Unifying the global coding sequence space enables the study of genes with unknown function across biomes

Chiara Vanni[1,2], Matthew S. Schechter[1,3], Silvia G. Acinas[4], Albert Barberán[5], Pier Luigi Buttigieg[6], Emilio O. Casamayor[7], Tom O. Delmont[8], Carlos M. Duarte[9], A. Murat Eren[3,10], Robert D. Finn[11], Renzo Kottmann[1], Alex Mitchell[11], Pablo Sanchez[4], Kimmo Siren[12], Martin Steinegger[13,14], Frank Oliver Glöckner[15,16,2], Antonio Fernandez-Guerra[1,17 *]

Affiliations

1 Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359, Bremen, Germany

2 Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany

3 Department of Medicine, University of Chicago, Chicago, IL 60637, USA

4 Department of Marine Biology and Oceanography, Institut de Ciènces del Mar, CSIC, Barcelona, Spain.

5 Department of Environmental Science, University of Arizona, Tucson, 85721 AZ, USA

6 Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany

7 Center for Advanced Studies of Blanes CEAB-CSIC, Spanish Council for Research, Blanes, Spain

8 Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France

9 Red Sea Research Centre (RSRC) and Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia

10 Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA 02543, USA

11 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

12 Section for Evolutionary Genomics, The GLOBE Institute, University of Copenhagen, Copenhagen, Denmark

13 School of Biological Sciences, Seoul National University, Seoul, 08826, South Korea

14 Institute of Molecular Biology and Genetics, Seoul National University, Seoul, 08826, South Korea

15 University of Bremen, MARUM, Leobener Str. 8, 28359 Bremen, Germany
Life Sciences and Chemistry, Campus Ring 1, 28759 Bremen, Germany

16 Computing Center, Helmholtz Center for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany

17 Lundbeck GeoGenetics Centre, The Globe Institute, University of Copenhagen, 1350 Copenhagen, Denmark

*Corresponding author: Antonio Fernandez-Guerra, antonio.fernandez-guerra@sund.ku.dk

Abstract

One of the biggest challenges in molecular biology is bridging the gap between the known and the unknown coding sequence space. This challenge is especially extreme in microbial systems, where between 40% and 60% of the predicted genes are of unknown function. Discarding this uncharacterized fraction should not be an option anymore. Here, we present a conceptual framework and a computational workflow that bridges this gap and provides a powerful strategy to contextualize the investigations of genes of unknown function. Our approach partitions the coding sequence space removing the known-unknown dichotomy, unifies genomic and metagenomic data and provides a framework to expand those investigations across environments and organisms. By analyzing 415,971,742 genes predicted from 1,749 metagenomes and 28,941 bacterial and archaeal genomes we showcase our approach and its application in ecological, evolutionary and biotechnological investigations. As a result, we put into perspective the extent of the unknown fraction, its diversity, and its relevance in genomic and environmental contexts. By identifying a target gene of unknown function for antibiotic resistance, we demonstrate how a contextualized unknown coding sequence space enables the generation of hypotheses that can be used to augment experimental data.

Main

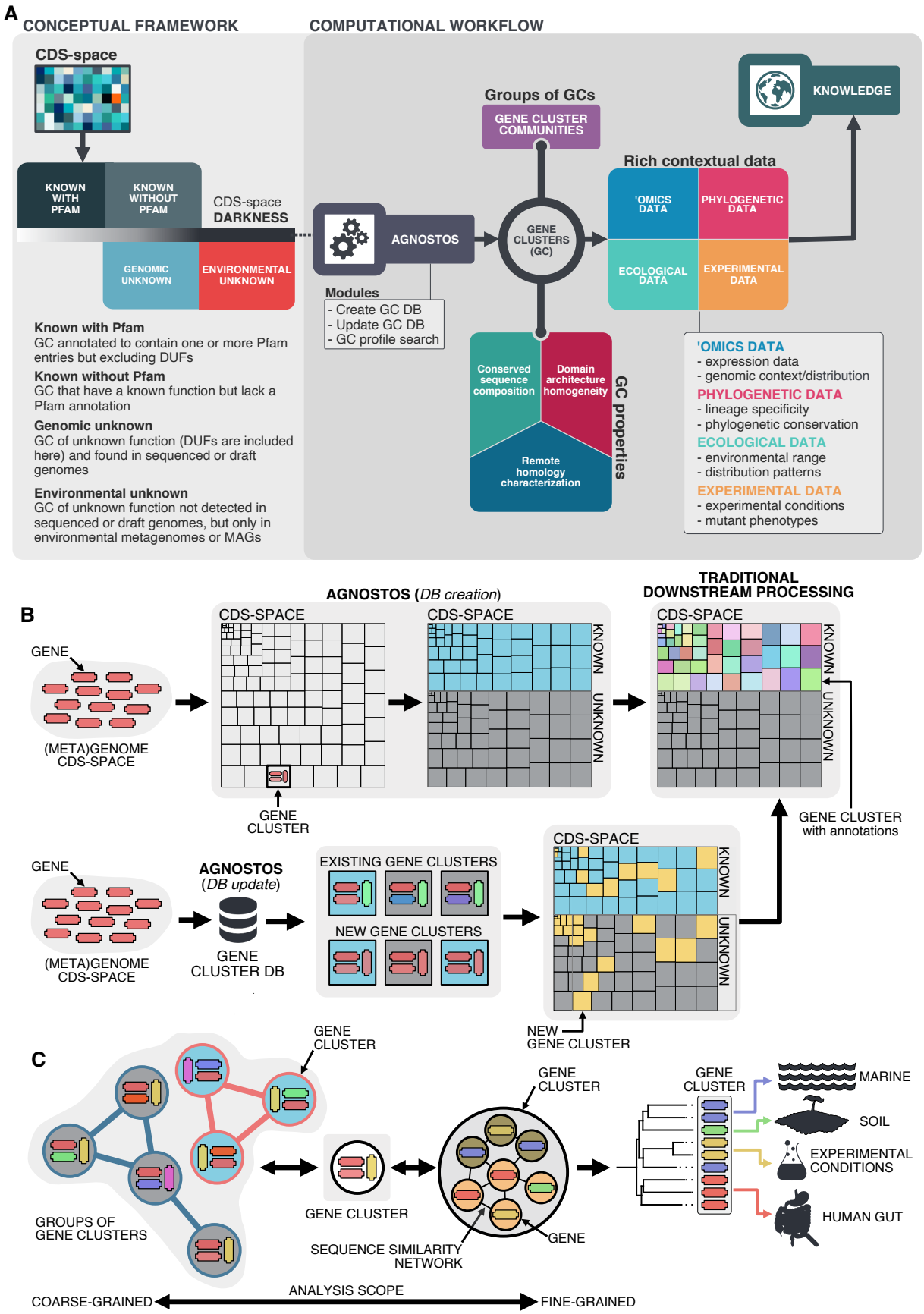
Thousands of isolate, single-cell, and metagenome-assembled genomes are guiding us towards a better understanding of life on Earth[1–8]. At the same time, the ever-increasing number of genomes and metagenomes unlocking uncharted regions of life's diversity[1,9,10] are providing new perspectives on the evolution of life[11,12]. However, our rapidly growing inventories of new genes have a glaring issue: between 40% and 60% cannot be assigned to a known function[13–16]. Current analytical approaches for genomic and metagenomic data[17–21] generally do not include this uncharacterized fraction in downstream analyses, constraining their results to

67 conserved pathways and housekeeping functions[18]. This inability to handle shades of the
68 unknown is an immense impediment to realizing the potential for discovery of microbiology and
69 molecular biology at large[13,22]. Predicting function from traditional single sequence similarity
70 appears to have yielded all it can[23–25], thus several groups have attempted to resolve gene
71 function by other means. Such efforts include combining biochemistry and crystallography[26];
72 using environmental co-occurrence[27]; by grouping those genes into evolutionarily related
73 families[28–31]; using remote homologies[32,33]; or more recently using deep learning
74 approaches[34,35]. In 2018, Price et al.[14] developed a high-throughput experimental pipeline that
75 provides mutant phenotypes for thousands of bacterial genes of unknown function being one of
76 the most promising methods to tackle the unknown. Despite their promise, experimental
77 methods are labor-intensive and require novel computational methods that could bridge the
78 existing gap between the known and unknown coding sequence space (CDS-space).
79 Here we present a conceptual framework and a computational workflow that closes the gap
80 between the known and unknown CDS-space by connecting genomic and metagenomic gene
81 clusters. Our approach adds context to vast amounts of unknown biology, providing an
82 invaluable resource to get a better understanding of the unknown functional fraction and boost
83 the current methods for its experimental characterization. The application of our approach to
84 415,971,742 genes predicted from 1,749 metagenomes and 28,941 bacterial and archaeal
85 genomes revealed that the unknown fraction (1) is smaller than expected, (2) is exceptionally
86 diverse, and (3) is phylogenetically more conserved and predominantly lineage-specific at the
87 Species level. Finally, we show how we can connect all the outputs produced by our approach
88 to augment the results from experimental data and add context to genes of unknown function
89 through hypothesis-driven molecular investigations.

Results

A conceptual framework and a computational workflow to unify the known and the unknown coding sequence space

We created the conceptual and technical foundations to unify the known and unknown CDS-space facilitating the integration of the genes of unknown function into ecological, evolutionary and biotechnological investigations. First, we conceptually partitioned the known and unknown fractions into (1) Known with Pfam annotations (K), (2) Known without Pfam annotations (KWP), (3) Genomic unknown (GU), and (4) Environmental unknown (EU) (Fig. 1A). The framework introduces a subtle change of paradigm compared to traditional approaches where our objective is to provide the best representation of the unknown space. We gear all our efforts towards finding sequences without any evidence of known homologies by pushing the search space beyond the *twilight zone* of sequence similarity[36]. With this objective in mind, we use gene clusters (GCs) instead of genes as the fundamental unit to compartmentalize the CDS-space owing to their unique characteristics (Fig. 1B). (1) GCs produce a structured CDS-space reducing its complexity (Fig. 1B), (2) are independent of the known and unknown fraction, (3) are conserved across environments and organisms, and (4) can be used to aggregate information from different sources (Fig. 1A). Moreover, the GCs provide a good compromise in terms of resolution for analytical purposes, and owing to their unique properties, one can perform analyses at different scales. For fine-grained analyses, we can exploit the gene associations within each GC; and for coarse-grained analyses, we can create groups of GCs based on their shared homologies (Fig. 1B).


A
CONCEPTUAL FRAMEWORK
COMPUTATIONAL WORKFLOW
CDS-space
KNOWN WITH PFAM
KNOWN WITHOUT PFAM
CDS-space DARKNESS
GENOMIC UNKNOWN
ENVIRONMENTAL UNKNOWN
Known with Pfam
GC annotated to contain one or more Pfam entries but excluding DUFs
Known without Pfam
GC that have a known function but lack a Pfam annotation
Genomic unknown
GC of unknown function (DUFs are included here) and found in sequenced or draft genomes
Environmental unknown
GC of unknown function not detected in sequenced or draft genomes, but only in environmental metagenomes or MAGs
AGNOSTOS
Modules
- Create GC DB
- Update GC DB
- GC profile search
Groups of GCs
GENE CLUSTER COMMUNITIES
GENE CLUSTERS (GC)
Conserved sequence composition
Domain architecture homogeneity
Remote homology characterization
GC properties
Rich contextual data
'OMICS DATA
PHYLOGENETIC DATA
ECOLOGICAL DATA
EXPERIMENTAL DATA
KNOWLEDGE
'OMICS DATA
- expression data
- genomic context/distribution
PHYLOGENETIC DATA
- lineage specificity
- phylogenetic conservation
ECOLOGICAL DATA
- environmental range
- distribution patterns
EXPERIMENTAL DATA
- experimental conditions
- mutant phenotypes
B
AGNOSTOS (DB creation)
TRADITIONAL DOWNSTREAM PROCESSING
GENE
(META)GENOME CDS-SPACE
CDS-SPACE
KNOWN
UNKNOWN
GENE CLUSTER
GENE CLUSTER with annotations
AGNOSTOS (DB update)
GENE CLUSTER DB
EXISTING GENE CLUSTERS
NEW GENE CLUSTERS
NEW GENE CLUSTER
C
GENE CLUSTER
GROUPS OF GENE CLUSTERS
SEQUENCE SIMILARITY NETWORK
GENE
MARINE
SOIL
EXPERIMENTAL CONDITIONS
HUMAN GUT
COARSE-GRAINED
ANALYSIS SCOPE
FINE-GRAINED

112

113 **Figure 1:** Conceptual framework to unify the known and unknown CDS-space and integration of the
114 framework in the current analytical workflows (A) Link between the conceptual framework and the
115 computational workflow to partition the CDS-space in the four conceptual categories. AGNOSTOS infers,
116 validates and refines the GCs and combines them in gene cluster communities (GCCs). Then, it classifies
117 them in one of the four conceptual categories based on their level of 'darkness'. Finally, we add context to
118 each GC based on several sources of information, providing a robust framework for the generation of
119 hypotheses that can be used to augment experimental data. (B) The computational workflow provides two
120 mechanisms to structure the CDS-space using GCs, de novo creation of the GCs (*DB creation*), or
121 integration of the dataset in an existing GC database (*DB update*). The structured CDS-space can then
122 be plugged into traditional analytical workflows to annotate the genes within each GC of the known
123 fraction. With AGNOSTOS, we provide the opportunity to easily integrate the unknown fraction into the
124 current microbiome analyses. C) The versatility of the GCs enables analyses at different scales
125 depending on the scope of our experiments. We can group GCs in gene cluster communities based on
126 their shared homologies to perform coarse-grained analyses. On the other hand, we can design fine-
127 grained analyses using the relationships between the genes in a GC, i.e., detecting network modules in
128 the GC inner sequence similarity network. Additionally, the fact that GCs are conserved across
129 environments, organisms and experimental conditions gives us access to an unprecedented amount of
130 information to design and interpret experimental data.
131

132 Driven by the concepts defined in the conceptual framework, we developed AGNOSTOS, a
133 computational workflow that infers, validates, refines, and classifies GCs in the four proposed
134 categories (Fig. 1A; Fig. 1B; Supp. Fig 1). AGNOSTOS provides two operational modules (*DB*
135 *creation* and *DB update*) to produce GCs with a highly conserved intra-homogeneous structure
136 (Fig. 1B), both in terms of sequence similarity and domain architecture homogeneity; it exhausts
137 any existing homology to known genes and provides a proper delimitation of the unknown CDS-
138 space before classifying each GC in one of the four categories (Methods). In the last step, we
139 decorate each GC with a rich collection of contextual data that we compile from different
140 sources, or that we generate by analyzing the GC contents in different contexts (Fig. 1A). For
141 each GC, we also offer several products that can be used for analytical purposes like improved
142 representative sequences, consensus sequences, sequence profiles for MMseqs2[37] and
143 HHblits[38], or the GC members as a sequence similarity network (Methods). To complement the
144 collection, we also provide a subset of what we define as *high-quality* GCs. The defining criteria
145 are (1) the representative is a complete gene and (2) more than one-third of genes within a GC
146 are complete genes.

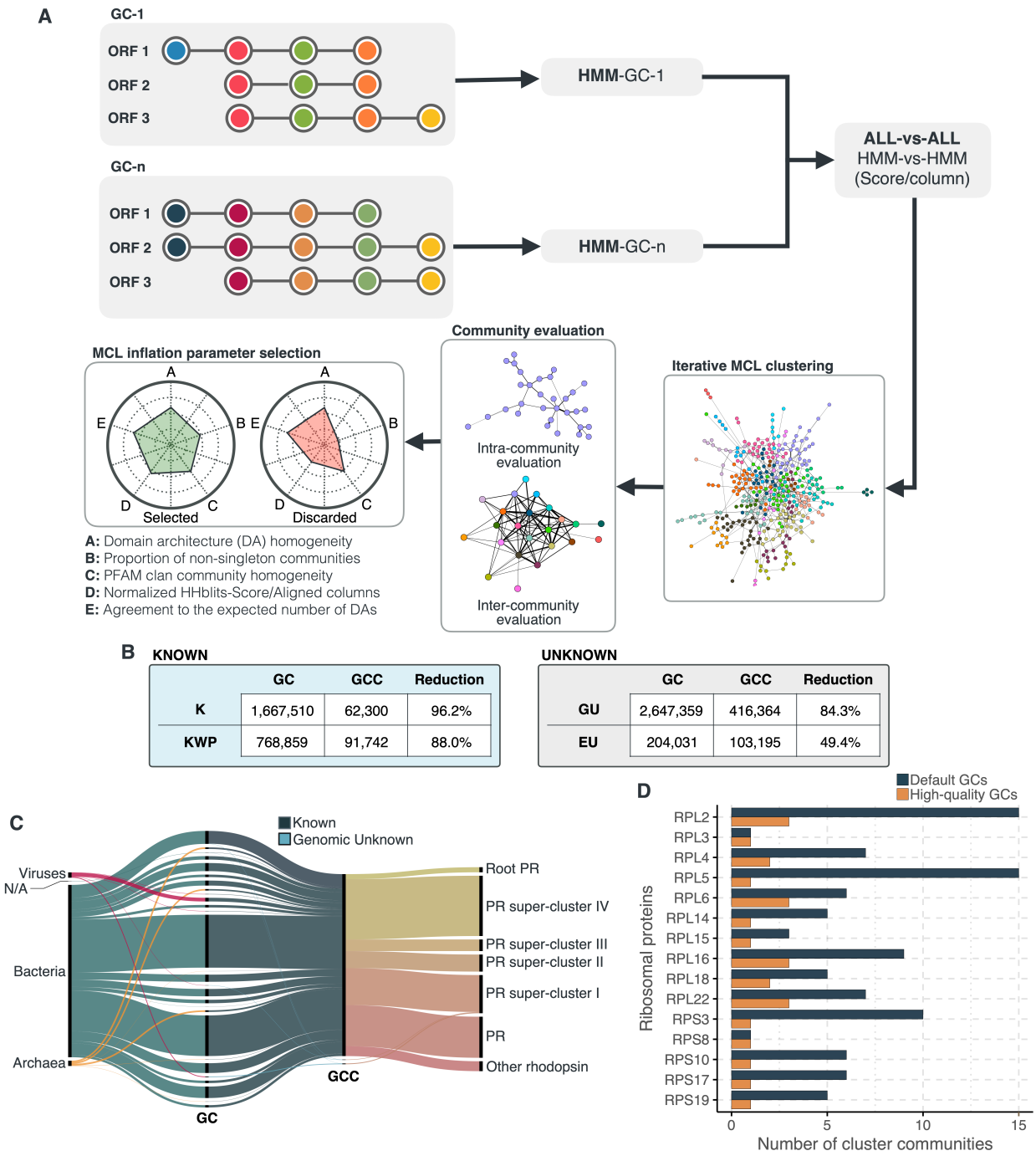
Partitioning and contextualizing the coding sequence space of genomes and metagenomes

We used our approach to explore the unknown CDS-space of 1,749 microbial metagenomes derived from human and marine environments, and 28,941 genomes from GTDB_r86 (Supp Fig 2A). The initial gene prediction of AGNOSTOS (Supp Fig 1) produced 322,248,552 genes from the environmental dataset and assigned Pfam annotations to 44% of them. Next, it clustered the predicted genes in 32,465,074 GCs. For the downstream processing, we kept 3,003,897 GCs (83% of the original genes) after filtering out any GC that contained less than ten genes[39] removing 9,549,853 clusters and 19,911,324 singletons (Supp Fig 2A; Supp. Note 1). The validation process selected 2,940,257 *good-quality* clusters (Fig. 1B; Supp. Table 1; Supp. Note 2), which resulted in 43% of them being members of the unknown CDS-space after the classification and remote homology refinement steps (Supp Fig 2A, Supp. Note 3). We build the link between the environmental and genomic CDS-space by expanding the final collection of GCs with the genes predicted from GTDB_r86 (Supp Fig 2A). Our environmental GCs already included 72% of the genes from GTDB_r86; 22% of them created 2,400,037 new GCs, and the rest 6% resulted in singleton GCs (Supp Fig 2A; Supp. Note 4; Supp. Note 5). The final dataset includes 5,287,759 GCs (Supp Fig 2A), with both datasets sharing only 922,599 GCs (Supp Fig 2B). The addition of the GTDB_r86 genes increased the proportion of GCs in the unknown CDS-space to 54%. As the final step, the workflow generated a subset of 203,217 *high-quality* GCs (Supp Table 2; Supp Fig 3). In these *high-quality* clusters, we identified 12,313 clusters potentially encoding for small proteins (<= 50 amino acids). Most of these GCs are unknown (66% of them), which agrees with recent findings on novel small proteins from metagenomes[40]. The KWP category contains the largest proportion of incomplete ORFs (Supp. Table 3), disrupting the detection and assignment of Pfam domains. But it also incorporates sequences with an unusual amino acid composition that has homology to proteins with high levels of

172 disorder in the DPD database[41] and that have characteristic functions of the intrinsically
173 disordered proteins[42] (IDP) like cellular processes and signaling as predicted by eggNOG
174 annotations (Supp. Table 4).
175 As part of the workflow, each GC is complemented with a rich set of information, as shown in
176 Fig 1A (Supp. Table 5; Supp Note 6).

177 Beyond the twilight zone, communities of gene clusters

178 The method we developed to group GCs in gene cluster communities (GCCs) (Fig. 2A) reduced
179 the final collection of GCs by 87%, producing 673,601 GCCs (Methods; Fig. 2B; Supp. Note 7).
180 We validated the ability of our approach in capturing remote homologies between related GCs
181 using two well-known gene families present in our environmental datasets, proteorhodopsins[43]
182 and bacterial ribosomal proteins[44]. Our dataset contained 64 GCs (12,184 genes) and 3 GCCs
183 (Supp Note 8) classified as proteorhodopsin (PR). One *Known* GCC contained 99% of the PR
184 annotated genes (Fig. 2C), with the only exception of 85 genes taxonomically annotated as viral
185 and assigned to the *PR Supercluster I*[45] within two GU communities (five GU gene clusters;
186 Supp Note 8). For the ribosomal proteins, the results were not so satisfactory. We identified
187 1,843 GCs (781,579 genes) and 98 GCCs. The number of GCCs is larger compared to the
188 expected number of ribosomal proteins families (16) used for validation. When we use *high-*
189 *quality* GCs (Supp. Note 8), we get closer to the expected number of GCCs (Fig. 2D). With this
190 subset, we identified 26 GCCs and 145 GCs (1,687 genes). The cross-validation of our method
191 against the approach used in Méheust et al.[44] (Supp. Note 9) confirms the intrinsic complexity of
192 analyzing metagenomic data. Both approaches showed a high agreement in the GCCs
193 identified (Supp. Table 9-1). Still, our method inferred fewer GCCs for each of the ribosomal
194 protein families (Supplementary Figure 9-3), coping better with the nuisances of a metagenomic
195 setup, like incomplete genes (Supp. Table 6).
196



	GC	GCC	Reduction
K	1,667,510	62,300	96.2%
KWP	768,859	91,742	88.0%

	GC	GCC	Reduction
GU	2,647,359	416,364	84.3%
EU	204,031	103,195	49.4%

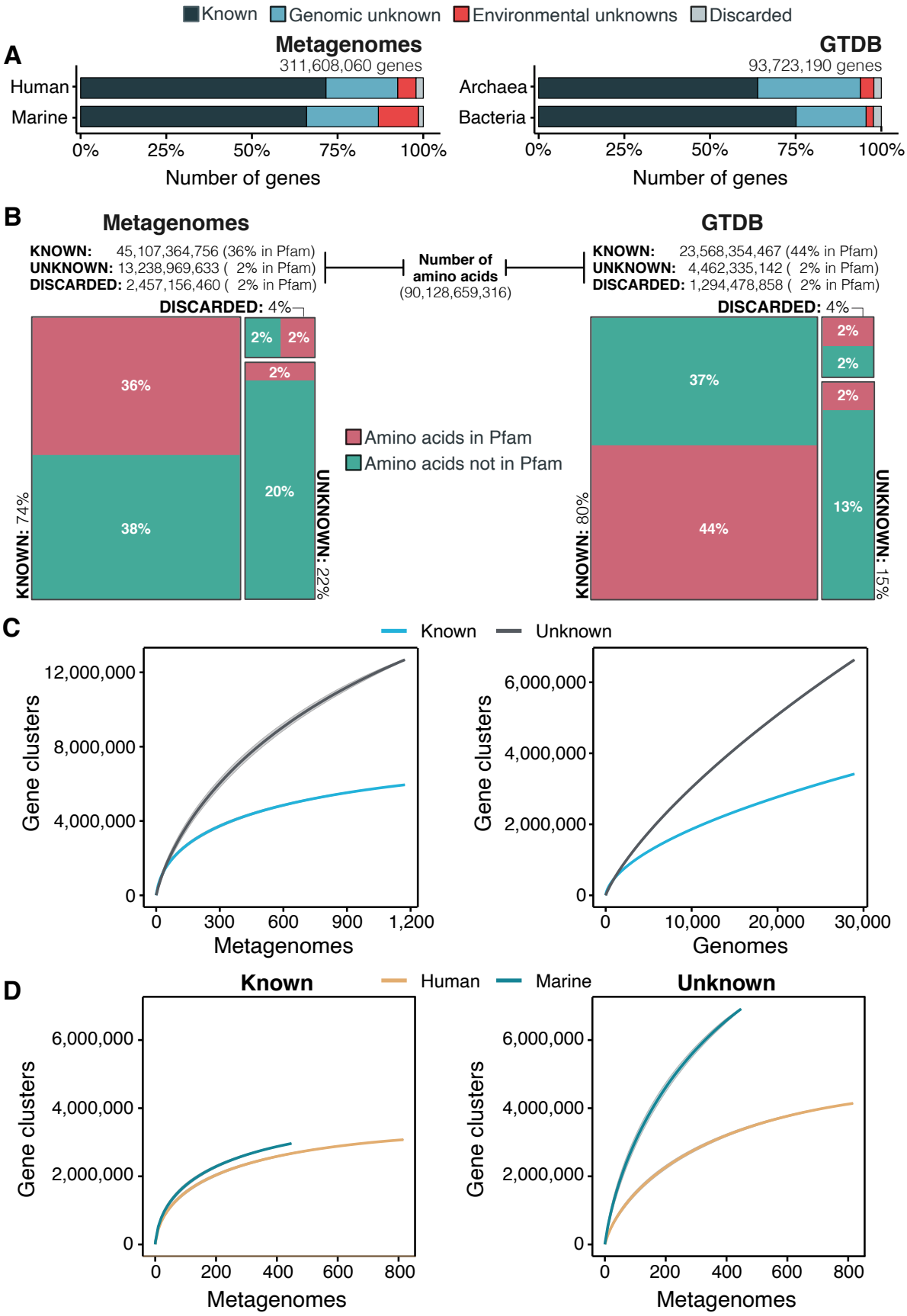
Figure 2: Overview and validation of the workflow to aggregate GCs in communities. (A) We inferred a gene cluster homology network using the results of an all-vs-all HMM gene cluster comparison with HHBLITS. The edges of the network are based on the HHblits-score/Aligned-columns. Communities are identified by an iterative screening of different MCL inflation parameters and evaluated using five different metrics that take into account the inter- and intra-community properties. (B) Comparison of the number of GCs and GCCs for each of the functional categories. (C) Validation of the GCCs inference based on the environmental genes annotated as proteorhodopsins. Ribbons in the alluvial plot are genes, and each stacked bar corresponds (from left to right) to the (1) gene taxonomic classification at the domain level,

206 (2) GC membership, (3) GCC membership and (4) MicRhoDE operational classification. (D) Validation of
207 the GCCs inference based on ribosomal proteins based on standard and high-quality GCs.

208 A smaller but highly diverse unknown coding sequence space

209 Combining clustering and remote homology searches reduce the extent of the unknown CDS-
210 space compared to what is reported by the traditional genomic and metagenomic analysis
211 approaches (Fig. 3A). Our workflow recruited as much as 71% of genes in human-related
212 metagenomic samples and 65% of the genes in marine metagenomes into the known CDS-
213 space. In both human and marine microbiomes, the genomic unknown fraction showed a similar
214 proportion of genes (21%, Fig. 3A). The number of genes corresponding to EU gene clusters is
215 higher in marine metagenomes; in total, 12% of the genes are part of this GC category. We
216 obtained a comparable result when we evaluated the genes from the GTDB_r86, 75% of
217 bacterial and 64% of archaeal genes were part of the known CDS-space. Archaeal genomes
218 contained more unknowns than those from Bacteria, where 30% of the genes are classified as
219 genomic unknowns in Archaea, and only 20% in Bacteria (Fig. 3A; Supp. Table 7). Our
220 approach allows us to go beyond genes, and for the first time we can provide a detailed
221 characterization of the CDS-space at the amino acid level. From the 90,128,659,316 amino
222 acids analyzed, the majority of the amino acids in metagenomes (74%) and in GTDB_r86 (80%)
223 are in the known CDS-space (Fig. 3B; Supp. Table 7). In both cases, approximately 40% of the
224 amino acids in the known CDS-space were part of a Pfam domain (Fig. 3B; Supp. Table 7). The
225 proportion of amino acids in the unknown CDS-space ranged from the 22% in metagenomes
226 and 15% in GTDB_r86. In both cases, only 2% of the amino acids in the unknown CDS-space
227 were covered by a Pfam domain. To evaluate the differences between both fractions of the
228 CDS-space, we calculated the accumulation rates of GCs and GCCs. For the metagenomic
229 dataset we used 1,264 metagenomes (18,566,675 GCs and 282,580 GCCs) and for the
230 genomic dataset 28,941 genomes (9,586,109 GCs and 496,930 GCCs). The rate of

231 accumulation of unknown GCs was three times higher than the known (2 times for the
232 genomic), and both cases were far from reaching a plateau (Fig. 3C). This is not the case for
233 the GCC accumulation curves (Supp Fig 4B), where they reached a plateau. The rate of
234 accumulation is largely determined by the number of singletons, and especially singletons from
235 EUs (Supp note 11 and Supp Fig 5). While the accumulation rate of known GCs between
236 marine and human metagenomes is almost identical, there are striking differences for the
237 unknown GCs (Fig. 3D). These differences are maintained even when we remove the virus-
238 enriched samples from the marine metagenomes (Supp Fig 4A). Although the marine
239 metagenomes include a large variety of environments, from coastal to the deep sea, the known
240 space remains quite constrained.
241 Despite only including marine and human metagenomes in our database, our coverage to other
242 databases and environments is quite comprehensive, with an overall coverage of 76% (Supp.
243 Note 12). The lowest covered biomes are freshwater, soil and human non-digestive as revealed
244 by the screening of MGnify[17] (release 2018_09; 11 biomes; 843,535,6116 proteins) where we
245 assigned 74% of the MGnify proteins into one of our categories (Supplementary Fig. 6).
246 Furthermore, as a result of this evaluation, we classified 20% of the FunkFams[29] and 44% of the
247 unknowns used by Price et al.[14] to the known fraction (Supp. Table 12-1).


Known
Genomic unknown
Environmental unknowns
Discarded
A
Metagenomes
311,608,060 genes
Human
Marine
0%
25%
50%
75%
100%
Number of genes
GTDB
93,723,190 genes
Archaea
Bacteria
0%
25%
50%
75%
100%
Number of genes
B
Metagenomes
KNOWN: 45,107,364,756 (36% in Pfam)
UNKNOWN: 13,238,969,633 (2% in Pfam)
DISCARDED: 2,457,156,460 (2% in Pfam)
Number of amino acids
(90,128,659,316)
GTDB
KNOWN: 23,568,354,467 (44% in Pfam)
UNKNOWN: 4,462,335,142 (2% in Pfam)
DISCARDED: 1,294,478,858 (2% in Pfam)
DISCARDED: 4%
36%
38%
2%
2%
2%
20%
KNOWN: 74%
UNKNOWN: 22%
Amino acids in Pfam
Amino acids not in Pfam
DISCARDED: 4%
37%
44%
2%
2%
2%
13%
KNOWN: 80%
UNKNOWN: 15%
C
Known
Unknown
Gene clusters
12,000,000
8,000,000
4,000,000
0
0
300
600
900
1,200
Metagenomes
Gene clusters
6,000,000
4,000,000
2,000,000
0
0
10,000
20,000
30,000
Genomes
D
Known
Human
Marine
Unknown
Gene clusters
6,000,000
4,000,000
2,000,000
0
0
200
400
600
800
Metagenomes
Gene clusters
6,000,000
4,000,000
2,000,000
0
0
200
400
600
800
Metagenomes

248

249 **Figure 3:** The extent of the known and unknown coding sequence space (A) Proportion of genes in the
250 known and unknown. (B) Amino acid distribution in the known and unknown CDS-space. (C)
251 Accumulation curves for the known and unknown CDS-space at the GC- level for the metagenomic and
252 genomic data. from TARA, MALASPINA, OSD2014 and HMP-I/II projects. (D) Collector curves comparing
253 the human and marine biomes. Colored lines represent the mean of 1000 permutations and shaded in
254 grey the standard deviation. Non-abundant singleton clusters were excluded from the accumulation
255 curves calculation.

256 The unknown coding sequence space has a limited ecological 257 distribution in human and marine environments

258 Although the role of the unknown fraction in the environment is still a mystery, the large number
259 of gene counts and abundance observed underlines its inherent ecological relevance (Fig. 4A).
260 In some metagenomes, the genomic unknown fraction can account for more than 40% of the
261 total gene abundance observed (Fig. 4A). The environmental unknown fraction is also relevant
262 in several samples, where singleton GCs are the majority (Fig. 4A). We identified two
263 metagenomes with an unusual composition in terms of environmental unknown singletons. The
264 marine metagenome corresponds to a sample from Lake Faro (OSD42), a meromictic saline
265 with a unique extreme environment where Archaea plays an important role[46]. The HMP
266 metagenome (SRS143565) corresponds to a human sample from the right cubital fossa from a
267 healthy female subject. To understand the unusual composition of this metagenome, we should
268 perform further analyses to discard potential technical artifacts like sample contamination. The
269 ratio between the unknown and known GCs revealed that the metagenomes located at the
270 upper left quadrant in Fig. 4B-C are enriched in GCs of unknown function. In human
271 metagenomes, we can distinguish between body sites, with the gastrointestinal tract, where
272 microbial communities are expected to be more diverse and complex, especially enriched with
273 genomic unknowns. The HMP metagenomes with the largest ratio of unknowns are those
274 samples identified to contain crAssphages[47,48] and HPV viruses[49] (Supp. Table 8; Supp. Fig. 7).
275 Consistently, in marine metagenomes (Fig. 4D) we can separate between size fractions, where

the highest ratio in genomic and environmental unknowns corresponds to the ones enriched with viruses and giant viruses.

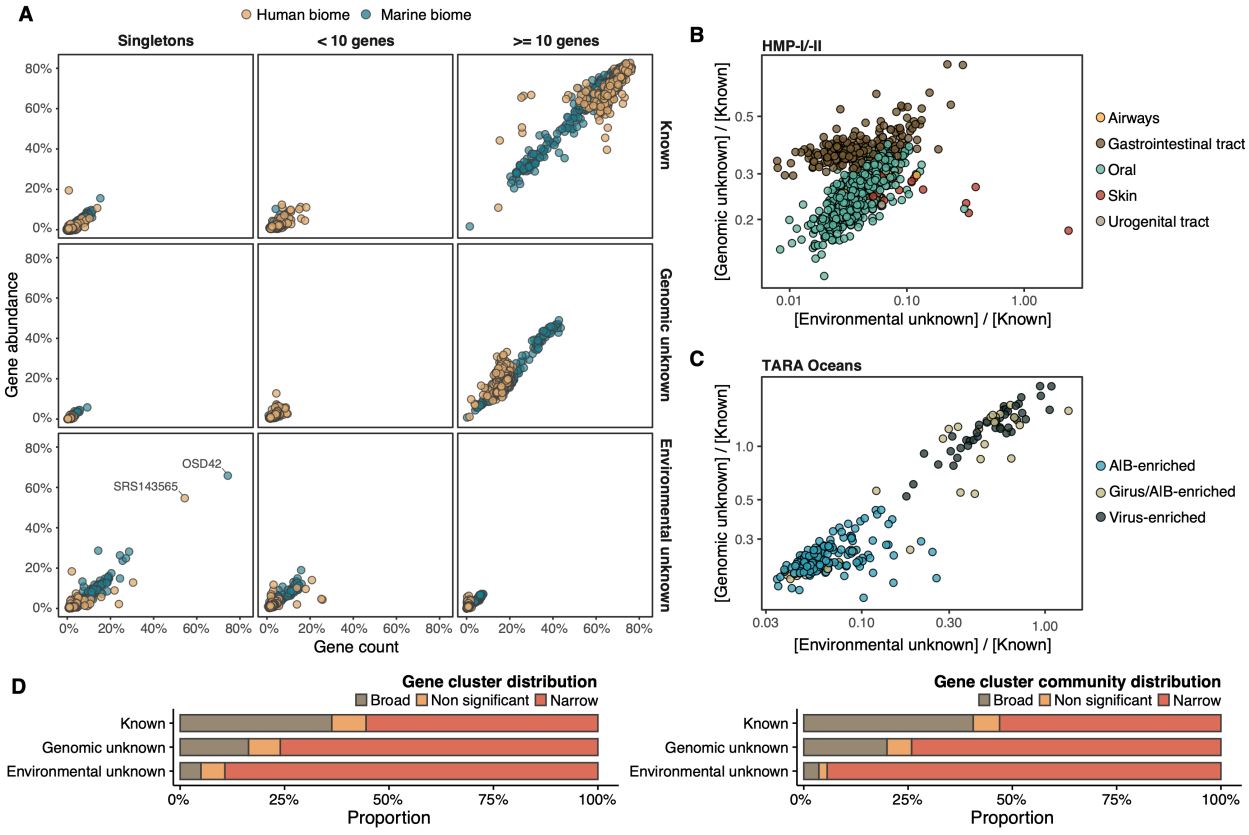


Figure 4: Distribution of the unknown coding sequence space in the human and marine metagenomes (A) Ratio between the proportion of the number of genes and their estimated abundances per cluster category and biome. Columns represented in the facet depicts three cluster categories based on the size of the clusters. (B) Relationship between the ratio of Genomic unknowns and Environmental unknowns in the HMP-I/II metagenomes. Gastrointestinal tract metagenomes are enriched in Genomic unknown coding sequences compared to the other body sites. (C) Relationship between the ratio of Genomic unknowns and Environmental unknowns in the TARA Oceans metagenomes. Girus and virus enriched metagenomes show a higher proportion of both unknown coding sequences (genomic and environmental) compared to the Archaea|Bacteria enriched fractions. (D) Environmental distribution of GCs and GCCs based on Levin's niche breadth index. We obtained the significance values after generating 100 null gene cluster abundance matrices using the *quasiswap* algorithm.

To complement the previous findings, we performed a large-scale analysis to investigate the GC occurrence patterns in the environment. The narrow distribution of the unknown fraction (Fig. 4D) suggests that these GCs might provide a selective advantage and be necessary for the

294 adaptation to specific environmental conditions. But the pool of broadly distributed
295 environmental unknowns is the most interesting result. We identified traces of potential
296 ubiquitous organisms left uncharacterized by traditional approaches, as more than 80% of these
297 GCs cannot be associated with a metagenome-assembled genome (MAG) (Supp Table 9,
298 Supp. Note 10).

299 The genomic unknown coding sequence space is lineage-specific

300 With the inclusion of the genomes from GTDB_r86, we have access to a phylogenomic
301 framework that can be used to assess how exclusive is a GC within a lineage (lineage-
302 specifity[50]) and how conserved is a GC across clades (phylogenetic conservation[51]). We
303 identified 781,814 lineage-specific GCs and 464,923 phylogenetically conserved ($P < 0.05$) GCs
304 in Bacteria (Supp. Table 10; Supp. Note 13 for Archaea). The number of lineage-specific GCs
305 increases with the Relative Evolutionary Distance[12] (Fig. 5A) and differences between the
306 known and the unknown fraction start to be evident at the Family level resulting in 4X more
307 unknown lineage-specific GCs at the Species level. The unknown GCs are more
308 phylogenetically conserved than the known (Fig. 5B, $p < 0.0001$), revealing the importance of
309 the genome's uncharacterized fraction. However, this is not the case for the lineage-specific and
310 phylogenetically conserved GCs, where the unknown GCs are less phylogenetically conserved
311 (Fig. 5B), agreeing with the large number of lineage-specific GCs at Genus and Species level.
312 To discard the possibility that the lineage-specific GCs of unknown function have a viral origin,
313 we screened all GTDB_r86 genomes for prophages. We only found 37,163 lineage-specific
314 GCs in prophage genomic regions, being 86% of them GCs of unknown function. After unveiling
315 the potential relevance of the GCs of unknown function in bacterial genomes, we identified
316 phyla in GTDB_r86 enriched with these types of clusters. A clear pattern emerged when we
317 partitioned the phyla based on the ratio of known to unknown GCs and vice versa (Fig. 5D), the
318 phyla with a larger number of MAGs are enriched in GCs of unknown function Figure 5D. Phyla

with a high proportion of non-classified GCs (those discarded during the validation steps) contain a small number of genomes and are primarily composed of MAGs. These groups of phyla highly enriched in unknowns and represented mainly by MAGs include newly described phyla such as *Cand.* Riflebacteria and *Cand.* Patescibacteria[10,52,53], both with the largest unknown to known ratio.

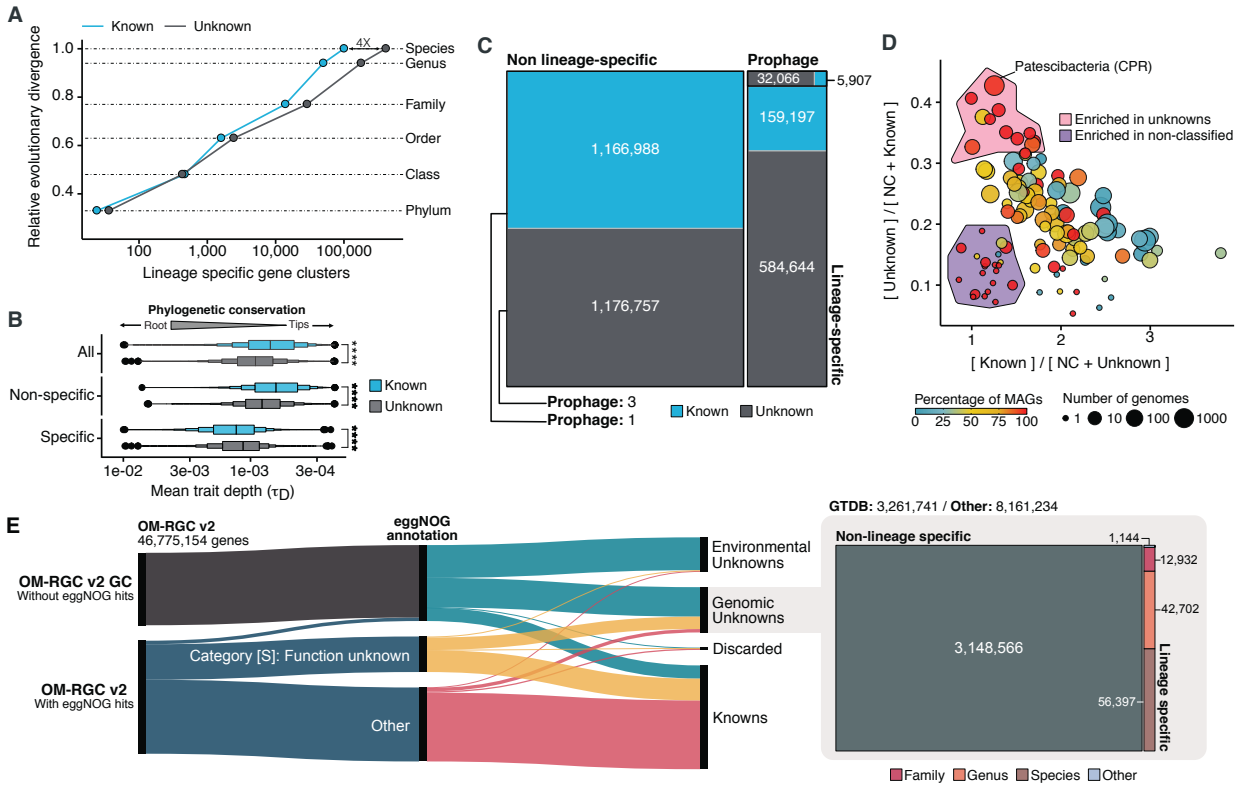


Figure 5: Phylogenomic exploration of the unknown coding sequence space. (A) Distribution of the lineage-specific GCs by taxonomic level. Lineage-specific unknown GCs are more abundant in the lower taxonomic levels (Genus, Species). (B) Phylogenetic conservation of the known and unknown coding sequence space in 27,372 bacterial genomes from GTDB_r86. We observe differences in the conservation between the known and the unknown coding sequence space for lineage- and non-lineage specific GCs (paired Wilcoxon rank-sum test; all p-values < 0.0001). (C) The majority of the lineage-specific clusters are part of the unknown coding sequence space, being a small proportion found in prophages present in the GTDB_r86 genomes. (D) Known and unknown coding sequence space of the 27,732 GTDB_r86 bacterial genomes grouped by bacterial phyla. Phyla are partitioned based on the ratio of known to unknown GCs and vice versa. Phyla enriched in MAGs have higher proportions in GCs of unknown function. Phyla with a high proportion of non-classified clusters (NC; discarded during the validation steps) tend to contain a small number of genomes. (E) The left side of the alluvial plot shows the uncharacterized (OM-RGC v2 GC) and characterized (OM-RGC v2) fraction of the gene catalog. The functional annotation is based on the eggNOG annotations provided by Salazar et al.[54]. The right side of

the alluvial plot shows the new organization of the OM-RGC v2 coding sequence space based on the approach described in this study. The treemap in the right links the metagenomic and genomic space adding context to the unknown fraction of the OM-RGC v2

We demonstrate the possibility to bridge genomic and metagenomic data and simultaneously unify the known and unknown CDS-space by integrating the new Ocean Microbial Reference Gene Catalog[54] (OM-RGC v2) in our database. We assigned 26,170,875 genes to known GCs, 11,422,975 to genomic unknowns, 8,661,221 to environmental unknown and 520,083 were discarded. From the 11,422,975 genes classified as genomic unknowns, we could associate 3,261,741 to a GTDB_r86 genome and we identified 113,175 as lineage-specific. The alluvial plot in Fig. 5E depicts the new organization of the OM-RGC v2 after being integrated into our framework, and how we can provide context to the two original types of unknowns in the OM-RGC (those annotated as category S in eggNOG[55] and those without known homologs in the eggNOG database[54]) that can lead to potential experimental targets at the organism level to complement the metatranscriptomic approach proposed by Salazar et al[54]..

A structured coding sequence space augments the interpretation of experimental data

We selected one of the experimental conditions tested in Price et al.[14] to demonstrate the potential of our approach to augment experimental data. We compared the fitness values in plain rich medium with added Spectinomycin dihydrochloride pentahydrate to the fitness in plain rich medium (LB) in *Pseudomonas fluorescens FW300-N2C3* (Fig. 6A). This antibiotic inhibits protein synthesis and elongation by binding to the bacterial 30S ribosomal subunit and interferes with the peptidyl tRNA translocation. We identified the gene with locus id AO356_08590 that presents a strong phenotype (fitness = -3.1; t = -9.1) and has no known function. This gene belongs to the genomic unknown GC GU_19737823. We can track this GC into the environment and explore the occurrence in the different samples we have in our

367 database. As expected, the GC is mostly found in non-human metagenomes (Fig. 6B) as
368 *Pseudomonas* are common inhabitants of soil and water environments[56]. However, finding this
369 GC also in human-related samples is very interesting, due to the potential association of *P.*
370 *fluorescens* and human disease where Crohn's disease patients develop serum antibodies to
371 this microbe[57]. We can add another layer of information to the selected GC by looking at the
372 associated remote homologs in the GCC GU_c_21103 (Fig. 6C). We identified all the genes in
373 the GTDB_r86 genomes that belong to the GCC GU_c_21103 (Supp. Table 11) and explored
374 their genomic neighborhoods. All members from GU_c_21103 are constrained to the class
375 *Gammaproteobacteria*, and interestingly GU_19737823 is mostly exclusive to the order
376 *Pseudomonadales*. The gene order in the different genomes analyzed is highly conserved,
377 finding GU_19737823 after the *rpsF*::*rpsR* operon and before *rplI*. *rpsF* and *rpsR* encode for
378 30S ribosomal proteins, the prime target of spectinomycin. The combination of the experimental
379 evidence and the associated data inferred by our approach provides strong support to generate
380 the hypothesis that the gene AO356_08590 might be involved in the resistance to
381 spectinomycin.
382

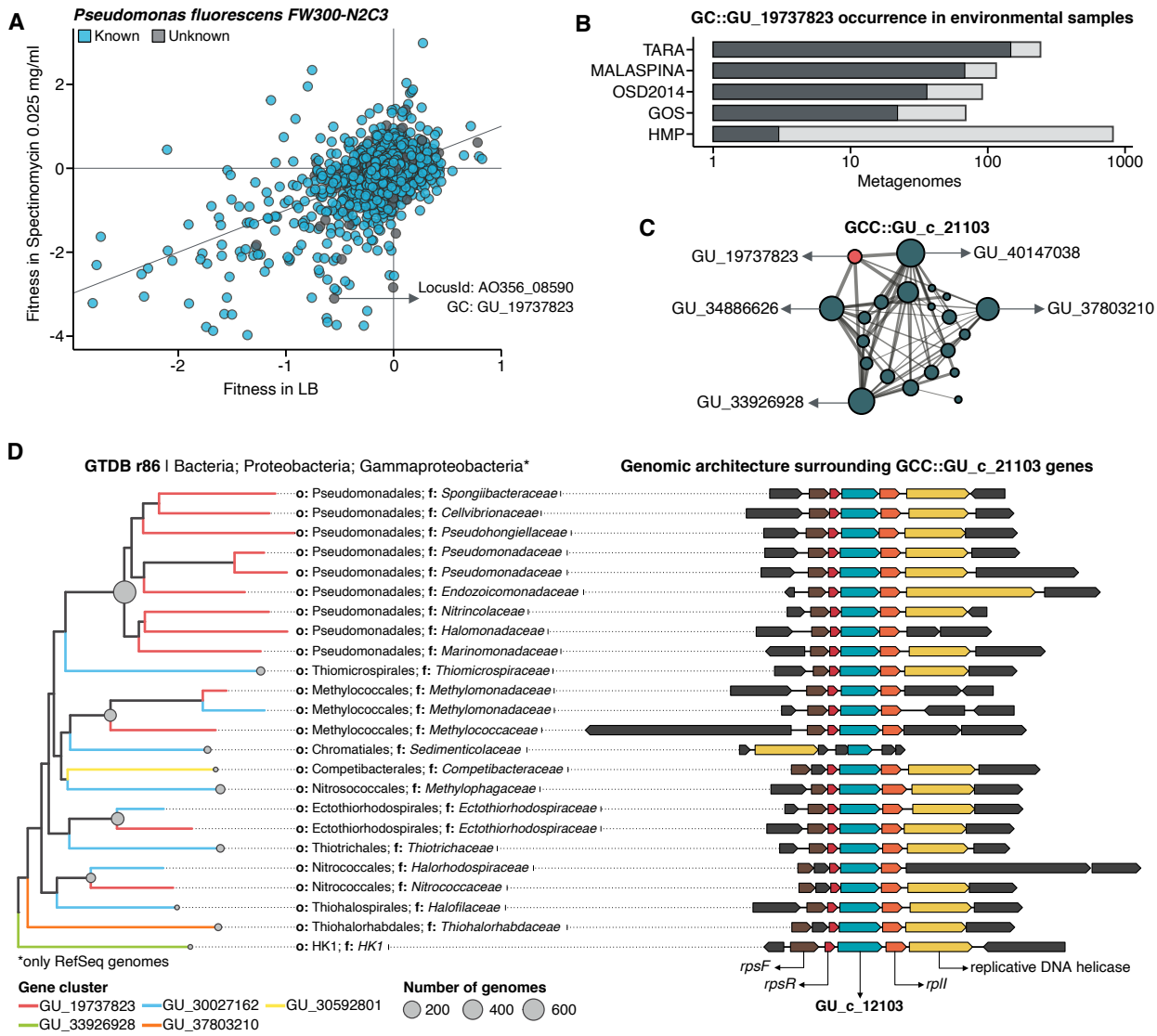


Figure 6: Augmenting experimental data with GCs of unknown function. (A) We used the fitness values from the experiments from Price et al.[14] to identify genes of unknown function that are important for fitness under certain experimental conditions. The selected gene belongs to the genomic unknown GC GU_19737823 and presents a strong phenotype (fitness = -3.1; t = -9.1) (B) Occurrence of GU_19737823 in the metagenomes used in this study. Darker bars depict the number of metagenomes where the GC is found. (C) GU_19737823 is a member of the GCC GU_c_21103. The network shows the relationships between the different GCs members of the gene cluster community GU_c_21103. The size of the node corresponds to the node degree of each GC. Edge thickness corresponds to the bitscore/column metric. Highlighted in red is GU_19737823. (D) We identified all the genes in the GTDB_r86 genomes that belong to the GCC GU_c_21103 and explored their genomic neighborhoods. GU_c_21103 members were constrained to the class Gammaproteobacteria, and GU_19737823 is mostly exclusive to the order Pseudomonadales. The gene order in the different genomes analyzed is highly conserved, finding GU_19737823 after the *rpsF::rpsR* operon and before *rplI. rpsF* and *rpsR* encode for the *30S ribosomal protein S6* and *30S ribosomal protein S18,* respectively. The GTDB_r86 subtree only shows RefSeq

399 genomes. Branch colors correspond to the different GCs found in GU_c_21103. The bubble plot depicts
400 the number of genomes with a gene that belongs to GU_c_21103.

401 Discussion

402 We present a new conceptual framework and computational workflow to unify the known and
403 unknown CDS-space. Using this framework, we performed an in-depth exploration of the
404 microbial unknown CDS-space, demonstrating that we can link the unknown fraction of
405 metagenomic studies to specific genomes and provide a powerful tool for hypothesis
406 generation. During the last years, the microbiome community has established a standard
407 operating procedure[18] for analyzing metagenomes that we can briefly summarize into (1)
408 assembly, (2) gene prediction, (3) gene catalog inference, (4) binning, and (5) characterization.
409 Thanks to recent computational developments[37,58], we envisioned an alternative to this workflow
410 to maximize the information used when analyzing genomic and metagenomic data. In addition,
411 we provide a mechanism to reconcile top-down and bottom-up approaches, thanks to the well-
412 structured CDS-space proposed by our framework. AGNOSTOS can create environmental- and
413 organism-specific variations of a seed GC database. Then, it integrates the predicted genes
414 from new genomes and metagenomes and dynamically creates and classifies new GCs with
415 those genes not integrated during the initial step (Fig. 1B). Afterward, the potential functions of
416 the known GCs can be carefully characterized by incorporating them into the traditional
417 workflows.
418 One of the most appealing characteristics of our approach is that the GCs provide unified
419 groups of homologous genes across environments and organisms indifferently if they belong to
420 the known or unknown CDS-space, and we can contextualize the unknown fraction using this
421 genomic and environmental information. Our combination of partitioning and contextualization
422 features a smaller unknown CDS-space than we expected. On average, for our genomic and
423 metagenomic data, only 30% of the genes fall in the unknown fraction. One hypothesis to

424 reconcile this surprising finding is that the methodologies to identify remotely homologous
425 sequences in large datasets were computationally prohibitive until recently. New methods[37,38],
426 like the ones used in AGNOSTOS, are enabling large scale distant homology searches. Still,
427 one has to apply conservative measures to control the trade-off between specificity and
428 sensitivity to avoid overclassification.
429 We found that most of the coding sequence space at the gene and amino acid level is known,
430 both in genomes and metagenomes. However, it presents a high diversity, as shown in the GC
431 accumulation curves highlighting the vast remaining untapped microbial fraction and its potential
432 importance for niche adaptation owing to its narrow ecological distribution. In a genomic context
433 and after ruling out the effect of prophages, the unknown fraction is predominantly Species'
434 lineage-specific and phylogenetically more conserved than the known fraction, supporting the
435 signal observed in the environmental data emphasizing that we should not ignore the unknown
436 fraction. It is worth noting that the high diversity observed in the unknowns only represents the
437 20% of the amino acids in the CDS-space we analyzed, and only 10% of this unknown amino
438 acid space is part of a Pfam domain (DUF and others). This contrasts with the numbers
439 observed in the known CDS-space, where Pfam domains include 50% of the amino acids. All
440 this evidence combined strengthens the hypothesis that the genes of unknown function,
441 especially the lineage-specific ones, might be associated with the mechanisms of microbial
442 diversification and niche adaptation due to the constant diversification of gene families and the
443 survival of new gene lineages[59,60].
444 Metagenome-assembled genomes are not only unveiling new regions of the microbial universe
445 (42% of the genomes in GTDB_r86), but they are also enriching genes of unknown function in
446 the tree of life. We investigated the unknown CDS-space of *Cand*. Patescibacteria, more
447 commonly known as Candidate Phyla Radiation (CPR), a phylum that has raised considerable
448 interest due to its unusual biology[10]. We provide a collection of 54,343 lineage-specific GCs of

449 unknown function at different taxonomic level resolutions (Supp. Table 12; Supp. Note 14),
450 which will be a valuable resource for the CPR advancement research efforts.
451 Our effort to tackle the unknown provides a pathway to unlock a large pool of likely relevant data
452 that remains untapped to analysis and discovery. By identifying a potential target gene of
453 unknown function for antibiotic resistance, we demonstrate the value of our approach and how it
454 can boost insights from model organism experiments. But severe challenges remain, such as
455 the dependence on the quality of the assemblies and their gene predictions, as shown by the
456 analysis of the ribosomal protein GCCs where many of the recovered genes are incomplete.
457 While sequence assembly has been an active area of research[61], this has not been the case for
458 gene prediction methods[61], which are becoming outdated[62] and cannot cope with the current
459 amount of data. Alternatives like protein-level assembly[63] combined with exploring the assembly
460 graphs' neighborhoods[64] become very attractive for our purposes. In any case, we still face the
461 challenge of discriminating between real and artifactual singletons[65]. Currently, there are no
462 methods available to provide a plausible solution and, at the same time, being scalable. We
463 devise a potential solution in the recent developments in unsupervised deep learning methods
464 where they use large corpora of proteins to define a language model *embedding* for protein
465 sequences[66]. These models could be applied to predict *embeddings* in singletons, which could
466 be clustered or used to determine their coding potential. Another issue is that we might be
467 creating more GCs than expected. We follow a conservative approach to avoid mixing multi-
468 domain proteins in GCs owing to the fragmented nature of the metagenome assemblies that
469 could result in the split of a GC. However, not only splitting can be a problem, but also lumping
470 unrelated genes or GCs owing to the use of remote homologies. Although the inference of
471 GCCs is using very sensitive methods to compare profile HMMs, low sequence diversity in GCs
472 can limit its effectiveness. Moreover, our approach is affected by the presence and propagation
473 of contamination in reference databases, a significant problem in 'omics[67,68]. In our case, we
474 only use Pfam as a source for annotation owing to its high-quality and manual curation process.

475 The categorization process of our GCs depends on the information from other databases, and
476 to minimize the potential impact of contamination, we apply methods that weight the annotations
477 of the identified homologs to discriminate if a GC belongs to the known or unknown CDS-space.
478 The results presented here prove that the integration and the analysis of the unknown fraction
479 are possible. We are unveiling a brighter future, not only for microbiome analyses but also for
480 boosting eukaryotic-related studies, thanks to the increasing number of projects including
481 metatranscriptomic data[8,69]. Furthermore, our work lays the foundations for further
482 developments of clear guidelines and protocols to define the different levels of unknown[70] and
483 should encourage the scientific community for a collaborative effort to tackle this challenge.
484

485 Material and methods

486 Genomic and metagenomic dataset

487 We used a set of 583 marine metagenomes from four of the major metagenomic surveys of the
488 ocean microbiome: Tara Oceans expedition (TARA)[2], Malaspina expedition[71], Ocean Sampling
489 Day (OSD)[3], and Global Ocean Sampling Expedition (GOS)[72]. We complemented this set with
490 1,246 metagenomes obtained from the Human Microbiome Project (HMP) phase I and II[73]. We
491 used the assemblies provided by TARA, Malaspina, OSD and HMP projects and the long
492 Sanger reads from GOS[74]. A total of 156M (156,422,969) contigs and 12.8M long-reads were
493 collected (Supp. Table 6).
494 For the genomic dataset, we used the 28,941 prokaryotic genomes (27,372 bacterial and 1,569
495 archaeal) from the Genome Taxonomy Database[12] (GTDB) Release 03-RS86 (19th August
496 2018).

497 ## Computational workflow development

498 We implemented a computation workflow based on Snakemake[75] for the easy processing of
499 large datasets in a reproducible manner. The workflow provides three different strategies to
500 analyze the data. The module *DB-creation* creates the gene cluster database, validates and
501 partitions the gene clusters (GCs) in the main functional categories. The module *DB-update*
502 allows the integration of new sequences (either at the contig or predicted gene level) in the
503 existing gene cluster database. In addition, the workflow has a *profile-search* function to quickly
504 screen samples using the gene cluster PSSM profiles in the database.

505 ## Metagenomic and genomic gene prediction

506 We used Prodigal (v2.6.3)[76] in metagenomic mode to predict the genes from the metagenomic
507 dataset. For the genomic dataset, we used the gene predictions provided by Annotree[50], since
508 they were obtained, consistently, with Prodigal v2.6.3. We identified potential spurious genes
509 using the *AntiFam* database[77]. Furthermore, we screened for '*shadow*' genes using the
510 procedure described in Yooseph et al.[78].

511 ## PFAM annotation

512 We annotated the predicted genes using the *hmmsearch* program from the *HMMER* package
513 (version: 3.1b2)[79] in combination with the Pfam database v31[80]. We kept the matches exceeding
514 the internal gathering threshold and presenting an independent e-value < 1e-5 and coverage >
515 0.4. In addition, we took into account multi-domain annotations, and we removed overlapping
516 annotations when the overlap is larger than 50%, keeping the ones with the smaller e-value.

517 ## Determination of the gene clusters

518 We clustered the metagenomic predicted genes using the cascaded-clustering workflow of the
519 MMseqs2 software[58] (“*--cov-mode 0 -c 0.8 --min-seq-id 0.3”*). We discarded from downstream
520 analyses the singletons and clusters with a size below a threshold identified after applying a
521 broken-stick model[81]. We integrated the genomic data into the metagenomic cluster database
522 using the “DB-update” module of the workflow. This module uses the *clusterupdate* module of
523 MMseqs2[37], with the same parameters used for the metagenomic clustering.

524 ## Quality-screening of gene clusters

525 We examined the GCs to ensure their high intra-cluster homogeneity. We applied two
526 methodologies to validate their cluster sequence composition and functional annotation
527 homogeneity. We identified non-homologous sequences inside each cluster combining the
528 identification of a new cluster representative sequence via a sequence similarity network (SSN)
529 analysis, and the investigation of intra-cluster multiple sequence alignments (MSAs), given the
530 new representative. Initially, we generated an SSN for each cluster, using the semi-global
531 alignment methods implemented in *PARASAIL*[82] (version 2.1.5). We trimmed the SSN using a
532 custom algorithm[83,84] that removes edges while maintaining the network structural integrity and
533 obtaining the smallest connected graph formed by a single component. Finally, the new cluster
534 representative was identified as the most central node of the trimmed SSN by the eigenvector
535 centrality algorithm, as implemented in igraph[85]. After this step, we built a multiple sequence
536 alignment for each cluster using *FAMSA*[86] (version 1.1). Then, we screened each cluster-MSA
537 for non-homologous sequences to the new cluster representative. Owing to computational
538 limitations, we used two different approaches to evaluate the cluster-MSAs. We used *LEON-*
539 *BIS*[87] for the clusters with a size ranging from 10 to 1,000 genes and OD-SEQ[88] for the clusters

540 with more than 1,000 genes. In the end, we applied a broken-stick model[81] to determine the
541 threshold to discard a cluster.
542 The predicted genes can have multi-domain annotations in different orders, therefore to validate
543 the consistency of intra-cluster Pfam annotations, we applied a combination of w-shingling[89] and
544 Jaccard similarity. We used w-shingling (k-shingle = 2) to group consecutive domain
545 annotations as a single object. We measured the homogeneity of the *shingle sets* (sets of
546 domains) between genes using the Jaccard similarity and reported the median similarity value
547 for each cluster. Moreover, we took into consideration the Clan membership of the Pfam
548 domains and that a gene might contain N-, C- and M-terminal domains for the functional
549 homogeneity validation. We discarded clusters with a median similarity < 1.
550 After the validation, we refined the gene cluster database removing the clusters identified to be
551 discarded and the clusters containing ≥ 30% *shadow genes*. Lastly, we removed the single
552 shadow, spurious and non-homologous genes from the remaining clusters (Supplementary Note
553 2).

554 Remote homology classification of gene clusters

555 To partition the validated GCs into the four main categories, we processed the set of GCs
556 containing Pfam annotated genes and the set of not annotated GCs separately. For the
557 annotated GCs, we inferred a consensus protein domain architecture (DA) (an ordered
558 combination of protein domains) for each annotated gene cluster. To identify each gene cluster
559 consensus DA, we created directed acyclic graphs connecting the Pfam domains based on their
560 topological order on the genes using *igraph*[85]. We collapsed the repetitions of the same domain.
561 Then we used the gene completeness as a positive-weighting value for the selection of the
562 cluster consensus DA. Within this step, we divided the GCs into “Knowns” (Known) if annotated
563 to at least one Pfam domains of known function (DKFs) and “Genomic unknowns” (GU) if
564 annotated entirely to Pfam domains of unknown function (DUFs).

565 We aligned the sequences of the non-annotated GCs with FAMSA[86] and obtained cluster
566 consensus sequences with the *hhconsensus* program from *HH-SUITE*[38]. We used the cluster
567 consensus sequences to perform a nested search against the UniRef90 database (release
568 2017_11)[90] and NCBI *nr* database (release 2017_12)[91] to retrieve non-Pfam annotations with
569 *MMSeqs2*[37] (“*-e 1e-05 --cov-mode 2 -c 0.6”*). We kept the hits within 60% of the Log(best-e-
570 value) and searched the annotations for any of the terms commonly used to define proteins of
571 unknown function (Supp. Table 12). We used a quorum majority voting approach to decide if a
572 gene cluster would be classified as *Genomic Unknown* or *Known without Pfams* based on the
573 annotations retrieved. We searched the consensus sequences without any homologs in the
574 UniRef90 database against NCBI *nr*. We applied the same approach and criteria described for
575 the first search. Ultimately, we classified as *Environmental Unknown* those GCs whose
576 consensus sequences did not align with any of the NCBI *nr* entries.
577 In addition, we developed some conservative measures to control the trade-off between
578 specificity and sensitivity for the remote homology searches such as (1) a modification of the
579 algorithm described in Hingamp et al.[92] to get a confident group of homologs to determine if a
580 query protein is known or unknown by a quorum majority voting approach (Supp Note 3); (2)
581 strict parameters in terms of iterations, bidirectional coverage and probability thresholds for the
582 HHblits alignments to minimize the inclusion of non-homologous sequences; and (3) avoid
583 providing annotations for our gene clusters, as we believe that annotation should be a careful
584 process done on a smaller scale and with experimental context.

585 Gene cluster remote homology refinement

586 We refined the *Environmental Unknown* GCs to ensure the lack of any characterization by
587 searching for remote homologies in the Uniclust database (release 30_2017_10) using the
588 HMM/HMM alignment method *HHblits*[93]. We created the HMM profiles with the *hhmake*
589 program from the *HH-SUITE*[38]. We only accepted those hits with an *HHblits-probability* ≥ 90%

590 and we re-classified them following the same majority vote approach as previously described.
591 The clusters with no hits remained as the refined set of EUs. We applied a similar refinement
592 approach to the KWP clusters to identify GCs with remote homologies to Pfam protein domains.
593 The KWP HMM profiles were searched against the Pfam *HH-SUITE* database (version 31),
594 using *HHblits*. We accepted hits with a probability ≥ 90% and a target coverage > 60% and
595 removed overlapping domains as described earlier. We moved the KWP with remote
596 homologies to known Pfams to the Known set, and those showing remote homologies to Pfam
597 DUFs to the GUs. The clusters with no hits remained as the refined set of KWP.

598 Gene cluster characterization

599 To retrieve the taxonomic composition of our clusters we applied the *MMseqs2 taxonomy*
600 program (version: b43de8b7559a3b45c8e5e9e02cb3023dd339231a), which allows computing
601 the lowest common ancestor through the implementation of the 2bLCA protocol [92]. We searched
602 all cluster genes against UniProtKB (release of January 2018) [94] using the following parameters
603 "*-e 1e-05 --cov-mode 0 -c 0.6"*. We parsed the results to keep only the hits within 60% of the
604 log10(best-e-value). To retrieve the taxonomic lineages, we used the R package *CHNOSZ*[95].
605 We measured the intra-cluster taxonomic admixture by applying the *entropy.empirical()* function
606 from the *entropy* R package[96]. This function estimates the Shannon entropy based on the
607 different taxonomic annotation frequencies. For each cluster, we also retrieved the cluster
608 consensus taxonomic annotation, which we defined as the taxonomic annotation of the majority
609 of the genes in the cluster.
610 In addition to the taxonomy, we evaluated the clusters' level of darkness and disorder using the
611 Dark Proteome Database (DPD)[41] as reference. We searched the cluster genes against the
612 DPD, applying the MMseqs2 search program[37] with "*-e 1e-20 --cov-mode 0 -c 0.6*". For each
613 cluster, we then retrieved the mean and the median level of darkness, based on the gene DPD
614 annotations.

615 ## High-quality clusters

616 We defined a subset of high-quality clusters based on the completeness of the cluster genes
617 and their representatives. We identified the minimum required percentage of complete genes
618 per cluster by a broken-stick model[81] applied to the percentage distribution. Then, we selected
619 the GCs found above the threshold and with a complete representative.

620 ## A set of non-redundant domain architectures

621 We estimated the number of potential domain architectures present in the *Known* GCs taking
622 into account the large proportion of fragmented genes in the metagenomic dataset and that
623 could inflate the number of potential domain architectures. To identify fragments of larger
624 domain architecture, we took into account their topological order in the genes. To reduce the
625 number of comparisons, we calculated the pairwise string cosine distance (q-gram = 3) between
626 domain architectures and discarded the pairs that were too divergent (cosine distance ≥ 0.9).
627 We collapsed a fragmented domain architecture to the larger one when it contained less than
628 75% of complete genes.

629 ## Inference of gene cluster communities

630 We aggregated distant homologous GCs into GCCs. The community inference approach
631 combined an all-vs-all HMM gene cluster comparison with Markov Cluster Algorithm (MCL)[97]
632 community identification. We started performing the inference on the Known GCs to use the
633 Pfam DAs as constraints. We aligned the gene cluster HMMs using HHblits[93] (-n 2 -Z 10000000
634 -B 10000000 -e 1) and we built a homology graph using the cluster pairs with probability ≥ 50%
635 and bidirectional coverage > 60%. We used the ratio between HHblits-bitscore and aligned-
636 columns as the edge weights (Supp. Note 9). We used MCL[97] (v. 12-068) to identify the
637 communities present in the graph. We developed an iterative method to determine the optimal

638 MCL inflation parameter that tries to maximize the relationship of five intra-/inter-community
639 properties: (1) the proportion of MCL communities with one single DA, based on the consensus
640 DAs of the cluster members; (2) the ratio of MCL communities with more than one cluster; (3)
641 the proportion of MCL communities with a PFAM clan entropy equal to 0; (4) the intra-
642 community HHblits-score/Aligned-columns score (normalized by the maximum value); and (5)
643 the number of MCL communities, which should, in the end, reflect the number of non-redundant
644 DAs. We iterated through values ranging from 1.2 to 3.0, with incremental steps of 0.1. During
645 the inference process, some of the GCs became orphans in the graph. We applied a three-step
646 approach to assigning a community membership to these GCs. First, we used less stringent
647 conditions (probability ≥ 50% and coverage >= 40%) to find homologs in the already existing
648 GCCs. Then, we ran a second iteration to find secondary relationships between the newly
649 assigned GCs and the missing ones. Lastly, we created new communities with the remaining
650 GCs. We repeated the whole process with the other categories (KWP, GU and EU), applying
651 the optimal inflation value found for the Known (2.2 for metagenomic and 2.5 for genomic data).

652 Gene cluster communities validation

653 We tested the biological significance of the GCCs using the phylogeny of proteorhodopsin[45]
654 (PR). We used the proteorhodopsin HMM profiles[43] to screen the marine metagenomic datasets
655 using *hmmsearch* (version 3.1b2)[79]. We kept the hits with a coverage > 0.4 and e-value <= 1e-
656 5. We removed identical duplicates from the sequences assigned to PR with CD-HIT[98] (v4.6)
657 and cleaned from sequences with less than 100 amino acids. To place the identified PR
658 sequences into the MicRhode[45] PR tree first, we optimized the initial tree parameters and
659 branch lengths with RAxML (v8.2.12)[99]. We used PaPaRA (v2.5)[100] to incrementally align the
660 query PR sequences against the MicRhode PR reference alignment and *pplacer*[101]
661 (v1.1.alpha19-0-g807f6f3) to place the sequences into the tree. Finally, we assigned the query
662 PR sequences to the MicRhode PR Superclusters based on the phylogenetic placement. We

663 further investigated the GCs annotated as viral (196 genes, 14 GC) comparing them to the six
664 newly discovered viral PRs[102] using Parasail[82] (-a sg_stats_scan_sse2_128_16 -t 8 -c 1 -x). As
665 an additional evaluation, we investigated the distributions of standard GCCs and HQ GCCs
666 within ribosomal protein families. We obtained the ribosomal proteins used for the analysis
667 combining the set of 16 ribosomal proteins from Méheust et al.[44] and those contained in the
668 collection of bacterial single-copy genes of Anvi'o[103]. Also, for the ribosomal proteins, we
669 compared the outcome of our method to the one proposed by Méheust et al.[44] (Supp. Note 9).

670 Metagenomic sample selection for downstream analyses

671 For the subsequent ecological analyses, we selected those metagenomes with a number of
672 genes larger or equal to the first quartile of the distribution of all the metagenomic gene counts.
673 (Supp. Table 13).

674 Gene cluster abundance profiles in genomes and metagenomes

675 We estimated abundance profiles for the metagenomic cluster categories using the read
676 coverage to each predicted gene as a proxy for abundance. We calculated the coverage by
677 mapping the reads against the assembly contigs using the *bwa-mem* algorithm from *BWA*
678 *mapper*[104]. Then, we used *BEDTOOLS*[105], to find the intersection of the gene coordinates to the
679 assemblies, and normalize the per-base coverage by the length of the gene. We calculated the
680 cluster abundance in a sample as the sum of the cluster gene abundances in that sample, and
681 the cluster category abundance in a sample as the sum of the cluster abundances. We obtained
682 the proportions of the different gene cluster categories applying a total-sum-scaling
683 normalization. For the genomic abundance profiles, we used the number of genes in the
684 genomes and normalized by the total gene counts per genome.

685 ## Rate of genomic and metagenomic gene clusters accumulation

686 We calculated the cumulative number of known and unknown GCs as a function of the number
687 of metagenomes and genomes. For each metagenome count, we generated 1000 random sets,
688 and we calculated the number of GCs and GCCs recovered. For this analysis, we used 1,246
689 HMP metagenomes and 358 marine metagenomes (242 from TARA and 116 from Malaspina).
690 We repeated the same procedure for the genomic dataset. We removed the singletons from the
691 metagenomic dataset with an abundance smaller than the mode abundance of the singletons
692 that got reclassified as good-quality clusters after integrating the GTDB data to minimize the
693 impact of potential spurious singletons. To complement those analyses, we evaluated the
694 coverage of our dataset by searching seven different state-of-the-art databases against our set
695 of metagenomic GC HMM profiles (Supp. Note 12).
696

697 ## Occurrence of gene clusters in the environment

698 We used 1,264 metagenomes from the TARA Oceans, MALASPINA Expedition, OSD2014 and
699 HMP-I/II to explore the properties of the unknown CDS-space in the environment. We applied
700 the Levins Niche Breadth (NB) index[106] to investigate the GCs and GCCs environmental
701 distributions. We removed the GCs and cluster communities with a mean relative abundance <
702 1e-5. We followed a divide-and-conquer strategy to avoid the computational burden of
703 generating the null-models to test the significance of the distributions owing to the large number
704 of metagenomes and GCs. First, we grouped similar samples based on the gene cluster content
705 using the Bray-Curtis dissimilarity[107] in combination with the *Dynamic Tree Cut*[108] R package.
706 We created 100 random datasets picking up one random sample from each group. For each of
707 the 100 random datasets, we created 100 random abundance matrices using the *nullmodel*
708 function of the *quasiswap* count method[109]. Then we calculated the *observed* NB and obtained

709 the 2.5% and 97.5% quantiles based on the randomized sets. We compared the observed and
710 quantile values for each gene cluster and defined it to have a *Narrow distribution* when the
711 *observed* was smaller than the 2.5% quantile and to have a *Broad distribution* when it was
712 larger than the 97.5% quantile. Otherwise, we classified the cluster as *Non-significant*[110]. We
713 used a majority voting approach to get a consensus distribution classification based on the ten
714 random datasets.

715 Identification of prophages in genomic sequences

716 We used PhageBoost (https://github.com/ku-cbd/PhageBoost/) to find gene regions in the
717 microbial genomes that result in high viral signals against the overall genome signal. We set the
718 following thresholds to consider a region prophage: minimum of 10 genes, maximum 5 gaps,
719 single-gene probability threshold 0.9. We further smoothed the predictions using Parzen rolling
720 windows of 20 periods and looked at the smoothed probability distribution across the genome.
721 We disregarded regions that had a summed smoothed probability less than 0.5, and those
722 regions that did differ from the overall population of the genes in a genome by using Kruskal–
723 Wallis rank test (p-value 0.001).

724 Lineage-specific gene clusters

725 We used the F1-score developed for AnnoTree[50] to identify the lineage-specific GCs and to
726 which rank they are specific. Following similar criteria to the ones used in Mendler et al.[50], we
727 considered a gene cluster to be lineage-specific if it is present in less than half of all genomes
728 and at least 2 with F1-score > 0.95.

Phylogenetic conservation of gene clusters

We calculated the phylogenetic conservation (τD) of each gene cluster using the *consenTRAIT*[51] function implemented in the R package *castor*[51]. We used a paired Wilcoxon rank-sum test to compare the average τD values for lineage-specific and non-specific GCs.

Evaluation of the OM-RGC v2 uncharacterized fraction

We integrated the 46,775,154 genes from the second version of the TARA Ocean Microbial Reference Gene Catalog (OM-RGC v2)[54] into our cluster database using the same procedure as for the genomic data. We evaluated the uncharacterized fraction and the genes classified into the eggNOG[55] category S within the context of our database.

Augmenting RB-TnSeq experimental data

We searched the 37,684 genes of unknown function associated with mutant phenotypes from Price et al.[14] against our gene cluster profiles. We kept the hits with e-value ≤ 1e-20 and a query coverage > 60%. Then we filtered the results to keep the hits within 90% of the Log(best-e-value), and we used a majority vote function to retrieve the consensus category for each hit. Lastly, we selected the best-hits based on the smallest e-value and the largest query and target coverage values. We used the fitness values from the RB-TnSeq experiments from Price et al. to identify genes of unknown function that are important for fitness under certain experimental conditions.

Code and data availability

The code used for the analyses in the manuscript is available at https://github.com/functional-dark-side/functional-dark-side.github.io/tree/master/scripts. The code to recreate the figures is available at https://github.com/functional-dark-side/vanni_et_al-figures. Detailed descriptions of

751 the different methods and results of this manuscript are available at
752 https://dark.metagenomics.eu. The workflow AGNOSTOS is available at
753 https://github.com/functional-dark-side/agnostos-wf, and its database can be downloaded from
754 https://doi.org/10.6084/m9.figshare.12459056.

755

756

757 Acknowledgements

758 The authors thankfully acknowledge the computer resources at MareNostrum and the technical
759 support provided by Barcelona Supercomputing Center (RES-AECT-2014-2-0085), the BMBF-
760 funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI)
761 (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B),
762 the University of Oxford Advanced Research Computing
763 (http://dx.doi.org/10.5281/zenodo.22558) and the MARBITS bioinformatics core at ICM-CSIC.
764 CV was supported by the Max Planck Society. AFG received funding from the European
765 Union's Horizon 2020 research and innovation program Blue Growth: Unlocking the potential of
766 Seas and Oceans under grant agreement no. 634486 (project acronym INMARE). AM was
767 supported by the Biotechnology and Biological Sciences Research Council [BB/M011755/1,
768 BB/R015228/1] and RDF by the European Molecular Biology Laboratory core funds. EOC was
769 supported by project INTERACTOMA RTI2018-101205-B-I00 from the Spanish Agency of
770 Science MICIU/AEI. SGA and PS received additional funding by the project MAGGY
771 (CTM2017-87736-R) from the Spanish Ministry of Economy and Competitiveness. The
772 Malaspina 2010 Expedition was supported by the Spanish Ministry of Economy and
773 Competitiveness (MINECO) through the Consolider-Ingenio program (ref. CSD2008-00077).
774 The authors thank Johannes Söding and Alex Bateman for helpful discussions.

Author Contributions

CV, MSS and AF-G performed the analyses and wrote the computational workflow. MS assisted with the clustering and remote homology searches. KS helped with the identification of prophages in genomic sequences. PLB and AB provided feedback and assisted with the ecological analyses. RDF and AM provided feedback and information on the MGnify and Pfam databases. CMD, PS and SGA provided the Malaspina metagenomes. TOD and AME analyzed data in the context of metagenome-assembled genomes. AF-G conceived the study and supervised the work. CV and AF-G wrote the manuscript. All authors read, edited and approved the final manuscript.

Competing Interests

The authors declare no competing interests.

References

1. Hug, L. A. *et al.* A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).
2. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
3. Kopf, A. *et al.* The ocean sampling day consortium. *Gigascience* **4**, 27 (2015).
4. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
5. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e20 (2019).
6. Pachiadaki, M. G. *et al.* Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell* **179**, 1623-1635.e11 (2019).
7. Cross, K. L. *et al.* Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. *Nat. Biotechnol.* **37**, 1314–1321 (2019).
8. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. *Cold Spring Harbor Laboratory* 2020.10.15.341214 (2020) doi:10.1101/2020.10.15.341214.
9. Eloe-Fadrosh, E. A. *et al.* Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. *Nat. Commun.* **7**, 10476 (2016).
10. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
11. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
12. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).

816 13. Bernard, G., Pathmanathan, J. S., Lannes, R., Lopez, P. & Bapteste, E. Microbial Dark
817 Matter Investigations: How Microbial Studies Transform Biological Knowledge and
818 Empirically Sketch a Logic of Scientific Discovery. *Genome Biol. Evol.* **10**, 707–715 (2018).
819 14. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function.
820 *Nature* **557**, 503–509 (2018).
821 15. Carradec, Q. *et al.* A global ocean atlas of eukaryotic genes. *Nat. Commun.* **9**, 373 (2018).
822 16. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut
823 microbiome. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0603-3.
824 17. Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*
825 **48**, D570–D578 (2020).
826 18. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun
827 metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
828 19. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and
829 metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
830 20. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology
831 Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
832 21. Chen, I.-M. A. *et al.* IMG/M v.5.0: an integrated data management and comparative
833 analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–
834 D677 (2019).
835 22. Hanson, A. D., Pribat, A., Waller, J. C. & Crécy-Lagard, V. de. ‘Unknown’proteins and
836 ‘orphan’enzymes: the missing half of the engineering parts list--and how to find it. *Biochem.*
837 *J* **425**, 1–11 (2010).
838 23. Arnold, F. H. Design by Directed Evolution. *Acc. Chem. Res.* **31**, 125–131 (1998).
839 24. Brandenberg, O. F., Fasan, R. & Arnold, F. H. Exploiting and engineering hemoproteins for
840 abiological carbene and nitrene transfer reactions. *Curr. Opin. Biotechnol.* **47**, 102–111
841 (2017).

842 25. Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem. Int. Ed*
843 *Engl.* **57**, 4143–4148 (2018).
844 26. Jaroszewski, L. *et al.* Exploration of uncharted regions of the protein universe. *PLoS Biol.* **7**,
845 (2009).
846 27. Buttigieg, L. P. *et al.* Ecogenomic Perspectives on Domains of Unknown Function:
847 Correlation-Based Exploration of Marine Metagenomes. *PLoS One* **8**, (2013).
848 28. Yooseph, S. *et al.* The Sorcerer II global ocean sampling expedition: Expanding the
849 universe of protein families. *PLoS Biol.* **5**, 0432–0466 (2007).
850 29. Wyman, S. K., Avila-Herrera, A., Nayfach, S. & Pollard, K. S. A most wanted list of
851 conserved microbial protein families with no known domains. *PLoS One* **13**, e0205749
852 (2018).
853 30. Brum, J. R. *et al.* Illuminating structural proteins in viral “dark matter” with metaproteomics.
854 *Proc. Natl. Acad. Sci. U. S. A.* **113**, 2436–2441 (2016).
855 31. Bateman, A., Coggill, P. & Finn, D. R. DUFs: Families in search of function. *Acta*
856 *Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **66**, 1148–1152 (2010).
857 32. Lobb, B., Kurtz, D. A., Moreno-Hagelsieb, G. & Doxey, A. C. Remote homology and the
858 functions of metagenomic dark matter. *Front. Genet.* **6**, 1–12 (2015).
859 33. Bitard-Feildel, T. & Callebaut, I. Exploring the dark foldable proteome by considering
860 hydrophobic amino acids topology. *Sci. Rep.* **7**, 41425 (2017).
861 34. Bileschi, M. L. *et al.* Using Deep Learning to Annotate the Protein Universe. *bioRxiv* 626507
862 (2019) doi:10.1101/626507.
863 35. Liu, X. L. Deep Recurrent Neural Network for Protein Function Prediction from Sequence.
864 *bioRxiv* 103994 (2017) doi:10.1101/103994.
865 36. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng. Des. Sel.* **12**, 85–94
866 (1999).
867 37. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the

868 analysis of massive data sets. *Nat. Biotechnol.* **advance on**, (2017).
869 38. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein
870 annotation. *BMC Bioinformatics* **20**, 473 (2019).
871 39. Skewes-Cox, P., Sharpton, T. J., Pollard, K. S. & DeRisi, J. L. Profile hidden Markov
872 models for the detection of viruses within metagenomic sequence data. *PLoS One* **9**,
873 e105067 (2014).
874 40. Sberro, H. *et al.* Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small,
875 Novel Genes. *Cell* **178**, 1245-1259.e14 (2019).
876 41. Perdigão, N., Rosa, A. C. & O'Donoghue, S. I. The Dark Proteome Database. *BioData Min.*
877 **10**, 1–11 (2017).
878 42. Habchi, J., Tompa, P., Longhi, S. & Uversky, V. N. Introducing protein intrinsic disorder.
879 *Chem. Rev.* **114**, 6561–6588 (2014).
880 43. Olson, D. K., Yoshizawa, S., Boeuf, D., Iwasaki, W. & DeLong, E. F. Proteorhodopsin
881 variability and distribution in the North Pacific Subtropical Gyre. *ISME J.* **12**, 1047–1060
882 (2018).
883 44. Méheust, R., Burstein, D., Castelle, C. J. & Banfield, J. F. The distinction of CPR bacteria
884 from other bacteria based on protein family content. *Nat. Commun.* **10**, 4173 (2019).
885 45. Boeuf, D., Audic, S., Brillet-Guéguen, L., Caron, C. & Jeanthon, C. MicRhoDE: a curated
886 database for the analysis of microbial rhodopsin diversity and evolution. *Database* **2015**,
887 (2015).
888 46. La Cono, V. *et al.* Partaking of Archaea to biogeochemical cycling in oxygen-deficient
889 zones of meromictic saline Lake Faro (Messina, Italy). *Environ. Microbiol.* **15**, 1717–1733
890 (2013).
891 47. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread
892 human gut virus crAssphage. *Nat Microbiol* **4**, 1727–1736 (2019).
893 48. Dubinkina, V. B., Ischenko, D. S., Ulyantsev, V. I., Tyakht, A. V. & Alexeev, D. G.

894 Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC*
895 *Bioinformatics* vol. 17 (2016).
896 49. Ma, Y. *et al.* Human papillomavirus community in healthy persons, defined by
897 metagenomics analysis of human microbiome project shotgun sequencing data sets. *J.*
898 *Virol.* **88**, 4786–4797 (2014).
899 50. Mendler, K. *et al.* AnnoTree: visualization and exploration of a functionally annotated
900 microbial tree of life. *Nucleic Acids Res.* **47**, 4442–4448 (2019).
901 51. Martiny, A. C., Treseder, K. & Pusch, G. Phylogenetic conservatism of functional traits in
902 microorganisms. *ISME J.* **7**, 830–838 (2013).
903 52. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter.
904 *Nature* **499**, 431–437 (2013).
905 53. Anantharaman, K. *et al.* Expanded diversity of microbial groups that shape the dissimilatory
906 sulfur cycle. *ISME J.* **12**, 1715–1728 (2018).
907 54. Salazar, G. *et al.* Gene Expression Changes and Community Turnover Differentially Shape
908 the Global Ocean Metatranscriptome. *Cell* **179**, 1068-1083.e21 (2019).
909 55. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically
910 annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids*
911 *Res.* **47**, D309–D314 (2019).
912 56. Heffernan, B., Murphy, C. D. & Casey, E. Comparison of planktonic and biofilm cultures of
913 Pseudomonas fluorescens DSM 8341 cells grown on fluoroacetate. *Appl. Environ.*
914 *Microbiol.* **75**, 2899–2907 (2009).
915 57. Scales, B. S., Dickson, R. P., LiPuma, J. J. & Huffnagle, G. B. Microbiology, genomics, and
916 clinical significance of the Pseudomonas fluorescens species complex, an unappreciated
917 colonizer of humans. *Clin. Microbiol. Rev.* **27**, 927–948 (2014).
918 58. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat.*
919 *Commun.* **9**, 2542 (2018).

920 59. Francino, M. P. The ecology of bacterial genes and the survival of the new. *Int. J. Evol.*
921 *Biol.* **2012**, 394026 (2012).
922 60. Muller, E. E. L. Determining Microbial Niche Breadth in the Environment for Better
923 Ecosystem Fate Predictions. *mSystems* **4**, (2019).
924 61. Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I. & Watson, M. A Review
925 of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. *Front.*
926 *Genet.* **8**, 23 (2017).
927 62. Ivanova, N. N. *et al.* Stop codon reassignments in the wild. *Science* **344**, 909–913 (2014).
928 63. Steinegger, M., Mirdita, M. & Söding, J. Protein-level assembly increases protein sequence
929 recovery from metagenomic samples manyfold. *Nat. Methods* **16**, 603–606 (2019).
930 64. Titus Brown, C. *et al.* Exploring neighborhoods in large metagenome assembly graphs
931 reveals hidden sequence diversity. *bioRxiv* 462788 (2018) doi:10.1101/462788.
932 65. Höps, W., Jeffryes, M. & Bateman, A. Gene Unprediction with Spurio: A tool to identify
933 spurious protein sequences. *F1000Res.* **7**, 261 (2018).
934 66. Heinzinger, M. *et al.* Modeling aspects of the language of life through transfer-learning
935 protein sequences. *BMC Bioinformatics* **20**, 723 (2019).
936 67. Breitwieser, F. P., Pertea, M., Zimin, A. & Salzberg, S. L. Human contamination in bacterial
937 genomes has created thousands of spurious proteins. *Genome Res.* (2019)
938 doi:10.1101/gr.245373.118.
939 68. Steinegger, M. & Salzberg, S. L. Terminating contamination: large-scale search identifies
940 more than 2,000,000 contaminated entries in GenBank. *Genome Biol.* **21**, 115 (2020).
941 69. Vorobev, A. *et al.* Transcriptome reconstruction and functional analysis of eukaryotic
942 marine plankton communities via high-throughput metagenomics and metatranscriptomics.
943 *Genome Res.* (2020) doi:10.1101/gr.253070.119.
944 70. Thomas, A. M. & Segata, N. Multiple levels of the unknown in microbiome research. *BMC*
945 *Biol.* **17**, 48 (2019).

946 71. Duarte, C. M. Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation
947 Expedition. *Limnol. Oceanog. Bull.* **24**, 11–14 (2015).
948 72. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic
949 through Eastern Tropical Pacific. *PLoS Biol.* **5**, 1–34 (2007).
950 73. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome
951 Project. *Nature* **550**, 61–66 (2017).
952 74. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors.
953 *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).
954 75. Köster, J. Reproducible data analysis with Snakemake. *F1000Res.* **7**, (2018).
955 76. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
956 identification. *BMC Bioinformatics* **11**, 119–119 (2010).
957 77. Eberhardt, R. Y. *et al.* AntiFam: a tool to help identify spurious ORFs in protein annotation.
958 *Database* **2012**, bas003–bas003 (2012).
959 78. Yooseph, S., Li, W. & Sutton, G. Gene identification and protein classification in microbial
960 metagenomic sequence data via incremental clustering. *BMC Bioinformatics* **9**, 1–13
961 (2008).
962 79. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity
963 searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
964 80. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future.
965 *Nucleic Acids Res.* **44**, D279–D285 (2016).
966 81. Bennett, K. D. Determination of the number of zones in a biostratigraphical sequence. *New*
967 *Phytol.* **132**, 155–170 (1996).
968 82. Daily, J. Parasail: SIMD C library for global, semi-global, and local pairwise sequence
969 alignments. *BMC Bioinformatics* **17**, 81–81 (2016).
970 83. Žure, M., Fernandez-Guerra, A., Munn, C. B. & Harder, J. Geographic distribution at
971 subspecies resolution level: closely related Rhodopirellula species in European coastal

sediments. *ISME J.* **11**, 478–489 (2017).

84. Chafee, M. *et al.* Recurrent patterns of microdiversity in a temperate coastal marine environment. *ISME J.* **12**, 237–252 (2018).

85. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* vol. Complex Systems 1695 (2006).

86. Deorowicz, S., Debudaj-Grabysz, A. & Gudyś, A. FAMSA: Fast and accurate multiple sequence alignment of huge protein families. *Sci. Rep.* **6**, 33964–33964 (2016).

87. Vanhoutreve, R. *et al.* LEON-BIS: multiple alignment evaluation of sequence neighbours using a Bayesian inference system. *BMC Bioinformatics* **17**, 271–271 (2016).

88. Jehl, P., Sievers, F. & Higgins, D. G. OD-seq: outlier detection in multiple sequence alignments. *BMC Bioinformatics* **16**, 269–269 (2015).

89. Broder, A. Z. On the resemblance and containment of documents. in *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)* 21–29 (IEEE, 1997).

90. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).

91. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46**, D8–D13 (2018).

92. Hingamp, P. *et al.* Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* **7**, 1678–1695 (2013).

93. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2012).

94. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699 (2018).

95. Dick, J. M. Calculation of the relative metastabilities of proteins using the CHNOSZ software package. *Geochem. Trans.* **9**, 10 (2008).

998 96. Hausser, J. & Strimmer, K. Entropy inference and the James-Stein estimator, with
999 application to nonlinear gene association networks. *arXiv [stat.ML]* (2008).
1000 97. van Dongen, S. & Abreu-Goodger, C. Using MCL to Extract Clusters from Networks. in
1001 *Bacterial Molecular Networks: Methods and Protocols* (eds. van Helden, J., Toussaint, A. &
1002 Thieffry, D.) 281–295 (Springer New York, 2012).
1003 98. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein
1004 or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
1005 99. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
1006 phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
1007 100. Berger, S. A. & Stamatakis, A. PaPaRa 2.0: a vectorized algorithm for probabilistic
1008 phylogeny-aware alignment extension. *Heidelberg Institute for Theoretical Studies,*
1009 *http://sco. h-its. org/exelixis/publica tions. html. Exelixis-RRDR-2012-2015* (2012).
1010 101. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and
1011 Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC*
1012 *Bioinformatics* **11**, 538 (2010).
1013 102. Needham, D. M. *et al.* A distinct lineage of giant viruses brings a rhodopsin photosystem to
1014 unicellular marine predators. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 20574–20583 (2019).
1015 103. Murat Eren, A. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics
1016 data. *PeerJ* **3**, e1319 (2015).
1017 104. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform.
1018 *Bioinformatics* **26**, 589–595 (2010).
1019 105. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
1020 features. *Bioinformatics* **26**, 841–842 (2010).
1021 106. Levins, R. THE STRATEGY OF MODEL BUILDING IN POPULATION BIOLOGY. *Am. Sci.*
1022 **54**, 421–431 (1966).
1023 107. Bray, J. R., Roger Bray, J. & Curtis, J. T. An Ordination of the Upland Forest Communities

of Southern Wisconsin. *Ecological Monographs* vol. 27 325–349 (1957).

108. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).

109. Miklós, I. & Podani, J. RANDOMIZATION OF PRESENCE–ABSENCE MATRICES: COMMENTS AND NEW ALGORITHMS. *Ecology* vol. 85 86–92 (2004).

110. Salazar, G. *et al.* Particle-association lifestyle is a phylogenetically conserved trait in bathypelagic prokaryotes. *Mol. Ecol.* **24**, 5692–5706 (2015).