Ribosomal protein phylogeography offers quantitative insights into the efficacy of genome-resolved surveys of microbial communities

Matthew S. Schechter[1,2,*], Florian Trigodet[3,4], Iva A. Veseli[3,4], Samuel E. Miller[5], Matthew L. Klein[2], Metehan Sever[3,4], Loïs Maignien[6], Tom O. Delmont[7,8], Samuel H. Light[2,9,*,†], A. Murat Eren[3,4,5,10,11,*,†]

[1]Committee on Microbiology, University of Chicago, Chicago, IL 60637, USA; [2]Duchossois Family Institute, University of Chicago, Chicago, IL 60637, USA; [3]Helmholtz Institute for Functional Marine Biodiversity, 26129 Oldenburg, Germany; [4]Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, 27570 Bremerhaven, Germany; [5]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA 02543, USA; [6]University of Brest, CNRS, IFREMER, EMR 6002 BIOMEX, Unité Biologie et Écologie des Écosystèmes Marins Profonds BEEP, F-29280 Plouzané, France; [7]Génomique Métabolique du Genoscope, Institut François Jacob, CEA, CNRS, University of Évry Val d'Essonne, Université Paris-Saclay, Evry, France; [8]Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara GOSEE, Paris, France; [9]Department of Microbiology, University of Chicago, Chicago, IL 60637, USA; [10]Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, 26129 Oldenburg, Germany; [11]Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany.

* Correspondence: meren@hifmb.de, samlight@uchicago.edu, and mschechter@uchicago.edu

† Co-senior authors: samlight@uchicago.edu and meren@hifmb.de

Running Title

Ribosomal protein phylogeography with EcoPhylo

Keywords

phylogeography, metagenomics, metagenome-assembled genomes, single amplified genomes

Supplementary Information

doi:10.6084/m9.figshare.28200050

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.

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 (C. T. 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. 2023; J. Chen et al. 2024a), (3) offers resources to search for novel biosynthetic capabilities and natural products (Paoli et al. 2022; J. Chen et al. 2024b), (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. 2024).

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) (L.-X. Chen et al. 2020), and (3) the use of microfluidics and cell sorting that enables the recovery of single amplified genomes (SAGs) (Woyke, Doud, and Schulz 2017). These strategies have not only been used in large-scale characterization of many of the Earth's 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 involves 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, Larkin, and Martiny 2023) or RNA and DNA polymerases to identify and guide the genomic recovery of major viral clades (Weinheimer and Aylward 2020; Gaïa 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 and Eisen 2008; Crits-Christoph et al. 2022), integrating ribosomal protein phylogenies with metagenomic read recruitment to track individual clades of microbes (Hug et al. 2013; 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.

Results

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 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 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.

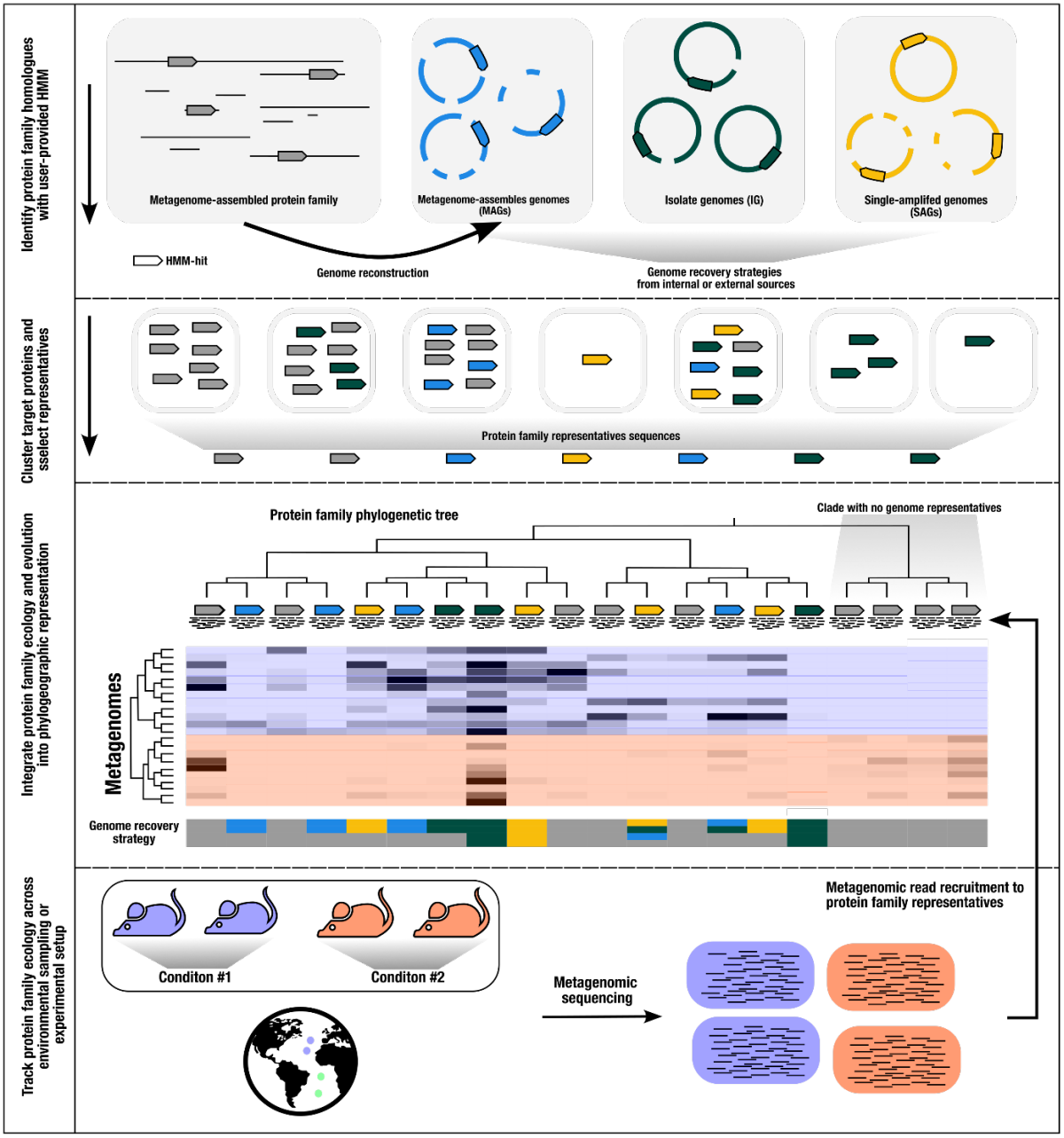

Figure 1: Schematic of the EcoPhylo workflow applied to a single gene family. The proposed workflow integrates biogeography from metagenomic read recruitment and protein phylogenetics to display the phylogeographical distribution of closely related lineages. When including genome sources, the workflow 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 gene family.

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 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.

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; T. 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 ribosomal 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. MAGs (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 2A, 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 2A, 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, Dewhirst, and Borisy 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 2A, Supplementary information). Overall, the ecological insights revealed by *rpL19* recapitulated known ecology of oral microbes (Mark Welch, Dewhirst, and Borisy 2019) and provided a framework to assess genome recovery rates.

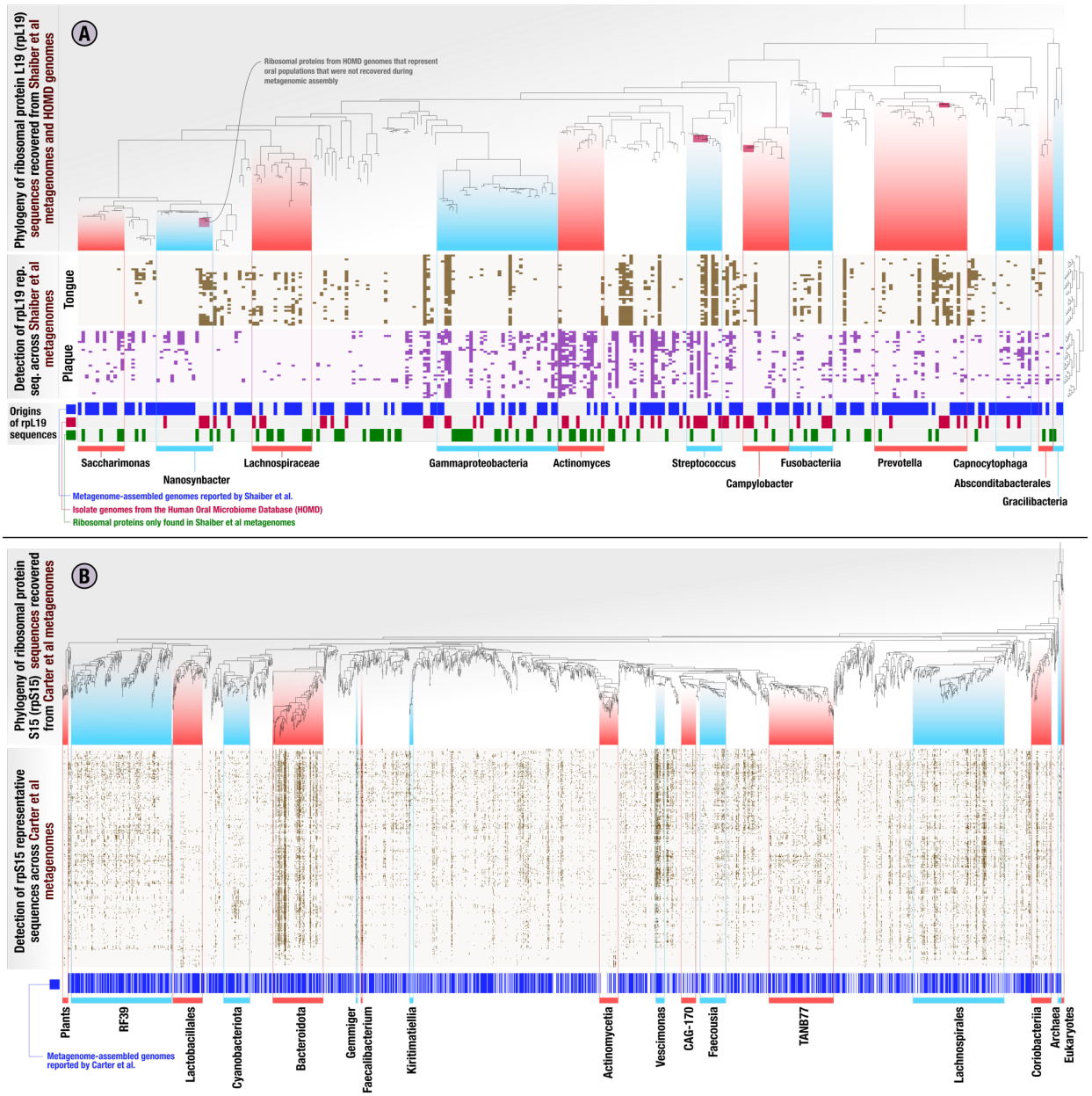

Figure 2: Ribosomal protein phylogeny and detection patterns across metagenomes from the human oral cavity and gut microbiomes. In the heatmaps in both panels, each column represents a ribosomal protein representative sequence, each row represents a metagenome, and each data point indicates whether a given ribosomal protein was detected in a given metagenome. The columns of heatmaps 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 the ribosomal protein detection patterns across metagenomes. The 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 that were 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. 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).

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 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-read metagenomic assembly approaches (Figure 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*: 62.8%, *rpS2*: 53.2%) (Figure 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 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 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 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 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 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 2), revealed a relatively high bacterial MAG recovery rate of 67.7% (*rpS16*: 72.8%, *rpL19*: 69.5%) (Figure 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 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.

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 catalogues for bacteria and archaea with new MAGs (Delmont et al. 2018b; Tully, Graham, and Heidelberg 2018; Paoli et al. 2022) and SAGs (Pachiadaki et al. 2019; Martínez-Pérez 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. 2019a; 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 (Sánchez 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 0.22µm to 3µm (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 (Martínez-Pérez 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).

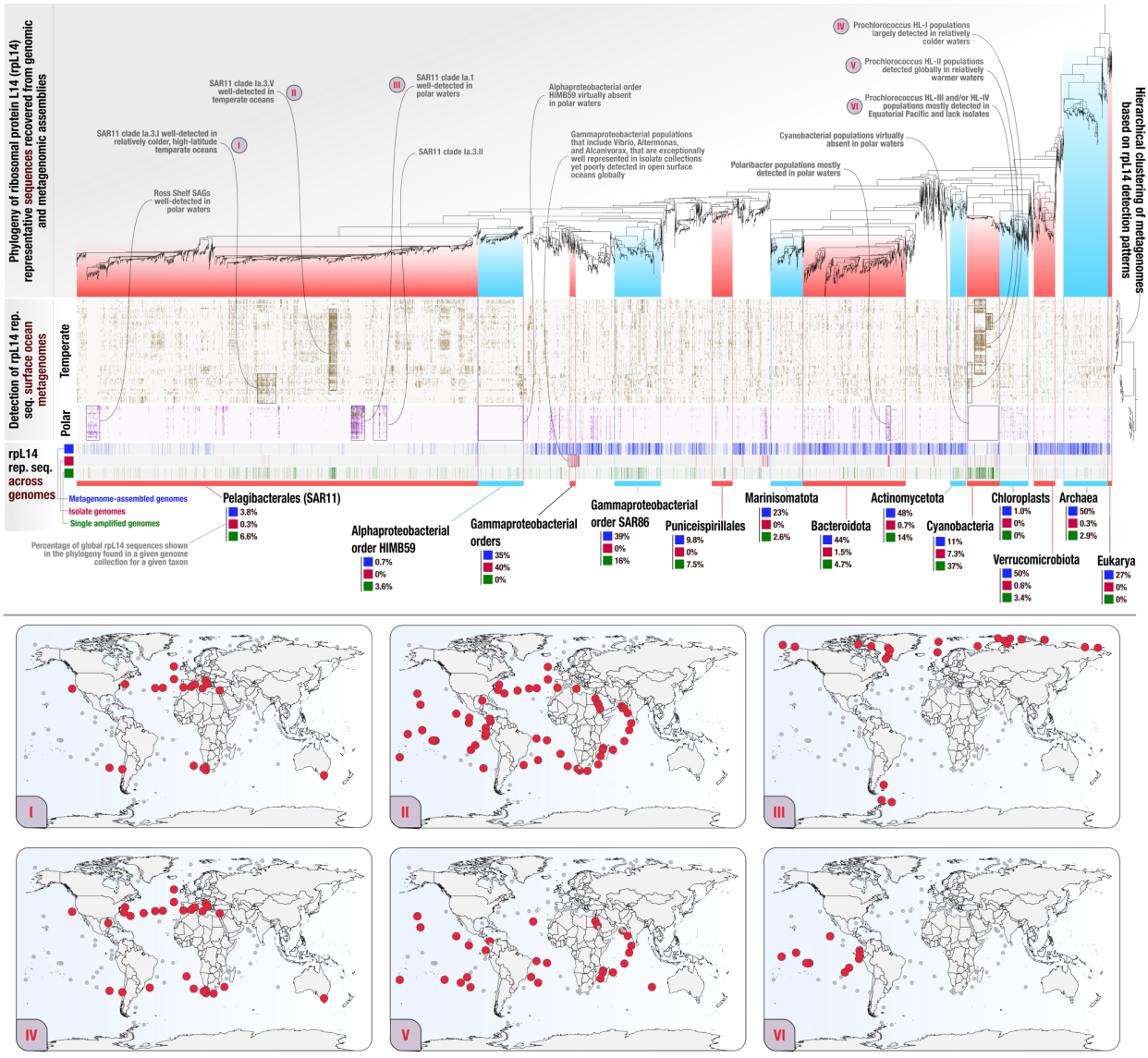

Figure 3: Ribosomal protein L14 phylogeny and detection patterns across metagenomes from the global surface ocean (depth < 30 m; size fraction: 0.22 to 1.6 µm, 0.22 to 3 µm). In the heatmap of panel (A), each column represents a ribosomal protein representative sequence, each row represents a metagenome, and each data point indicates whether a given ribosomal protein was detected in a given metagenome. 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).

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, 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). Furthermore, the *rpL14* phylogeography captured well-understood biogeographical patterns of prevalent pelagic taxa (Figure 3), in agreement with previous studies that showed the dominance of SAR11 subclade Ia.3.V in temperate waters (Delmont et al. 2019b) and the exclusivity of cold-adapted SAR11 clades Ia.1 and Ia.3.II to polar regions (M. V. Brown et al. 2012; Delmont et al. 2019a). It also

corroborated the global distribution of *Prochlorococcus* HL-II in temperate waters (Johnson et al. 2006; Biller et al. 2015; Ustick, Larkin, and Martiny 2023) and the contrasting distribution of this group with *Prochlorococcus* HL-III and *Prochlorococcus* HL-IV, which are mainly found in the Equatorial Pacific (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, 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, 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 *Verrucomicrobiota* (*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 (L.-X. Chen et al. 2020).

De novo characterization of *rpL14* sequences with EcoPhylo uncovers the vast diversity within *Pelagibacterales* compared to the other clades (Figure 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, revealing yet another representation of its immense phylogenetic diversity (Morris et al. 2002; M. V. 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 ~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), and highlights the importance of ongoing cultivation efforts to improve its genomic representation (Freel et al. 2024).

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). 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, Aluwihare, and Azam 2014; Manck et al. 2022; Henríquez-Castillo et al. 2022; Z. Lu et al. 2024; Halloran et al. 2025)*, and *Alcanivorax* (Sabirova et al. 2008; Naether et al. 2013; Manoj Prasad et al. 2019; M. Prasad et al. 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, 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*, including *HIMB59* at 5.06% (*rpS8*: 5.61%, *rpS11*: 5.29%) and *SAR11* at 6.35% (*rpS8*: 7.56%, *rpS11*:

7.82%) (Figure 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 (Martínez-Pérez 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, Supplementary Figure 1), despite being sampled from only 9 surface ocean sampling sites (Martínez-Pérez 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), 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.

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 (J. 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

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

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

Materials and Methods

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 (Köster 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 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 Söding 2017), (5) use the translated representative sequences to calculate a multiple sequence alignment (MSA) using (Edgar 2004) with the `-maxiters 2` flag (Edgar 2004), trim the alignment by removing columns of the alignment with trimAL with the '-gappyout' flag (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 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, Dehal, and Arkin 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, (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.

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

606 families to limit non-specific read recruitment while maximizing taxonomic resolution. We applied
607 the EcoPhylo workflow across the three CAMI biome synthetic genomic/metagenomic datasets
608 (Marine, Plant-associated, and Strain-madness). As an initial step, we identified the top five most
609 frequent ribosomal gene families that were detected in single-copy in the associated genomic
610 collections for each synthetic metagenomic dataset. We then conducted a parameter grid search,
611 spanning 95%-100% nucleotide similarity parameter grid search (Meyer et al. 2022). Next, we
612 measured the amount of non-specific read recruitment in each EcoPhylo iteration, i.e. reads with
613 equal mapping scores between their primary and secondary alignments (multi-mapped reads),
614 with the following Samtools command: `samtools view $sample | grep XS:i | cut -f12-13 | sed
615 's/..:i://g' | awk '$1==$2' | wc -l`. The percentage of non-specific read recruitment was calculated
616 by dividing the number of multi-mapped reads by the total number of reads mapped to the
617 representative dataset. With this, we identified that nucleotide clustering thresholds greater than
618 97% began to show signs of non-specific read recruitment (Supplementary information).

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

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:10.6084/m9.figshare.12217805, doi:10.6084/m9.figshare.12217961).

For the Hadza tribe human gut microbiome analysis we followed the data download guidelines 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 (Martínez-Pérez et al. 2022) and isolate genomes for SAR11 and *Prochlorococcus* (Delmont and Eren 2018; Delmont et al. 2019a). 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. 2019a). The final input surface ocean genome dataset contained 1,474 SAGs, 1,723 isolate genomes, and 7,282 MAGs.

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: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 PRJEB45951 and PRJEB5245228 (Paoli et al. 2022; Sánchez et al. 2023). All metagenomes and associated assembly accessions can be found at Supplementary Table 1.

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, Dohm, and Himmelbauer 2011). All Snakemake workflows in this manuscript leveraged Snakemake v7.32.4 (Köster and Rahmann 2012).

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, Reuter, and Drost 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".

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).

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` in (Eddy 2011)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

718 from `hmmsearch --domtblout` output file leading us to to identify an 80% HMM hit model
719 coverage as an optimal filtering threshold to identify ribosomal proteins. Next, we restarted the
720 workflow but re-modified the second rule parameter `anvi-script-filter-hmm-hits-table --min-model-
721 coverage 0.8` to filter for HMMs hits with at least 80% model alignment coverage. Finally, we
722 extracted all ribosomal gene families from the genome dataset with anvi'o program `anvi-script-
723 gen-hmm-hits-matrix-across-genomes` and visualized the genome detection and SCG copy
724 number across the dataset in in R using the Tidyverse (Wickham et al. 2019).

725 Detection of whole genomes in metagenomic data

726 In some cases, ribosomal proteins clustering at 97% brought together large groups of highly
727 similar isolate genomes. To identify the specific genome that is detected in the metagenomic
728 datasets, we re-clustered the target EcoPhylo protein at 98% to resolve sequence clusters and
729 thus increase the number of representative sequences. We then used the whole genomes
730 associated with the new, larger set of representative proteins to explore their distribution in
731 metagenomes by performing the anvi'o metagenomic workflow (Shaiber et al. 2020). Our
732 threshold for detection of a whole-genome in metagenomic data was 50% (percent of genome
733 covered by at least one read from metagenomic read recruitment), which was found to be efficient
734 for human oral cavity microbes (Utter et al. 2020).

735 Genome recovery rate estimations

736 Genome recovery rates were estimated to measure which individual or combination of genome
737 types (MAGs, SAGs, isolate genomes) most effectively sampled clades in the ribosomal protein
738 phylogenetic trees calculated during the EcoPhylo workflow. To calculate genome recovery rates
739 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`.

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 2 and *rpL14* in Figure 3).

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.

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).

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.

Competing interests

Authors declare that they have no conflicts of interest.

References

Al-Shayeb, Basem, Petr Skopintsev, Katarzyna M. Soczek, Elizabeth C. Stahl, Zheng Li, Evan Groover, Dylan Smock, et al. 2022. “Diverse Virus-Encoded CRISPR-Cas Systems Include Streamlined Genome Editors.” *Cell* 185 (24): 4574–86.e16.

Baker-Austin, Craig, Joaquin Trinanes, Narjol Gonzalez-Escalona, and Jaime Martinez-Urtaza. 2017. “Non-Cholera Vibrios: The Microbial Barometer of Climate Change.” *Trends in Microbiology* 25 (1): 76–84.

Biller, Steven J., Paul M. Berube, Keven Dooley, Madeline Williams, Brandon M. Satinsky, Thomas Hackl, Shane L. Hogle, et al. 2018. “Marine Microbial Metagenomes Sampled across Space and Time.” *Scientific Data* 5 (September):180176.

Biller, Steven J., Paul M. Berube, Debbie Lindell, and Sallie W. Chisholm. 2015. “Prochlorococcus: The Structure and Function of Collective Diversity.” *Nature Reviews. Microbiology* 13 (1): 13–27.

Bowen, William H., Robert A. Burne, Hui Wu, and Hyun Koo. 2018. “Oral Biofilms: Pathogens, Matrix, and Polymicrobial Interactions in Microenvironments.” *Trends in Microbiology* 26 (3): 229–42.

Bowers, Robert M., Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T. B. K. Reddy, Frederik Schulz, et al. 2017. “Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea.” *Nature Biotechnology* 35 (8): 725–31.

Brown, Christopher T., Laura A. Hug, Brian C. Thomas, Itai Sharon, Cindy J. Castelle, Andrea Singh, Michael J. Wilkins, Kelly C. Wrighton, Kenneth H. Williams, and Jillian F. Banfield. 2015. “Unusual Biology across a Group Comprising More than 15% of Domain Bacteria.” *Nature* 523 (7559): 208–11.

Brown, Mark V., Federico M. Lauro, Matthew Z. DeMaere, Les Muir, David Wilkins, Torsten Thomas, Martin J. Riddle, et al. 2012. “Global Biogeography of SAR11 Marine Bacteria.” *Molecular Systems Biology* 8 (1): 595.

Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost. 2021. “Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND.” *Nature Methods* 18 (4): 366–68.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. “trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” *Bioinformatics* 25 (15): 1972–73.

Carter, Matthew M., Matthew R. Olm, Bryan D. Merrill, Dylan Dahan, Surya Tripathi, Sean P. Spencer, Feiqiao B. Yu, et al. 2023. “Ultra-Deep Sequencing of Hadza Hunter-Gatherers Recovers Vanishing Gut Microbes.” *Cell*, June. https://doi.org/10.1016/j.cell.2023.05.046.

Chang, Tianyi, Gregory S. Gavelis, Julia M. Brown, and Ramunas Stepanauskas. 2024. “Genomic Representativeness and Chimerism in Large Collections of SAGs and MAGs of Marine Prokaryoplankton.” *Microbiome* 12 (1): 126.

Chen, Jianwei, Yangyang Jia, Ying Sun, Kun Liu, Changhao Zhou, Chuan Liu, Denghui Li, et al. 2024a. “Global Marine Microbial Diversity and Its Potential in Bioprospecting.” *Nature* 633 (8029): 371–79.

———. 2024b. “Global Marine Microbial Diversity and Its Potential in Bioprospecting.” *Nature*, September. https://doi.org/10.1038/s41586-024-07891-2.

Chen, Lin-Xing, Karthik Anantharaman, Alon Shaiber, A. Murat Eren, and Jillian F. Banfield. 2020. “Accurate and Complete Genomes from Metagenomes.” *Genome Research* 30 (3): 315–33.

Chen, Tsute, Wen-Han Yu, Jacques Izard, Oxana V. Baranova, Abirami Lakshmanan, and

826 Floyd E. Dewhirst. 2010. “The Human Oral Microbiome Database: A Web Accessible
827 Resource for Investigating Oral Microbe Taxonomic and Genomic Information.” *Database:*
828 *The Journal of Biological Databases and Curation* 2010 (0): baq013.
829 Cornman, Andre, Jacob West-Roberts, Antonio Pedro Camargo, Simon Roux, Martin
830 Beracochea, Milot Mirdita, Sergey Ovchinnikov, and Yunha Hwang. 2024. “The OMG
831 Dataset: An Open MetaGenomic Corpus for Mixed-Modality Genomic Language Modeling.”
832 *bioRxiv*. https://doi.org/10.1101/2024.08.14.607850.
833 Crits-Christoph, Alexander, Spencer Diamond, Basem Al-Shayeb, Luis Valentin-Alvarado, and
834 Jillian F. Banfield. 2022. “A Widely Distributed Genus of Soil Acidobacteria Genomically
835 Enriched in Biosynthetic Gene Clusters.” *ISME Communications* 2 (1): 1–8.
836 Cross, Karissa L., James H. Campbell, Manasi Balachandran, Alisha G. Campbell, Connor J.
837 Cooper, Ann Griffen, Matthew Heaton, et al. 2019. “Targeted Isolation and Cultivation of
838 Uncultivated Bacteria by Reverse Genomics.” *Nature Biotechnology* 37 (11): 1314–21.
839 Delmont, Tom O., and A. Murat Eren. 2018. “Linking Pangenomes and Metagenomes: The
840 Prochlorococcus Metapangenome.” *PeerJ* 6 (January):e4320.
841 Delmont, Tom O., Evan Kiefl, Ozsel Kilinc, Ozcan C. Esen, Ismail Uysal, Michael S. Rappé,
842 Steven Giovannoni, and A. Murat Eren. 2019a. “Single-Amino Acid Variants Reveal
843 Evolutionary Processes That Shape the Biogeography of a Global SAR11 Subclade.” *eLife*
844 8 (September). https://doi.org/10.7554/eLife.46497.
845 ———. 2019b. “Single-Amino Acid Variants Reveal Evolutionary Processes That Shape the
846 Biogeography of a Global SAR11 Subclade.” *eLife* 8 (September).
847 https://doi.org/10.7554/eLife.46497.
848 Delmont, Tom O., Christopher Quince, Alon Shaiber, Özcan C. Esen, Sonny Tm Lee, Michael
849 S. Rappé, Sandra L. McLellan, Sebastian Lücker, and A. Murat Eren. 2018a. “Nitrogen-
850 Fixing Populations of Planctomycetes and Proteobacteria Are Abundant in Surface Ocean
851 Metagenomes.” *Nature Microbiology* 3 (7): 804–13.
852 ———. 2018b. “Nitrogen-Fixing Populations of Planctomycetes and Proteobacteria Are
853 Abundant in Surface Ocean Metagenomes.” *Nature Microbiology* 3 (7): 804–13.
854 Dewhirst, Floyd E., Tuste Chen, Jacques Izard, Bruce J. Paster, Anne C. R. Tanner, Wen-Han
855 Yu, Abirami Lakshmanan, and William G. Wade. 2010. “The Human Oral Microbiome.”
856 *Journal of Bacteriology* 192 (19): 5002–17.
857 Diamond, Spencer, Peter F. Andeer, Zhou Li, Alexander Crits-Christoph, David Burstein, Karthik
858 Anantharaman, Katherine R. Lane, et al. 2019. “Mediterranean Grassland Soil C-N
859 Compound Turnover Is Dependent on Rainfall and Depth, and Is Mediated by Genomically
860 Divergent Microorganisms.” *Nature Microbiology* 4 (8): 1356–67.
861 Dixon, Philip. 2003. “VEGAN, a Package of R Functions for Community Ecology.” *Journal of*
862 *Vegetation Science: Official Organ of the International Association for Vegetation Science*
863 14 (6): 927–30.
864 Durrant, Matthew G., Alison Fanton, Josh Tycko, Michaela Hinks, Sita S. Chandrasekaran,
865 Nicholas T. Perry, Julia Schaepe, et al. 2023. “Systematic Discovery of Recombinases for
866 Efficient Integration of Large DNA Sequences into the Human Genome.” *Nature*
867 *Biotechnology* 41 (4): 488–99.
868 Eddy, Sean R. 2011. “Accelerated Profile HMM Searches.” *PLoS Computational Biology* 7 (10):
869 e1002195.
870 Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High
871 Throughput.” *Nucleic Acids Research* 32 (5): 1792–97.
872 Emerson, Joanne B., Brian C. Thomas, Walter Alvarez, and Jillian F. Banfield. 2016.
873 “Metagenomic Analysis of a High Carbon Dioxide Subsurface Microbial Community
874 Populated by Chemolithoautotrophs and Bacteria and Archaea from Candidate Phyla.”
875 *Environmental Microbiology* 18 (6): 1686–1703.
876 Eren, A. Murat, and Jillian F. Banfield. 2024. “Modern Microbiology: Embracing Complexity

through Integration across Scales." *Cell* 187 (19): 5151–70.
Eren, A. Murat, Gary G. Borisy, Susan M. Huse, and Jessica L. Mark Welch. 2014. "Oligotyping Analysis of the Human Oral Microbiome." *Proceedings of the National Academy of Sciences of the United States of America* 111 (28): E2875–84.
Eren, A. Murat, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. 2015. "Anvi'o: An Advanced Analysis and Visualization Platform for 'Omics Data." *PeerJ* 3 (October):e1319.
Eren, A. Murat, Evan Kiefl, Alon Shaiber, Iva Veseli, Samuel E. Miller, Matthew S. Schechter, Isaac Fink, et al. 2021. "Community-Led, Integrated, Reproducible Multi-Omics with Anvi'o." *Nature Microbiology* 6 (1): 3–6.
Eren, A. Murat, Joseph H. Vineis, Hilary G. Morrison, and Mitchell L. Sogin. 2013. "A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology." *PloS One* 8 (6): e66643.
Escapa, Isabel F., Tsute Chen, Yanmei Huang, Prasad Gajare, Floyd E. Dewhirst, and Katherine P. Lemon. 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). https://doi.org/10.1128/mSystems.00187-18.
Freel, Kelle C., Sarah J. Tucker, Evan B. Freel, Stephen J. Giovannoni, A. Murat Eren, and Michael S. Rappe. 2024. "New Isolate Genomes and Global Marine Metagenomes Resolve Ecologically Relevant Units of SAR11." *Microbiology*. bioRxiv. https://www.biorxiv.org/content/10.1101/2024.12.24.630191v2.
Gaïa, Morgan, Lingjie Meng, Eric Pelletier, Patrick Forterre, Chiara Vanni, Antonio Fernandez-Guerra, Olivier Jaillon, et al. 2023. "Mirusviruses Link Herpesviruses to Giant Viruses." *Nature* 616 (7958): 783–89.
Giovannoni, Stephen J. 2017. "SAR11 Bacteria: The Most Abundant Plankton in the Oceans." *Annual Review of Marine Science* 9 (January):231–55.
Halloran, Kathryn H., Rogier Braakman, Allison Coe, Gretchen Swarr, Melissa C. Kido Soule, Sallie W. Chisholm, and Elizabeth B. Kujawinski. 2025. "Uptake of Prochlorococcus-Derived Metabolites by Alteromonas Macleodii MIT1002 Shows High Levels of Substrate Specificity." *Microbiology*. bioRxiv. https://www.biorxiv.org/content/10.1101/2025.01.10.632383v1.
Hamilton, Trinity L., Roderick J. Bovee, Sarah R. Sattin, Wiebke Mohr, William P. Gilhooly 3rd, Timothy W. Lyons, Ann Pearson, and Jennifer L. Macalady. 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." *Frontiers in Microbiology* 7 (April):598.
Henríquez-Castillo, Carlos, Alvaro M. Plominsky, Salvador Ramírez-Flandes, Anthony D. Bertagnolli, Frank J. Stewart, and Osvaldo Ulloa. 2022. "Metaomics Unveils the Contribution of Alteromonas Bacteria to Carbon Cycling in Marine Oxygen Minimum Zones." *Frontiers in Marine Science* 9 (September). https://doi.org/10.3389/fmars.2022.993667.
He, Xuesong, Jeffrey S. McLean, Anna Edlund, Shibu Yooseph, Adam P. Hall, Su-Yang Liu, Pieter C. Dorrestein, et al. 2015. "Cultivation of a Human-Associated TM7 Phylotype Reveals a Reduced Genome and Epibiotic Parasitic Lifestyle." *Proceedings of the National Academy of Sciences of the United States of America* 112 (1): 244–49.
Hosokawa, Masahito, and Yohei Nishikawa. 2024. "Tools for Microbial Single-Cell Genomics for Obtaining Uncultured Microbial Genomes." *Biophysical Reviews* 16 (1): 69–77.
Huang, Sijun, Steven W. Wilhelm, H. Rodger Harvey, Karen Taylor, Nianzhi Jiao, and Feng Chen. 2012. "Novel Lineages of Prochlorococcus and Synechococcus in the Global Oceans." *The ISME Journal* 6 (2): 285–97.

928 Hug, Laura A., Cindy J. Castelle, Kelly C. Wrighton, Brian C. Thomas, Itai Sharon, Kyle R.
929 Frischkorn, Kenneth H. Williams, Susannah G. Tringe, and Jillian F. Banfield. 2013.
930 “Community Genomic Analyses Constrain the Distribution of Metabolic Traits across the
931 Chloroflexi Phylum and Indicate Roles in Sediment Carbon Cycling.” *Microbiome* 1 (1): 22.
932 Hwang, Yunha, Andre L. Cornman, Elizabeth H. Kellogg, Sergey Ovchinnikov, and Peter R.
933 Girguis. 2024. “Genomic Language Model Predicts Protein Co-Regulation and Function.”
934 *Nature Communications* 15 (1): 2880.
935 Hyatt, Doug, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren
936 J. Hauser. 2010. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site
937 Identification.” *BMC Bioinformatics* 11 (March):119.
938 Imachi, Hiroyuki, Masaru K. Nobu, Nozomi Nakahara, Yuki Morono, Miyuki Ogawara, Yoshihiro
939 Takaki, Yoshinori Takano, et al. 2020. “Isolation of an Archaeon at the Prokaryote–
940 eukaryote Interface.” *Nature*. https://doi.org/10.1038/s41586-019-1916-6.
941 Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and
942 Srinivas Aluru. 2018. “High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals
943 Clear Species Boundaries.” *Nature Communications* 9 (1): 5114.
944 Jiang, Cheng-Ying, Libing Dong, Jian-Kang Zhao, Xiaofang Hu, Chaohua Shen, Yuxin Qiao,
945 Xinyue Zhang, et al. 2016. “High-Throughput Single-Cell Cultivation on Microfluidic Streak
946 Plates.” *Applied and Environmental Microbiology* 82 (7): 2210–18.
947 Johnson, Zackary I., Erik R. Zinser, Allison Coe, Nathan P. McNulty, E. Malcolm S. Woodward,
948 and Sallie W. Chisholm. 2006. “Niche Partitioning among Prochlorococcus Ecotypes along
949 Ocean-Scale Environmental Gradients.” *Science (New York, N.Y.)* 311 (5768): 1737–40.
950 Kauffman, Kathryn M., Fatima A. Hussain, Joy Yang, Philip Arevalo, Julia M. Brown, William K.
951 Chang, David VanInsberghe, et al. 2018. “A Major Lineage of Non-Tailed dsDNA Viruses
952 as Unrecognized Killers of Marine Bacteria.” *Nature* 554 (7690): 118–22.
953 Kent, Alyssa G., Steven E. Baer, Céline Mouginot, Jeremy S. Huang, Alyse A. Larkin, Michael
954 W. Lomas, and Adam C. Martiny. 2019. “Parallel Phylogeography of Prochlorococcus and
955 Synechococcus.” *The ISME Journal* 13 (2): 430–41.
956 Kent, Alyssa G., Chris L. Dupont, Shibu Yooseph, and Adam C. Martiny. 2016. “Global
957 Biogeography of Prochlorococcus Genome Diversity in the Surface Ocean.” *The ISME*
958 *Journal* 10 (8): 1856–65.
959 Kessel, Julia C. van, and Andrew Camilli. 2024. “Vibrio Cholerae: A Fundamental Model System
960 for Bacterial Genetics and Pathogenesis Research.” *Journal of Bacteriology* 206 (11):
961 e0024824.
962 Köster, Johannes, and Sven Rahmann. 2012. “Snakemake--a Scalable Bioinformatics Workflow
963 Engine.” *Bioinformatics (Oxford, England)* 28 (19): 2520–22.
964 Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.”
965 *Nature Methods* 9 (4): 357–59.
966 Lu, Jennifer, Florian P. Breitwieser, Peter Thielen, and Steven L. Salzberg. 2017. “Bracken:
967 Estimating Species Abundance in Metagenomics Data.” *PeerJ. Computer Science* 3
968 (e104): e104.
969 Lu, Zhiying, Elizabeth Entwistle, Matthew D. Kuhl, Alexander R. Durrant, Marcelo Malisano
970 Barreto Filho, Anuradha Goswami, and J. Jeffrey Morris. 2024. “Coevolution of Marine
971 Phytoplankton and Alteromonas Bacteria in Response to pCO2 and Co-Culture.” *The ISME*
972 *Journal*, December, wrae259.
973 Ma, Bin, Caiyu Lu, Yiling Wang, Jingwen Yu, Kankan Zhao, Ran Xue, Hao Ren, et al. 2023. “A
974 Genomic Catalogue of Soil Microbiomes Boosts Mining of Biodiversity and Genetic
975 Resources.” *Nature Communications* 14 (1): 7318.
976 Malmstrom, Rex R., Sébastien Rodrigue, Katherine H. Huang, Libusha Kelly, Suzanne E. Kern,
977 Anne Thompson, Sara Roggensack, Paul M. Berube, Matthew R. Henn, and Sallie W.
978 Chisholm. 2013. “Ecology of Uncultured Prochlorococcus Clades Revealed through Single-

Cell Genomics and Biogeographic Analysis." *The ISME Journal* 7 (1): 184–98.
Manck, Lauren E., Jiwoon Park, Benjamin J. Tully, Alfonso M. Poire, Randelle M. Bundy, Christopher L. Dupont, and Katherine A. Barbeau. 2022. "Petrobactin, a Siderophore Produced by Alteromonas, Mediates Community Iron Acquisition in the Global Ocean." *The ISME Journal* 16 (2): 358–69.
Manghi, Paolo, Aitor Blanco-Míguez, Serena Manara, Amir NabiNejad, Fabio Cumbo, Francesco Beghini, Federica Armanini, et al. 2023. "MetaPhlAn 4 Profiling of Unknown Species-Level Genome Bins Improves the Characterization of Diet-Associated Microbiome Changes in Mice." *Cell Reports* 42 (5): 112464.
Mark Welch, Jessica L., Floyd E. Dewhirst, and Gary G. Borisy. 2019. "Biogeography of the Oral Microbiome: The Site-Specialist Hypothesis." *Annual Review of Microbiology* 73 (September):335–58.
Martínez-Pérez, Clara, Chris Greening, Sean K. Bay, Rachael J. Lappan, Zihao Zhao, Daniele De Corte, Christina Hulbe, et al. 2022. "Phylogenetically and Functionally Diverse Microorganisms Reside under the Ross Ice Shelf." *Nature Communications* 13 (1): 117.
Matheus Carnevali, Paula B., Adi Lavy, Alex D. Thomas, Alexander Crits-Christoph, Spencer Diamond, Raphaël Méheust, Matthew R. Olm, et al. 2021. "Meanders as a Scaling Motif for Understanding of Floodplain Soil Microbiome and Biogeochemical Potential at the Watershed Scale." *Microbiome* 9 (1): 121.
McMurdie, Paul J., and Susan Holmes. 2013. "Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data." *PloS One* 8 (4): e61217.
Meyer, Fernando, Adrian Fritz, Zhi-Luo Deng, David Koslicki, Till Robin Lesker, Alexey Gurevich, Gary Robertson, et al. 2022. "Critical Assessment of Metagenome Interpretation: The Second Round of Challenges." *Nature Methods* 19 (4): 429–40.
Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era." *Molecular Biology and Evolution* 37 (5): 1530–34.
Minoche, André E., Juliane C. Dohm, and Heinz Himmelbauer. 2011. "Evaluation of Genomic High-Throughput Sequencing Data Generated on Illumina HiSeq and Genome Analyzer Systems." *Genome Biology* 12 (11): R112.
Morris, Robert M., Michael S. Rappé, Stephanie A. Connon, Kevin L. Vergin, William A. Siebold, Craig A. Carlson, and Stephen J. Giovannoni. 2002. "SAR11 Clade Dominates Ocean Surface Bacterioplankton Communities." *Nature* 420 (6917): 806–10.
Naether, Daniela J., Slavtscho Slawtschew, Sebastian Stasik, Maria Engel, Martin Olzog, Lukas Y. Wick, Kenneth N. Timmis, and Hermann J. Heipieper. 2013. "Adaptation of the Hydrocarbonoclastic Bacterium Alcanivorax Borkumensis SK2 to Alkanes and Toxic Organic Compounds: A Physiological and Transcriptomic Approach." *Applied and Environmental Microbiology* 79 (14): 4282–93.
Nayfach, Stephen, Beltran Rodriguez-Mueller, Nandita Garud, and Katherine S. Pollard. 2016. "An Integrated Metagenomics Pipeline for Strain Profiling Reveals Novel Patterns of Bacterial Transmission and Biogeography." *Genome Research* 26 (11): 1612–25.
Nguyen, Eric, Michael Poli, Matthew G. Durrant, Armin W. Thomas, Brian Kang, Jeremy Sullivan, Madelena Y. Ng, et al. 2024. "Sequence Modeling and Design from Molecular to Genome Scale with Evo." https://doi.org/10.1101/2024.02.27.582234.
Olm, Matthew R., Alexander Crits-Christoph, Spencer Diamond, Adi Lavy, Paula B. Matheus Carnevali, and Jillian F. Banfield. 2020. "Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries." *mSystems* 5 (1). https://doi.org/10.1128/mSystems.00731-19.
Pachiadaki, Maria G., Julia M. Brown, Joseph Brown, Oliver Bezuidt, Paul M. Berube, Steven J. Biller, Nicole J. Poulton, et al. 2019. "Charting the Complexity of the Marine Microbiome

1030 through Single-Cell Genomics." *Cell* 179 (7): 1623–35.e11.
1031 Paoli, Lucas, Hans-Joachim Ruscheweyh, Clarissa C. Forneris, Florian Hubrich, Satria Kautsar,
1032 Agneya Bhushan, Alessandro Lotti, et al. 2022. "Biosynthetic Potential of the Global Ocean
1033 Microbiome." *Nature* 607 (7917): 111–18.
1034 Parks, Donovan H., Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J.
1035 Mussig, and Philip Hugenholtz. 2020. "A Complete Domain-to-Species Taxonomy for
1036 Bacteria and Archaea." *Nature Biotechnology* 38 (9): 1079–86.
1037 Parks, Donovan H., Maria Chuvochina, Christian Rinke, Aaron J. Mussig, Pierre-Alain
1038 Chaumeil, and Philip Hugenholtz. 2022. "GTDB: An Ongoing Census of Bacterial and
1039 Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete
1040 Genome-Based Taxonomy." *Nucleic Acids Research* 50 (D1): D785–94.
1041 Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J.
1042 Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. 2017. "Recovery of
1043 Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life."
1044 *Nature Microbiology* 2 (11): 1533–42.
1045 Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica
1046 Armanini, Francesco Beghini, et al. 2019. "Extensive Unexplored Human Microbiome
1047 Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age,
1048 Geography, and Lifestyle." *Cell* 176 (3): 649–62.e20.
1049 Pedler, Byron E., Lihini I. Aluwihare, and Farooq Azam. 2014. "Single Bacterial Strain Capable
1050 of Significant Contribution to Carbon Cycling in the Surface Ocean." *Proceedings of the*
1051 *National Academy of Sciences of the United States of America* 111 (20): 7202–7.
1052 Prasad, Manoj, Nozomu Obana, Kaori Sakai, Toshiki Nagakubo, Shun Miyazaki, Masanori
1053 Toyofuku, Jacques Fattaccioli, Nobuhiko Nomura, and Andrew S. Utada. 2019. "Point
1054 Mutations Lead to Increased Levels of c-Di-GMP and Phenotypic Changes to the Colony
1055 Biofilm Morphology in Alcanivorax Borkumensis SK2." *Microbes and Environments* 34 (1):
1056 104–7.
1057 Prasad, M., N. Obana, S-Z Lin, S. Zhao, K. Sakai, C. Blanch-Mercader, J. Prost, et al. 2023.
1058 "Alcanivorax Borkumensis Biofilms Enhance Oil Degradation by Interfacial Tubulation."
1059 *Science (New York, N.Y.)* 381 (6659): 748–53.
1060 Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. "FastTree 2--Approximately
1061 Maximum-Likelihood Trees for Large Alignments." *PloS One* 5 (3): e9490.
1062 Rusch, Douglas B., Adam C. Martiny, Christopher L. Dupont, Aaron L. Halpern, and J. Craig
1063 Venter. 2010. "Characterization of Prochlorococcus Clades from Iron-Depleted Oceanic
1064 Regions." *Proceedings of the National Academy of Sciences of the United States of*
1065 *America* 107 (37): 16184–89.
1066 Ruscheweyh, Hans-Joachim, Alessio Milanese, Lucas Paoli, Nicolai Karcher, Quentin Clayssen,
1067 Marisa Isabell Keller, Jakob Wirbel, et al. 2022. "Cultivation-Independent Genomes Greatly
1068 Expand Taxonomic-Profiling Capabilities of mOTUs across Various Environments."
1069 *Microbiome* 10 (1): 212.
1070 Sabirova, Julia S., Tatyana N. Chernikova, Kenneth N. Timmis, and Peter N. Golyshin. 2008.
1071 "Niche-Specificity Factors of a Marine Oil-Degrading Bacterium Alcanivorax Borkumensis
1072 SK2." *FEMS Microbiology Letters* 285 (1): 89–96.
1073 Salazar, Guillem, Lucas Paoli, Adriana Alberti, Jaime Huerta-Cepas, Hans-Joachim
1074 Ruscheweyh, Miguelangel Cuenca, Christopher M. Field, et al. 2019. "Gene Expression
1075 Changes and Community Turnover Differentially Shape the Global Ocean
1076 Metatranscriptome." *Cell* 179 (5): 1068–83.e21.
1077 Sánchez, Pablo, Felipe H. Coutinho, Marta Sebastián, Massimo C. Pernice, Raquel Rodríguez-
1078 Martínez, Guillem Salazar, Francisco Miguel Cornejo-Castillo, et al. 2024. "Marine
1079 Picoplankton Metagenomes and MAGs from Eleven Vertical Profiles Obtained by the
1080 Malaspina Expedition." *Scientific Data* 11 (1): 154.

Sánchez, Pablo, Marta Sebastián, Massimo Pernice, Raquel Rodríguez-Martínez, Stephane Pesant, Susana Agustí, Takashi Gojobori, et al. 2023. “Marine Picoplankton Metagenomes from Eleven Vertical Profiles Obtained by the Malaspina Expedition in the Tropical and Subtropical Oceans.” *bioRxiv*. https://doi.org/10.1101/2023.02.06.526790.

Schattenhofer, Martha, Bernhard M. Fuchs, Rudolf Amann, Mikhail V. Zubkov, Glen A. Tarran, and Jakob Pernthaler. 2009. “Latitudinal Distribution of Prokaryotic Picoplankton Populations in the Atlantic Ocean.” *Environmental Microbiology* 11 (8): 2078–93.

Schmidt, Thomas S. B., Anthony Fullam, Pamela Ferretti, Askarbek Orakov, Oleksandr M. Maistrenko, Hans-Joachim Ruscheweyh, Ivica Letunic, et al. 2024. “SPIRE: A Searchable, Planetary-Scale mIcrobiome REsource.” *Nucleic Acids Research* 52 (D1): D777–83.

Schroer, William Francis. 2023. “Metabolite Transport and Its Role in Marine Microbial Interactions.” University of Georgia. https://www.proquest.com/docview/2917419506.

Septer, Alecia N., and Karen L. Visick. 2024. “Lighting the Way: How the Vibrio Fischeri Model Microbe Reveals the Complexity of Earth’s ‘Simplest’ Life Forms.” *Journal of Bacteriology* 206 (5): e0003524.

Shaiber, Alon, Amy D. Willis, Tom O. Delmont, Simon Roux, Lin-Xing Chen, Abigail C. Schmid, Mahmoud Yousef, et al. 2020. “Functional and Genetic Markers of Niche Partitioning among Enigmatic Members of the Human Oral Microbiome.” *Genome Biology* 21 (1): 292.

Spang, Anja, Jimmy H. Saw, Steffen L. Jørgensen, Katarzyna Zaremba-Niedzwiedzka, Joran Martijn, Anders E. Lind, Roel van Eijk, Christa Schleper, Lionel Guy, and Thijs J. G. Ettema. 2015. “Complex Archaea That Bridge the Gap between Prokaryotes and Eukaryotes.” *Nature* 521 (7551): 173–79.

Steinegger, Martin, and Johannes Söding. 2017. “MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets.” *Nature Biotechnology*, October. https://doi.org/10.1038/nbt.3988.

Stepanauskas, Ramunas, Elizabeth A. Fergusson, Joseph Brown, Nicole J. Poulton, Ben Tupper, Jessica M. Labonté, Eric D. Becraft, et al. 2017. “Improved Genome Recovery and Integrated Cell-Size Analyses of Individual Uncultured Microbial Cells and Viral Particles.” *Nature Communications* 8 (1): 84.

Sul, Woo Jun, Thomas A. Oliver, Hugh W. Ducklow, Linda A. Amaral-Zettler, and Mitchell L. Sogin. 2013. “Marine Bacteria Exhibit a Bipolar Distribution.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (6): 2342–47.

Sunagawa, Shinichi, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, et al. 2015. “Ocean Plankton. Structure and Function of the Global Ocean Microbiome.” *Science* 348 (6237): 1261359.

Thomas, Mridul K., Colin T. Kremer, Christopher A. Klausmeier, and Elena Litchman. 2012. “A Global Pattern of Thermal Adaptation in Marine Phytoplankton.” *Science* 338 (6110): 1085–88.

Tully, Benjamin J., Elaina D. Graham, and John F. Heidelberg. 2018. “The Reconstruction of 2,631 Draft Metagenome-Assembled Genomes from the Global Oceans.” *Scientific Data* 5 (January):170203.

Ustick, Lucas J., Alyse A. Larkin, and Adam C. Martiny. 2023. “Global Scale Phylogeography of Functional Traits and Microdiversity in Prochlorococcus.” *The ISME Journal* 17 (10): 1671–79.

Utter, Daniel R., Gary G. Borisy, A. Murat Eren, Colleen M. Cavanaugh, and Jessica L. Mark Welch. 2020. “Metapangenomics of the Oral Microbiome Provides Insights into Habitat Adaptation and Cultivar Diversity.” *Genome Biology* 21 (1): 293.

Watterson, William J., Melikhan Tanyeri, Andrea R. Watson, Candace M. Cham, Yue Shan, Eugene B. Chang, A. Murat Eren, and Savaş Tay. 2020. “Droplet-Based High-Throughput Cultivation for Accurate Screening of Antibiotic Resistant Gut Microbes.” *eLife* 9 (June). https://doi.org/10.7554/eLife.56998.

Weinheimer, Alaina R., and Frank O. Aylward. 2020. “A Distinct Lineage of Caudovirales That Encodes a Deeply Branching Multi-Subunit RNA Polymerase.” *Nature Communications* 11 (1): 4506.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the Tidyverse.” *Journal of Open Source Software* 4 (43): 1686.

Woodcroft, Ben J., Samuel T. N. Aroney, Rossen Zhao, Mitchell Cunningham, Joshua A. M. Mitchell, Linda Blackall, and Gene W. Tyson. 2024. “SingleM and Sandpiper: Robust Microbial Taxonomic Profiles from Metagenomic Data.” *bioRxiv*. https://doi.org/10.1101/2024.01.30.578060.

Wood, Derrick E., and Steven L. Salzberg. 2014. “Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments.” *Genome Biology* 15 (3): R46.

Woyke, Tanja, Devin F. R. Doud, and Frederik Schulz. 2017. “The Trajectory of Microbial Single-Cell Sequencing.” *Nature Methods* 14 (11): 1045–54.

Wu, Martin, and Jonathan A. Eisen. 2008. “A Simple, Fast, and Accurate Method of Phylogenomic Inference.” *Genome Biology* 9 (10): R151.