Open Source Softw.*, 4(43):1686.

Wilbert, S. A., Mark Welch, J. L., and Borisy, G. G. (2020). Spatial ecology of the human tongue dorsum microbiome. *Cell Rep.*, 30(12):4003–4015.e3.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data*, 3(1):160018.

Winter, S. E., Thiennimitr, P., Winter, M. G., Butler, B. P., Huseby, D. L., Crawford, R. W., Russell, J. M., Bevins, C. L., Adams, L. G., Tsolis, R. M., et al. (2010). Gut inflammation provides a respiratory electron acceptor for salmonella. *Nature*, 467(7314):426–429.

Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews*, 51(2):221–271.

Woese, C. R. and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090.

Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, 15(3):R46.

Woodcroft, B. J., Aroney, S. T. N., Zhao, R., Cunningham, M., Mitchell, J. A. M., Blackall, L., and Tyson, G. W. (2024). SingleM and sandpiper: Robust microbial taxonomic profiles from metagenomic data. *bioRxiv*, page 2024.01.30.578060.

Woolfson, A. (2015). Origins of life: An improbable journey. *Nature*, 520(7549):617–618.

Woting, A. and Blaut, M. (2016). The intestinal microbiota in metabolic disease. *Nutrients*, 8(4):202.

Woyke, T., Doud, D. F. R., and Schulz, F. (2017). The trajectory of microbial single-cell sequencing. *Nat. Methods*, 14(11):1045–1054.

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., Kunin, V., Goodwin, L., Wu, M., Tindall, B. J., et al. (2009). A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature*, 462(7276):1056–1060.

Wu, D., Seshadri, R., Kyrpides, N. C., and Ivanova, N. N. (2025). A metagenomic perspective on the microbial prokaryotic genome census. *Science Advances*, 11(3).

Yang, Z. (1995). A space-time process model for the evolution of DNA sequences. *Genetics*, 139(2):993–1005.

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., and Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.*, 40(Web Server issue):W445–51.

Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., Eisen, J. A., Heidelberg, K. B., Manning, G., Li, W., Jaroszewski, L., Cieplak, P., Miller, C. S., Li, H., Mashiyama, S. T., Joachimiak, M. P., van Belle, C., Chandonia, J.-M., Soergel, D. A., Zhai, Y., Natarajan, K., Lee, S., Raphael, B. J., Bafna, V., Friedman, R., Brenner, S. E., Godzik, A., Eisenberg, D., Dixon, J. E., Taylor, S. S., Strausberg, R. L., Frazier, M., and Venter, J. C. (2007). The sorcerer ii global ocean sampling expedition: Expanding the universe of protein families. *PLoS Biology*, 5(3):e16.

Zhou, Z., Tran, P. Q., Adams, A. M., Kieft, K., Breier, J. A., Fortunato, C. S., Sheik, C. S., Huber, J. A., Li, M., Dick, G. J., and Anantharaman, K. (2023). Sulfur cycling connects microbiomes and biogeochemistry in deep-sea hydrothermal plumes. *ISME J.*, 17(8):1194–1207.