Quantitative insights into the efficacy of genome-resolved surveys of microbial communities through ribosomal protein phylogeography and EcoPhylo

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.28847705

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 apply this approach to ribosomal proteins to quantify phylogeny-aware genome recovery rates in two genome-resolved investigations of the human gut and oral cavity. Our results demonstrate that 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 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 (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

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

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

88 One approach to gaining insight into microbial life underrepresented in genome catalogues
89 involves the use of marker genes in metagenomic assemblies. *De novo* assembly, in which
90 individual sequencing reads are stitched together to recover much longer contiguous segments
91 of DNA (contigs), is common to the vast majority of genome recovery efforts. While in most cases
92 contigs only represent fragments of genomes, they still explain a much greater genomic context
93 than unassembled reads and give access to entire open reading frames, including
94 phylogenetically informative marker genes. Employing such phylogenetically informative genes
95 assembled from metagenomes enables both broad surveys of taxonomic diversity and fine-scale
96 phylogenetic analyses. For example, RNA and DNA polymerase genes have been used to resolve
97 viral clades and guide targeted genome recovery (Weinheimer and Aylward 2020; Gaïa et al.
98 2023). Additionally, marker genes have also been used to estimate the proportion of phylogenetic
99 diversity contributed by genomes and isolate genomes in metagenomic assemblies (D. Wu et al.
100 2025). Furthermore, when combined with metagenomic read recruitment, such marker genes also
101 reveal the biogeography of individual taxa, as demonstrated by studies leveraging the rpoC1 gene
102 to uncover phylogeographic patterns in marine bacteria (Kent et al. 2019; Ustick, Larkin, and
103 Martiny 2023).

104 Among all phylogenetically informative genes, ribosomal proteins represent a special class as
105 they (1) occur as a single-copy gene in genomes across the tree of life, (2) are consistently
106 assembled even for complex or relatively rare populations in metagenomes due to their relatively
107 short length, and (3) contain enough phylogenetic information to delineate distinct branches of life
108 at relatively high levels of resolution (Olm et al. 2020). Recognizing their utility, many studies have
109 leveraged individual ribosomal proteins to analyze community composition (M. Wu and Eisen
110 2008; Crits-Christoph et al. 2022), integrating ribosomal protein phylogenies with metagenomic
111 read recruitment to track individual clades of microbes (Hug et al. 2013; Emerson et al. 2016;
112 Hamilton et al. 2016; Diamond et al. 2019; Matheus Carnevali et al. 2021). Ribosomal proteins

113 are thus ideally suited gene markers for tracking microbial populations underrepresented within
114 genome collections.

115 Here we present EcoPhylo, a workflow to simultaneously visualize the phylogenetic relationships
116 and biogeographical distribution patterns of sequences that match any given gene family from
117 genomes and metagenomes, and demonstrate its application to the phylogeography of ribosomal
118 proteins for quantification of genome recovery rates across biomes. Our results show that bringing
119 together multi-domain ribosomal protein phylogenies with distribution patterns of individual clades
120 across environments in a single interface offers a valuable data analysis and visualization strategy
121 to benchmark genome recovery efforts scaling from individual projects to global surveys of large
122 genome collections and metagenomes.

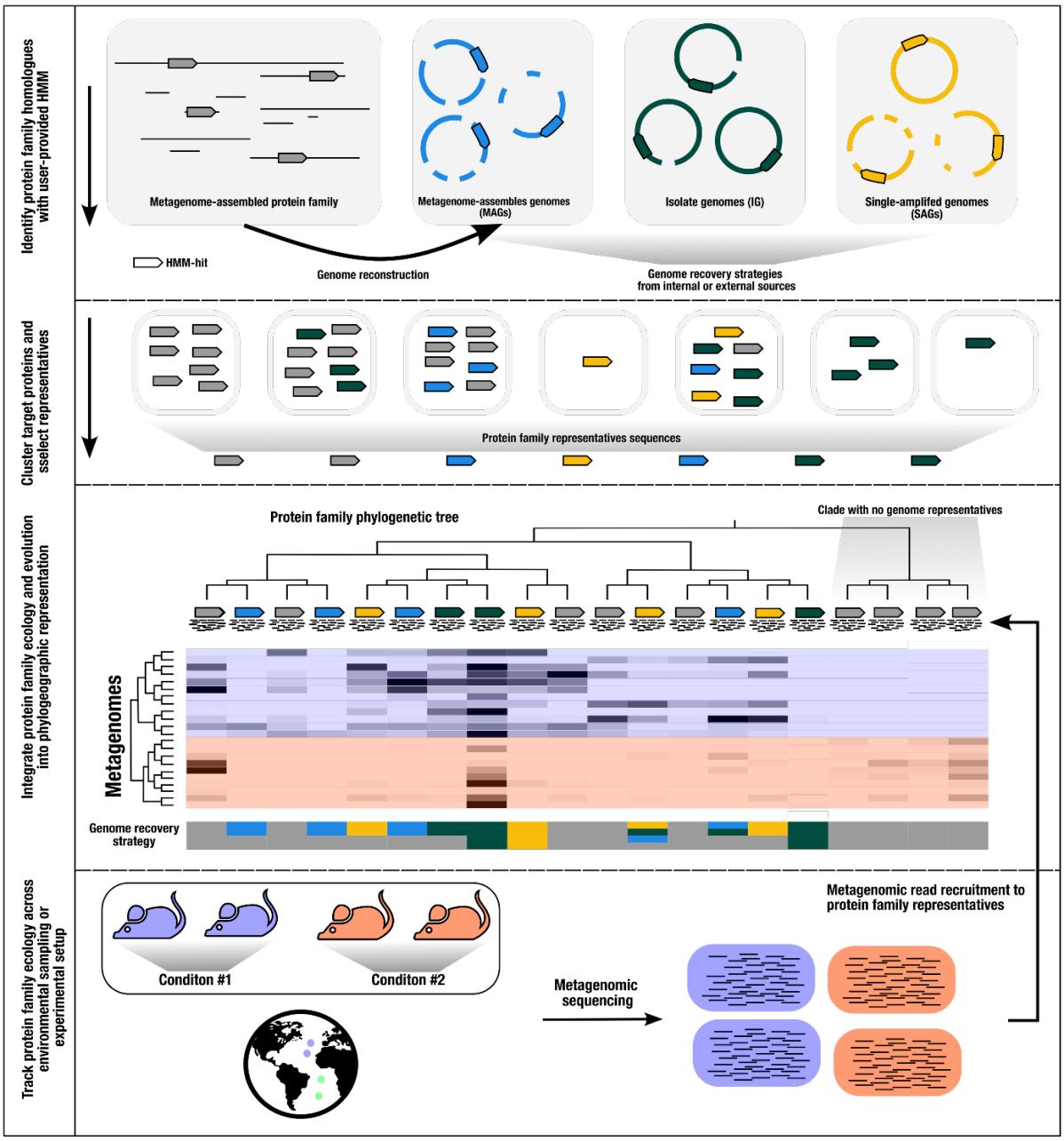
123 Results

124 EcoPhylo enables integrated surveys of gene family 125 phylogeography

126 EcoPhylo implements a computational workflow to integrate the phylogeny and biogeography of
127 any given gene family and enables its users to track the distribution patterns and evolutionary
128 relationships between homologous genes across environments and/or experimental conditions
129 (Figure 1, also see Materials and Methods).

130 When applied to phylogenetically tractable single-copy core genes, such as ribosomal proteins,
131 in tandem with metagenomes and a genome collection, EcoPhylo identifies populations
132 assembled in metagenomes but absent in the genomic collections (and *vice versa*), highlighting
133 the ecological and evolutionary relevance of organisms detected through metagenomic
134 assemblies but lacking genomic representation (Figure 1). This allows for the quantification of
135 genome recovery rates of different methods (e.g., isolate genomes, MAGs, SAGs) across taxa

136 and provides a means to investigate phylogenetic and ecological features of organisms without
137 genomic representation. Importantly, the unbroken link between genes and contigs enables
138 downstream targeted binning efforts when necessary.



139

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

143 which genome recovery strategies are more effective for sampling specific taxa. Although this manuscript focuses on
144 ribosomal proteins, the proposed workflow is generalizable to any gene family.

145 Using ribosomal proteins to *de novo* characterize the phylogenetic makeup of microbiomes and
146 benchmark genome recovery rates has numerous advantages. However, these advantages also
147 pose noteworthy challenges. Ribosomal proteins are short protein sequences (~300 amino acids),
148 which substantially limits their ability to resolve deep phylogenetic branching patterns.
149 Furthermore, their evolution is subject to strong purifying selection, as a result, the average
150 nucleotide identity (ANI) threshold often used to define 'species' boundaries between whole
151 genomes is 95% (Jain et al. 2018) increases to 99% for ribosomal protein sequences (Olm et al.
152 2020). Therefore, ribosomal proteins are more vulnerable than other genes to non-specific read-
153 recruitment from closely related proteins within metagenomes. To identify criteria for reliably
154 resolving taxa, we started our investigation by developing a series of benchmarks to optimize the
155 use of ribosomal proteins in EcoPhylo with appropriate parameters to maximize the ecological
156 and evolutionary signal they can offer while minimizing non-specific read recruitment. These
157 benchmarks, which are detailed in the Supplementary information (1) inspected hidden Markov
158 model (HMM) alignment coverage thresholds to accurately detect ribosomal proteins in genomes
159 and metagenomes; (2) examined the copy number distribution of ribosomal protein HMMs across
160 archaeal and bacterial genomes to only consider single-copy candidates; and (3) explored
161 nucleotide similarity thresholds to cluster ribosomal gene sequences to maximize the taxonomic
162 resolution of representative sequences while maintaining sufficient nucleotide distance between
163 distinct representative sequences to reduce non-specific read recruitment from metagenomes
164 (Supplementary information).

165 Based on these considerations, we implemented routines and adjusted default EcoPhylo
166 parameters to (1) use a minimum of 80% model coverage for ribosomal protein HMMs for a match;
167 (2) filter for complete open reading frame sequences to remove assembly artifacts; and (3) cluster
168 HMM hits with target coverage to ensure grouping of extended open reading frames and leverage

169 97% nucleotide similarity as the most appropriate clustering threshold to minimize non-specific
170 read recruitment (Supplementary information). We also compared broad ecological insights
171 recovered from EcoPhylo to state-of-the-art taxonomic profiling tools, confirming that this
172 framework offered qualitatively comparable results (Supplementary information). Altogether,
173 these evaluation and optimization steps yielded EcoPhylo default parameters to obtain
174 representative ribosomal protein sequences that are suitable for investigations of the phylogeny,
175 biogeography, and genome recovery of populations they describe. To demonstrate the utility of
176 the EcoPhylo workflow, we performed two case studies exploring microbial ecology and genome
177 recovery across individual studies from the human oral cavity (Shaiber et al. 2020) and gut (Carter
178 et al. 2023).

179 Ribosomal proteins quantify and contextualize genome recovery 180 rates from metagenomes

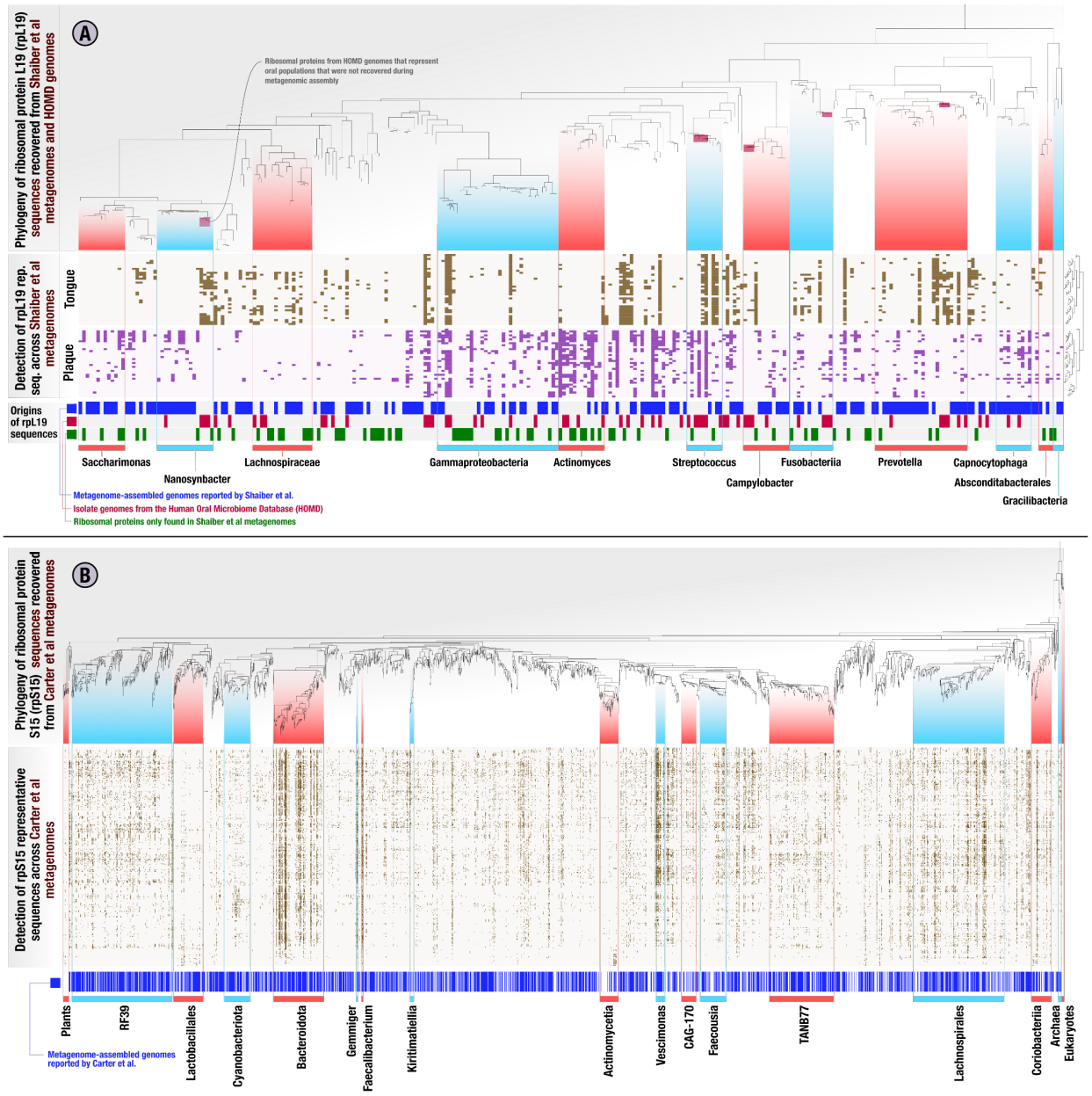
181 Thanks to its diverse physiological properties that promote a variety of chemical gradients and
182 surfaces (Bowen et al. 2018), the human oral cavity is home to diverse communities of microbes
183 (Dewhirst et al. 2010). The human oral microbiome is a relatively well-characterized environment
184 with a wealth of isolate genomes accessible through the Human Oral Microbiome Database
185 (HOMD) (Escapa et al. 2018; T. Chen et al. 2010), and numerous genome-resolved
186 metagenomics surveys that have captured representative genomes of microbial clades that have
187 largely eluded cultivation efforts. Using EcoPhylo we first focused on a genome-resolved
188 metagenomics survey which reconstructed multiple high-quality MAGs from tongue and plaque
189 samples from the human oral cavity (Shaiber et al. 2020). While Shaiber et al. (2020) reported
190 numerous genomes for elusive taxa, such as *Saccharimonadia* (TM7), *Absconditabacteria* (SR1),
191 and *Gracilibacteria* (GN02), the genome-resolved metagenomic workflow failed to reconstruct
192 MAGs that resolved to some of the best-represented organisms in culture collections from the

193 oral cavity, such as members of the genus *Streptococcus*, (Escapa et al. 2018), which was
194 represented by only two MAGs in Shaiber et al. (2020). This discrepancy compelled us to combine
195 isolate genomes from the HOMD together with metagenomes and MAGs from Shaiber et al.
196 (2020), to investigate whether EcoPhylo could reveal the differential recovery of genomes through
197 distinct recovery approaches.

198 We started our analysis by combining 790 non-redundant MAGs and 14 metagenomic co-
199 assemblies of tongue and plaque metagenomes reported by Shaiber et al (2020) with 8,615
200 isolate genomes we obtained from the HOMD (Supplementary Table 1). To characterize these
201 data, we elected to use EcoPhylo with *rpL19* HMM, since it was the most frequent ribosomal
202 protein with an average length of 393 nucleotides across all genomes in our collection, occurring
203 in 98.59% of the HOMD genomes and 81.81% of the Shaiber et al. MAGs (Supplementary
204 information). To assess the generalizability of observations made from *rpL19*, we also ran
205 EcoPhylo on the same dataset with *rpS15* and *rpS2*, with the average length of 275 and 781
206 nucleotides, respectively (Supplementary Table 2, Supplementary information).

207 The EcoPhylo analysis of the *rpL19* genes found in the genomes and metagenomic assemblies
208 resulted in a phylogenetic tree with 277 non-redundant bacterial representative sequences
209 (Figure 2A, Supplementary Table 3). Hierarchical clustering of metagenomes based on the
210 detection patterns of these *rpL19* sequences organized metagenomes into tongue and plaque
211 sampling sites *de novo* (Figure 2A, Supplementary Figure 1), demonstrating that a single
212 ribosomal gene family is able to capture the known ecological differences between these habitats.
213 Many closely related *rpL19* genes that resolved to prevalent oral taxa, such as *Prevotella* and
214 *Steptococcus*, showed within-genus differences in site specificity, a previously observed
215 phenomenon (Eren et al. 2014) that is attributed to divergent accessory genomes (Mark Welch,
216 Dewhirst, and Borisy 2019; Utter et al. 2020). Multiple ribosomal protein representative sequences
217 recruited reads from tongue as well as plaque metagenomes, also matching prior observations of

218 cosmopolitan taxa (Figure 2A, Supplementary information). Overall, the ecological insights
219 revealed by *rpL19* recapitulated known ecology of oral microbes (Mark Welch, Dewhirst, and
220 Borisy 2019) and provided a framework to assess genome recovery rates.



221

222 **Figure 2: Ribosomal protein phylogeny and detection patterns across metagenomes from the human oral**
223 **cavity and gut microbiomes.** In the heatmaps in both panels, each column represents a ribosomal protein
224 representative sequence, each row represents a metagenome, and each data point indicates whether a given ribosomal
225 protein was detected in a given metagenome. The columns of heatmaps are ordered by a tree which represents a

226 phylogenetic analysis of all ribosomal protein representative sequences, and the rows are ordered by a hierarchical
227 clustering dendrogram that is calculated based on the ribosomal protein detection patterns across metagenomes. The
228 panel **(A)** represents the EcoPhylo analysis of *rpL19* sequences across Shaiber et al. (2020) metagenome-assembled
229 genomes (MAGs), Shaiber et al. (2020) oral metagenomes, and HOMD genomes, and includes three additional rows
230 that indicate the origin of a given ribosomal protein, whether it is a metagenome-assembled genome (MAG, blue),
231 HOMD isolate genome (red), or only recovered from metagenomic assemblies with no representation in genomes
232 (green). Smaller red boxes in the phylogenetic tree mark microbial clades that were absent in the collection of MAGs
233 and assemblies reported by Shaiber et al. (2020), but detected in Shaiber et al. (2020) metagenomes solely due to the
234 inclusion of HOMD isolate genomes. The panel **(B)** represents the EcoPhylo analysis of *rpS15* sequences across the
235 Carter et al. (2023) metagenome-assembled genomes (MAGs) and Carter et al. (2023) gut metagenomes from a Hadza
236 tribe, and includes an additional row that indicates whether a MAG was reported for a given ribosomal protein (blue).

237 EcoPhylo tracks the origins of each sequence in each sequence cluster. Some *rpL19* clusters,
238 representatives of which are shown in the phylogenetic tree in Figure 2A, were composed of
239 sequences found only in metagenomic assemblies and not in MAGs or isolate genomes,
240 highlighting clades present in the environment but not in genome collections. Other *rpL19* clusters
241 only contained sequences represented in HOMD isolate genomes; despite their consistent
242 detection in oral samples through metagenomic read recruitment, they were absent in
243 metagenomic assemblies or MAGs, highlighting clades that are less accessible to short-read
244 metagenomic assembly approaches (Figure 2). To calculate genome recovery rates for any given
245 taxon, we divided the number of sequence clusters that contained a sequence from a given
246 genome recovery method by the total number of representative sequences EcoPhylo reported for
247 that taxon (Materials and Methods). This analysis revealed that 60.3% of the bacterial populations
248 defined by *rpL19* gene clusters that were detected in metagenomic reads also appeared in MAGs.
249 In other words, the overall bacterial MAG recovery rate in the study by Shaiber et al. (2020) was
250 60.3% (*rpS15*: 62.8%, *rpS2*: 53.2%) (Figure 2A, Supplementary Table 3, Supplementary Table
251 4). However, this rate of recovery was not uniform across individual taxa. EcoPhylo revealed
252 higher MAG recovery rates for taxa such as *Saccharimonas* at 69.2% (*rpS15*: N/A, *rpS2:* 63.6%),

253 and *Prevotella* at 76.9% (*rpS15*: 82.6%, *rpS2*: 84%). In contrast, the MAG recovery was lower for
254 populations in other clades, including *Gammaproteobacteria* and *Fusobacteriia*, with MAG
255 recovery rates of 47.1% (*rpS15*: 58.1%, *rpS2*: 44.8%) and 41.7% (*rpS15*: 46.2%, *rpS2*: 31.2%),
256 respectively (Figure 2A, Supplementary Table 3, Supplementary Table 4). The MAG recovery
257 rate was particularly low for *Streptococcus* at 30% (*rpS15*: 15.4%, *rpS2*: 10%), consistent with
258 the presence of only two MAGs in Shaiber et al. (2020). However, the MAG recovery rate for
259 *Actinomyces* was also very low at 23.1% (*rpS15*: 36.4%, *rpS2*: 13.3%) despite the
260 characterization of nine *Actinomyces* MAGs by Shaiber et al. (2020) reveals a large number of
261 distinct *Actinomyces* populations missed by MAGs even though they were present in the
262 assemblies (Figure 2A, Supplementary Table 3, Supplementary Table 4). Overall, this analysis
263 not only confirmed that MAG recovery rates are not uniform across microbial clades, but also
264 showed that quantification of these rates is possible and may yield unexpected insights into the
265 extent of diversity that is not represented in the final set of MAGs for some clades.

266 The inclusion of genomes from the HOMD increased the number of *rpL19* sequence clusters that
267 contained genomes in this dataset, i.e., the total genome recovery rate, from 60.3% to 73.3%
268 (*rpS15*: 74.8%, *rpS2*: 81.3%), and led to the representation of 35 additional microbial clades for
269 which the metagenomic sequencing and analysis workflow implemented in Shaiber et al. (2020)
270 did not assemble. As with MAGs, the improved detection of taxa among HOMD genomes was
271 not uniform across clades (Figure 2A). For example, HOMD genomes offered genomic context
272 for five additional *Streptococcus* populations, increasing the genome recovery rate from 30% with
273 MAGs only, up to 80% when including the HOMD collection. When taking into account both MAGs
274 and isolate genomes, the overall genome recovery rate of Shaiber et al. (2020) from a human
275 oral microbiome dataset determined by EcoPhylo was 73.3%, showing that ribosomal protein
276 phylogeography is an effective means to quantify genome recovery statistics for individual
277 studies. Conversely, EcoPhylo results showed that 26.7% of the individual clades that could be

278 detected through the presence of *rpL19* sequences in assemblies of Shaiber et al. (2020)
279 metagenomes lacked genomic representation in both Shaiber et al. (2020) MAGs and HOMD
280 isolates (Figure 2A). Clades that were solely detected through their assembled yet not binned
281 ribosomal proteins increased the detection of populations of *Lachnospiraceae*, *Actinomyces*,
282 *Gammaproteobacteria*, and *Patescibacteria* (Figure 2A). As EcoPhylo clusters ribosomal proteins
283 at 97% nucleotide similarity, a conservative threshold that underestimates biodiversity by often
284 grouping genomes with gANI below 95% (Olm et al. 2020), unbinned populations have the
285 potential to contain novel genomic diversity.

286 Next, we applied EcoPhylo to another genome-resolved metagenomics study that recently
287 characterized the gut microbiome of a Hadza hunter-gatherer tribe with a deep sequencing effort
288 by Carter et al. (2023), in which the authors reported nearly 50,000 redundant bacterial and
289 archaeal MAGs from 338 metagenomes with an average of 76 million paired-end reads
290 (Supplementary Table 1). EcoPhylo analysis of this dataset with *rpS15* with an average length of
291 276 nucleotides, along with *rpS16* and *rpL19*, with the average length of 297 and 370 nucleotides
292 respectively (Supplementary Figure 2), revealed a relatively high bacterial MAG recovery rate of
293 67.7% (*rpS16*: 72.8%, *rpL19*: 69.5%) (Figure 2B, Supplementary Table 2). While there were some
294 clades, such as *Actinomycetia*, for which the genome recovery rate was as low as 31.9% (*rpS16*:
295 32.4%, *rpL19*: 33.3%), the high MAG recovery rate was generally uniform across all major taxa,
296 highlighting how deep metagenomic sequencing yield relatively higher genome recovery rates
297 (Supplementary Table 5, Supplementary Table 6, Supplementary Information).

298 Through these analyses, we are able to demonstrate that the MAGs obtained by Carter et al.
299 (2023) more comprehensively represents the populations captured by their metagenomic
300 assemblies of the human gut compared to the MAGs obtained by Shaiber et al. (2020) given their
301 metagenomic assemblies of the oral cavity (Figure 2B, Supplementary Table 5, Supplementary
302 Table 6, Supplementary Information). The ability to make such a statement highlights the utility of

303 EcoPhylo at providing quantitative insights into the efficacy of genome-resolved surveys
304 independent of biomes while offering a phylogenetic and biogeographical context for the
305 populations that were detected in the assemblies.

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

310 Discussion

311 Our work illuminates the efficiencies of current genome recovery methods and their ability to
312 sample genomes from various microbiomes. By leveraging phylogenetically informative marker
313 genes detected in metagenomic assemblies, such as ribosomal proteins, that are absent from
314 final genome collections, EcoPhylo provides a robust framework for benchmarking genome
315 recovery rates across multiple genome acquisition methods and contextualizing the ecological
316 and evolutionary of genome collections with naturally occurring microbial populations. Our study
317 examined two microbiome projects that used multiple genome recovery strategies (MAGs and
318 isolate genomes) to survey the human oral cavity and gut. Overall, we found that the EcoPhylo
319 workflow can quantitatively measure genome recovery rates and analyze heterogeneous genome
320 collections to assess the efficacy of distinct recovery methods at the level of individual
321 phylogenetic clades. In the oral microbiome, we found clade-specific biases of genome recovery
322 rates across MAGs and isolate genomes. For example, isolate genomes effectively sampled
323 *Streptococcus* while MAGs efficiently sampled *Prevotella*. However, in the human gut, we
324 observed that deep metagenomic sequencing and subsequent generation of MAGs yielded higher
325 genome recovery rates. Previous studies have shown that human gut samples exhibit relatively

326 little genomic novelty (D. Wu et al. 2025), further supporting the effectiveness of genome-resolved
327 metagenomics for acquiring genomes in this biome. Across both case studies, EcoPhylo
328 generated insights into multi-domain ribosomal protein phylogeography and provided a valuable
329 interactive data visualization strategy to evaluate the underlying microbial ecology of
330 metagenomic sequencing projects.

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

346 While the phylogeography of ribosomal proteins offers valuable insights into genome recovery,
347 these genes have notable limitations. First, ribosomal proteins assembled from metagenomes
348 only capture a fraction of the microbial diversity in a sample, restricting genome recovery metrics
349 to taxa detectable by metagenomics. Next, rates of evolution as well as the likelihood to be
350 recovered through metagenomic assembly will differ across ribosomal protein families,

351 complicating direct quantitative comparisons between different ribosomal proteins and in some
352 cases will require surveying multiple ribosomal proteins to ensure the generalizability of
353 observations from a single ribosomal protein. Additionally, individual ribosomal protein trees will
354 have less phylogenetic power compared to concatenated ribosomal protein trees or longer marker
355 genes. Although this may lead to suboptimal organism phylogenetics, the efficient organization of
356 ribosomal proteins yields informative insights into the diversity of clades within a sample.
357 Furthermore, when working with incomplete genomes, such as MAGs or SAGs, a single
358 ribosomal gene family will rarely be detected across the entire genome collection and thus only a
359 subset of genomes will be contextualized per protein. Yet the inherent trade-offs of using
360 incomplete genomes (x ≥ 50% and less than 10% contamination) highlight ongoing challenges in
361 genome recovery, as stricter completeness thresholds would further reduce the number of
362 genomes available for analysis.

363 The modular design and customizable parameters of EcoPhylo allows users to go beyond
364 ribosomal proteins and leverage other gene families tailored for specific analyses which can
365 improve phylogenetics and the detection of specific taxa. For example, RNApolA and RNApolB
366 have been leveraged for phylogeny-guided binning leading to the discovery of missing branches
367 in viral evolution (Gaïa et al. 2023). Furthermore, phylogeography of functional protein families
368 can be leveraged as proxies for microbial metabolism, e.g. phylogeography of ABC transporters
369 can aid in modeling cryptic fluxes of microbial metabolites (Schroer 2023). The EcoPhylo workflow
370 provides a platform for future microbiome projects to benchmark their genome recovery rates
371 upon release of genome collections. Ribosomal protein phylogeography in tandem with reporting
372 read recruitment percentages to representative genome collections, provides comprehensive
373 insights into genome recovery rates given the biodiversity detected in metagenomes. The

374 identification of microbial diversity hotspots with novel genomes using multi-biome databases is
375 a growing trend in the field (Woodcroft et al. 2024; D. Wu et al. 2025). While these approaches
376 enable broad-scale insights into habitats and samples with low genome recovery, EcoPhylo
377 provides a direct link to the assembled genes within contigs of taxa with unrecovered genomes
378 and illuminates phylogeny-guided genome recovery on a project-level. This immediate traceability
379 to the underlying genomic fragments in metagenomic assemblies can catalyze targetted genome
380 recovery. Future studies can leverage the strategy implemented in EcoPhylo to reanalyze existing
381 metagenomic assemblies to identify missing clades or develop tailored methods to optimize
382 overall genome recovery efforts by taking advantage of the increasing availability of genomes and
383 metagenomes.

384 Materials and Methods

385 The EcoPhylo workflow

386 EcoPhylo is a computational workflow implemented in the open-source software ecosystem anvi'o
387 (Eren et al. 2015, 2021) using the Python programming language and the workflow management
388 system, Snakemake (Köster and Rahmann 2012). The primary purpose of EcoPhylo is to offer
389 an integrated means to study phylogenetic relationships and ecological distribution patterns of
390 sequences that match to any gene family based on user-provided hidden Markov model (HMM)
391 searches from genomic and metagenomic assemblies. A minimal command line instruction to
392 start an EcoPhylo run is `anvi-run-workflow -w ecophylo -c config.json`, where `anvi-run-workflow`
393 is a program in anvi'o that runs various workflows, and `config.json` is a JSON formatted
394 configuration file that describes file paths (such as the locations of genomes and/or
395 metagenomes) and other parameters (such as the HMM to be used for a homology search, and

396 sequence identity cutoffs). Comprehensive user documentation for EcoPhylo is available at
397 https://anvio.org/m/ecophylo.

398 The minimum input for the EcoPhylo is a gene family hidden Markov model (HMM) and a dataset
399 of genomic and/or metagenomic assemblies. EcoPhylo identifies and clusters target genes or
400 translated proteins across assemblies to yield a non-redundant, representative set of open
401 reading frames (ORFs). Next, an amino acid phylogenetic tree is calculated with the translated
402 representative ORFs yielding the evolutionary history captured by homologues from input
403 assemblies. An additional user input to the workflow is a metagenomic sequencing dataset
404 representing ecological sampling or an experimental setup. With this input, the workflow performs
405 metagenomic read recruitment against the representative ORFs to yield ecological insights into
406 the gene family. Finally, the separate data types are integrated into a phylogeographic
407 representation of the gene family (Figure 1).

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

415 An application of the EcoPhylo workflow with default settings will (1) identify gene families with
416 the program `hmmsearch` in (Eddy 2011) using the user-provided HMM model, (2) annotate
417 affiliate hmm-hits with taxonomic names with `anvi-run-scg-taxonomy` when applicable, (3)
418 remove hmm-hits with less than 80% HMM model alignment coverage and incomplete ORFs with
419 the anvi'o program `anvi-script-filter-hmm-hits` with parameters `--min-model-coverage 0.8` and

420 `--filter-out-partial-gene-calls` to minimize the inclusion of non-target sequences and spurious
421 HMM hits, (4) dereplicate the resulting DNA sequences at 97% gANI and pick cluster
422 representatives using MMseqs2 (Steinegger and Söding 2017), (5) use the translated
423 representative sequences to calculate a multiple sequence alignment (MSA) using (Edgar 2004)
424 with the `-maxiters 2` flag (Edgar 2004), trim the alignment by removing columns of the alignment
425 with trimAL with the '-gappyout' flag (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009), (6)
426 remove sequences that have more than 50% gaps using the anvi'o program `anvi-script-reformat-
427 fasta`, (7) calculate a phylogenetic tree using FastTree (Price, Dehal, and Arkin 2010) with the
428 flag `-fastest`, (8) perform metagenomic read recruitment analysis and profiling of non-translated
429 representative sequences using the anvi'o metagenomic workflow (Shaiber et al. 2020), which by
430 default relies upon Bowtie2 (Langmead and Salzberg 2012), (9) generate miscellaneous data to
431 annotate the representative sequences including taxonomy with `anvi-estimate-scg-taxonomy`
432 for ribosomal proteins, cluster size, and sequence length, and finally (10) generate anvi'o artifacts
433 that give integrated access to the phylogenetic tree of all representative sequences and read
434 recruitment results that can be visualized using the anvi'o interactive interface and/or further
435 processed for specific downstream analyses using any popular data analysis environment such
436 as R and/or Python.

437 The workflow that resulted in the recovery and characterization of ribosomal proteins in our
438 manuscript used the following additional steps: (1) we removed input reference genomes that
439 were not detected in at least one of the input metagenomes above a detection value of 0.9 with
440 their ribosomal protein (we kept all MAGs originating from the samples themselves) to only
441 visualize detected populations, (2) we manually curated the ribosomal protein tree when
442 necessary to remove sequences that appeared to be chimeric and those that formed spurious
443 long branches likely originating from metagenomic assembly artifacts, and/or mitochondrial or
444 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 families to limit non-specific read recruitment while maximizing taxonomic resolution. We applied the EcoPhylo workflow across the three CAMI biome synthetic genomic/metagenomic datasets (Marine, Plant-associated, and Strain-madness). As an initial step, we identified the top five most frequent ribosomal gene families that were detected in single-copy in the associated genomic collections for each synthetic metagenomic dataset. We then conducted a parameter grid search, spanning 95%-100% nucleotide similarity parameter grid search (Meyer et al. 2022). Next, we measured the amount of non-specific read recruitment in each EcoPhylo iteration, i.e. reads with equal mapping scores between their primary and secondary alignments (multi-mapped reads), with the following Samtools command: `samtools view $sample | grep XS:i | cut -f12-13 | sed 's/..:i://g' | awk '$1==$2' | wc -l`. The percentage of non-specific read recruitment was calculated by dividing the number of multi-mapped reads by the total number of reads mapped to the representative dataset. With this, we identified that nucleotide clustering thresholds greater than 97% began to show signs of non-specific read recruitment (Supplementary information).

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

479 Genome collections

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

487 For the Hadza tribe human gut microbiome analysis we followed the data download guidelines
488 shared by Carter et al. (2023) to obtain genomes from doi:10.5281/zenodo.7782708. Carter et al.
489 (2023) formed clusters at 95% gANI by including additional genomes outside of the MAGs they
490 have reconstructed from the Hazda gut metagenomes. To exclusively analyze microbial genomes

491 affiliated with the Hadza metagenomes, we filtered for cluster representatives with cluster
492 members that contained at least one Hadzda adult or infant MAG which produced 2,437
493 representative Bacterial and Archaeal MAGs.

494 Metagenome and metagenomic assembly datasets

495 To explore the phylogeography of ribosomal proteins, we used used 71 tooth and plaque
496 metagenomes from the human oral cavity which were downloaded from the NCBI BioProject
497 PRJNA625082 (Shaiber et al. 2020) along with associated co-assemblies
498 (doi:10.6084/m9.figshare.12217799). Next, to explore deep sequencing in the human gut
499 microbiome we used 388 metagenomes and assemblies from infant and adult members of the
500 Hadza tribe (doi:10.5281/zenodo.7782708) using the FTP links shared in from the file
501 `Supplemental_Table_S1.csv` and NCBI BioProject PRJEB49206 (Carter et al. 2023). All
502 metagenomes and associated assembly accessions can be found at Supplementary Table 1.

503 Preprocessing of genomic and metagenomic assemblies and 504 metagenomic short reads

505 Metagenomic and genomic assemblies were preprocessed with the anvi'o contigs workflow with
506 the program `anvi-run-workflow -w contigs` to predict open-reading frames with Prodigal (V2.6.3)
507 and identify SCGs for taxonomic inference with `anvi-run-scg-taxonomy` (Hyatt et al. 2010;
508 Shaiber et al. 2020). No contig size filters were implemented during this process to include
509 ribosomal proteins located on small contigs. To limit detection of misassemblies in downstream
510 analyses, only ribosomal proteins with complete open-reading frames (as predicted by Prodigal)
511 were analyzed with EcoPhylo (Hyatt et al. 2010). Additionally, metagenomic samples were quality
512 controlled with the anvi'o metagenomics workflow with the program `anvi-run-workflow -w

513 metagenomics` (Shaiber et al. 2020). This workflow uses the tool `iu-filter-quality-minoche` (Eren
514 et al. 2013), which implements methods described in (Minoche, Dohm, and Himmelbauer 2011).
515 All Snakemake workflows in this manuscript leveraged Snakemake v7.32.4 (Köster and Rahmann
516 2012).

517 Gene-level taxonomy of ribosomal proteins

518 To assign gene level taxonomy to ribosomal proteins, the EcoPhylo workflow relies upon the
519 anvi'o tools `anvi-run-scg-taxonomy` and `anvi-estimate-scg-taxonomy`, which leverage the
520 genomes and their taxonomy made available by the GTDB (Parks et al. 2022) to identify
521 taxonomic affiliations of genes that match to any of the ribosomal proteins L1, L13, L14, L16, L17,
522 L19, L2, L20, L21p, L22, L27A, L3, L4, L5, S11, S15, S16, S2, S6, S7, S8, or S9. During the
523 workflow, EcoPhylo uses `anvi-run-scg-taxonomy` to search for ribosomal genes annotated within
524 each anvi'o contigs database against the downloaded marker gene dataset with DIAMOND
525 v0.9.14 (Buchfink, Reuter, and Drost 2021). Later in the workflow, EcoPhylo runs `anvi-estimate-
526 scg-taxonomy --metagenome-mode` on the representative set of ribosomal proteins, which
527 assigns a consensus taxonomy to each sequence. The program `anvi-estimate-scg-taxonomy`
528 does not provide a taxonomic annotation if the ribosomal protein is less than 90% similar to any
529 of the ribosomal proteins found in GTDB genomes. In some cases, ribosomal proteins without
530 taxonomic annotation can be manually annotated with taxonomy based on the annotated
531 sequences that surround them in the phylogenetic tree, as we described in the section
532 "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

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

562 Detection of whole genomes in metagenomic data

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

572 Genome recovery rate estimations

573 Genome recovery rates were estimated to measure which individual or combination of genome
574 types (MAGs, SAGs, isolate genomes) most effectively sampled clades in the ribosomal protein
575 phylogenetic trees calculated during the EcoPhylo workflow. To calculate genome recovery rates
576 for any given taxon, we divided the number of sequence clusters that contained a sequence from

577 a given genome recovery method to the total number of representative sequences EcoPhylo
578 reported for that taxon. Taxonomic assignments of sequence cluster representatives were
579 determined with `anvi-estimate-scg-taxonomy`.

580 Taxonomic binning to improve genome recovery estimations

581 A subset of ribosomal proteins lacked taxon assignments from `anvi-estimate-scg-taxonomy` due
582 to their sequence similarity being x < 90% to GTDB genomes (See methods section: Gene-level
583 taxonomy of ribosomal proteins). Using the `anvi-interactive` interface, we examined the
584 placement of these proteins in the EcoPhylo ribosomal protein phylogenetic tree and manually
585 assigned taxon names based on the taxonomic affiliations of neighboring sequences.
586 Unannotated sequences were assigned taxonomy only when phylogenetic clustering
587 demonstrated clear consistency among neighboring sequences. These refined taxonomic
588 annotations were used to improve estimations of genome recovery in the main figures (*rpL19* and
589 *rpS15* in Figure 2).

590 Data and code availability

591 The URL https://merenlab.org/data/ecophylo-ribosomal-proteins/ serves all code and data
592 needed to reproduce our study. Additionally, all anvi'o artifacts that give interactive access to
593 EcoPhylo interfaces are publicly available at doi:10.6084/m9.figshare.28207481. Publicly
594 available genomes and metagenomes we used in our study are listed in the Supplementary
595 Tables. Supplementary Tables as well as Supplementary Information text are available at
596 doi:10.6084/m9.figshare.28847705.

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.

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.

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* 186 (14): 3111–24.e13.

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 Floyd E. Dewhirst. 2010. “The Human Oral Microbiome Database: A Web Accessible Resource for Investigating Oral Microbe Taxonomic and Genomic Information.” *Database: The Journal of Biological Databases and Curation* 2010 (0): baq013.

Cornman, Andre, Jacob West-Roberts, Antonio Pedro Camargo, Simon Roux, Martin Beracochea, Milot Mirdita, Sergey Ovchinnikov, and Yunha Hwang. 2024. “The OMG Dataset: An Open MetaGenomic Corpus for Mixed-Modality Genomic Language Modeling.” *bioRxiv*. https://doi.org/10.1101/2024.08.14.607850.

Crits-Christoph, Alexander, Spencer Diamond, Basem Al-Shayeb, Luis Valentin-Alvarado, and Jillian F. Banfield. 2022. “A Widely Distributed Genus of Soil Acidobacteria Genomically Enriched in Biosynthetic Gene Clusters.” *ISME Communications* 2 (1): 1–8.

Cross, Karissa L., James H. Campbell, Manasi Balachandran, Alisha G. Campbell, Connor J. Cooper, Ann Griffen, Matthew Heaton, et al. 2019. “Targeted Isolation and Cultivation of Uncultivated Bacteria by Reverse Genomics.” *Nature Biotechnology* 37 (11): 1314–21.

Delmont, Tom O., Christopher Quince, Alon Shaiber, Özcan C. Esen, Sonny Tm Lee, Michael S. Rappé, Sandra L. McLellan, Sebastian Lücker, and A. Murat Eren. 2018. “Nitrogen-

Fixing Populations of Planctomycetes and Proteobacteria Are Abundant in Surface Ocean Metagenomes." *Nature Microbiology* 3 (7): 804–13.

Dewhirst, Floyd E., Tuste Chen, Jacques Izard, Bruce J. Paster, Anne C. R. Tanner, Wen-Han Yu, Abirami Lakshmanan, and William G. Wade. 2010. "The Human Oral Microbiome." *Journal of Bacteriology* 192 (19): 5002–17.

Diamond, Spencer, Peter F. Andeer, Zhou Li, Alexander Crits-Christoph, David Burstein, Karthik Anantharaman, Katherine R. Lane, et al. 2019. "Mediterranean Grassland Soil C-N Compound Turnover Is Dependent on Rainfall and Depth, and Is Mediated by Genomically Divergent Microorganisms." *Nature Microbiology* 4 (8): 1356–67.

Dixon, Philip. 2003. "VEGAN, a Package of R Functions for Community Ecology." *Journal of Vegetation Science: Official Organ of the International Association for Vegetation Science* 14 (6): 927–30.

Durrant, Matthew G., Alison Fanton, Josh Tycko, Michaela Hinks, Sita S. Chandrasekaran, Nicholas T. Perry, Julia Schaepe, et al. 2023. "Systematic Discovery of Recombinases for Efficient Integration of Large DNA Sequences into the Human Genome." *Nature Biotechnology* 41 (4): 488–99.

Eddy, Sean R. 2011. "Accelerated Profile HMM Searches." *PLoS Computational Biology* 7 (10): e1002195.

Edgar, Robert C. 2004. "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput." *Nucleic Acids Research* 32 (5): 1792–97.

Emerson, Joanne B., Brian C. Thomas, Walter Alvarez, and Jillian F. Banfield. 2016. "Metagenomic Analysis of a High Carbon Dioxide Subsurface Microbial Community Populated by Chemolithoautotrophs and Bacteria and Archaea from Candidate Phyla." *Environmental Microbiology* 18 (6): 1686–1703.

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.

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.

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.

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.

Hug, Laura A., Cindy J. Castelle, Kelly C. Wrighton, Brian C. Thomas, Itai Sharon, Kyle R. Frischkorn, Kenneth H. Williams, Susannah G. Tringe, and Jillian F. Banfield. 2013. "Community Genomic Analyses Constrain the Distribution of Metabolic Traits across the Chloroflexi Phylum and Indicate Roles in Sediment Carbon Cycling." *Microbiome* 1 (1): 22.

Hwang, Yunha, Andre L. Cornman, Elizabeth H. Kellogg, Sergey Ovchinnikov, and Peter R. Girguis. 2024. "Genomic Language Model Predicts Protein Co-Regulation and Function." *Nature Communications* 15 (1): 2880.

Hyatt, Doug, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11 (March):119.

Imachi, Hiroyuki, Masaru K. Nobu, Nozomi Nakahara, Yuki Morono, Miyuki Ogawara, Yoshihiro Takaki, Yoshinori Takano, et al. 2020. "Isolation of an Archaeon at the Prokaryote–eukaryote Interface." *Nature*. https://doi.org/10.1038/s41586-019-1916-6.

Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018. "High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries." *Nature Communications* 9 (1): 5114.

Jiang, Cheng-Ying, Libing Dong, Jian-Kang Zhao, Xiaofang Hu, Chaohua Shen, Yuxin Qiao, Xinyue Zhang, et al. 2016. "High-Throughput Single-Cell Cultivation on Microfluidic Streak Plates." *Applied and Environmental Microbiology* 82 (7): 2210–18.

Kent, Alyssa G., Steven E. Baer, Céline Mouginot, Jeremy S. Huang, Alyse A. Larkin, Michael W. Lomas, and Adam C. Martiny. 2019. "Parallel Phylogeography of Prochlorococcus and Synechococcus." *The ISME Journal* 13 (2): 430–41.

Köster, Johannes, and Sven Rahmann. 2012. "Snakemake--a Scalable Bioinformatics Workflow Engine." *Bioinformatics (Oxford, England)* 28 (19): 2520–22.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.

Lu, Jennifer, Florian P. Breitwieser, Peter Thielen, and Steven L. Salzberg. 2017. "Bracken: Estimating Species Abundance in Metagenomics Data." *PeerJ. Computer Science* 3 (e104): e104.

Ma, Bin, Caiyu Lu, Yiling Wang, Jingwen Yu, Kankan Zhao, Ran Xue, Hao Ren, et al. 2023. "A Genomic Catalogue of Soil Microbiomes Boosts Mining of Biodiversity and Genetic Resources." *Nature Communications* 14 (1): 7318.

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.

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.

765 Meyer, Fernando, Adrian Fritz, Zhi-Luo Deng, David Koslicki, Till Robin Lesker, Alexey
766 Gurevich, Gary Robertson, et al. 2022. “Critical Assessment of Metagenome Interpretation:
767 The Second Round of Challenges.” *Nature Methods* 19 (4): 429–40.
768 Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D.
769 Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. “IQ-TREE 2: New Models and
770 Efficient Methods for Phylogenetic Inference in the Genomic Era.” *Molecular Biology and*
771 *Evolution* 37 (5): 1530–34.
772 Minoche, André E., Juliane C. Dohm, and Heinz Himmelbauer. 2011. “Evaluation of Genomic
773 High-Throughput Sequencing Data Generated on Illumina HiSeq and Genome Analyzer
774 Systems.” *Genome Biology* 12 (11): R112.
775 Nayfach, Stephen, Beltran Rodriguez-Mueller, Nandita Garud, and Katherine S. Pollard. 2016.
776 “An Integrated Metagenomics Pipeline for Strain Profiling Reveals Novel Patterns of
777 Bacterial Transmission and Biogeography.” *Genome Research* 26 (11): 1612–25.
778 Nguyen, Eric, Michael Poli, Matthew G. Durrant, Armin W. Thomas, Brian Kang, Jeremy
779 Sullivan, Madelena Y. Ng, et al. 2024. “Sequence Modeling and Design from Molecular to
780 Genome Scale with Evo.” https://doi.org/10.1101/2024.02.27.582234.
781 Olm, Matthew R., Alexander Crits-Christoph, Spencer Diamond, Adi Lavy, Paula B. Matheus
782 Carnevali, and Jillian F. Banfield. 2020. “Consistent Metagenome-Derived Metrics Verify
783 and Delineate Bacterial Species Boundaries.” *mSystems* 5 (1).
784 https://doi.org/10.1128/mSystems.00731-19.
785 Pachiadaki, Maria G., Julia M. Brown, Joseph Brown, Oliver Bezuidt, Paul M. Berube, Steven J.
786 Biller, Nicole J. Poulton, et al. 2019. “Charting the Complexity of the Marine Microbiome
787 through Single-Cell Genomics.” *Cell* 179 (7): 1623–35.e11.
788 Paoli, Lucas, Hans-Joachim Ruscheweyh, Clarissa C. Forneris, Florian Hubrich, Satria Kautsar,
789 Agneya Bhushan, Alessandro Lotti, et al. 2022. “Biosynthetic Potential of the Global Ocean
790 Microbiome.” *Nature* 607 (7917): 111–18.
791 Parks, Donovan H., Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J.
792 Mussig, and Philip Hugenholtz. 2020. “A Complete Domain-to-Species Taxonomy for
793 Bacteria and Archaea.” *Nature Biotechnology* 38 (9): 1079–86.
794 Parks, Donovan H., Maria Chuvochina, Christian Rinke, Aaron J. Mussig, Pierre-Alain
795 Chaumeil, and Philip Hugenholtz. 2022. “GTDB: An Ongoing Census of Bacterial and
796 Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete
797 Genome-Based Taxonomy.” *Nucleic Acids Research* 50 (D1): D785–94.
798 Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J.
799 Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. 2017. “Recovery of
800 Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life.”
801 *Nature Microbiology* 2 (11): 1533–42.
802 Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica
803 Armanini, Francesco Beghini, et al. 2019. “Extensive Unexplored Human Microbiome
804 Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age,
805 Geography, and Lifestyle.” *Cell* 176 (3): 649–62.e20.
806 Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. “FastTree 2--Approximately
807 Maximum-Likelihood Trees for Large Alignments.” *PloS One* 5 (3): e9490.
808 Ruscheweyh, Hans-Joachim, Alessio Milanese, Lucas Paoli, Nicolai Karcher, Quentin Clayssen,
809 Marisa Isabell Keller, Jakob Wirbel, et al. 2022. “Cultivation-Independent Genomes Greatly
810 Expand Taxonomic-Profiling Capabilities of mOTUs across Various Environments.”
811 *Microbiome* 10 (1): 212.
812 Schmidt, Thomas S. B., Anthony Fullam, Pamela Ferretti, Askarbek Orakov, Oleksandr M.
813 Maistrenko, Hans-Joachim Ruscheweyh, Ivica Letunic, et al. 2024. “SPIRE: A Searchable,
814 Planetary-Scale mIcrobiome REsource.” *Nucleic Acids Research* 52 (D1): D777–83.
815 Schroer, William Francis. 2023. “Metabolite Transport and Its Role in Marine Microbial

Interactions." University of Georgia. https://www.proquest.com/docview/2917419506.
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.
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, Dongying, Rekha Seshadri, Nikos C. Kyrpides, and Natalia N. Ivanova. 2025. "A Metagenomic Perspective on the Microbial Prokaryotic Genome Census." *Science Advances* 11 (3): eadq2166.
Wu, Martin, and Jonathan A. Eisen. 2008. "A Simple, Fast, and Accurate Method of Phylogenomic Inference." *Genome Biology* 9 (10): R151.