Light into the darkness: Unifying the known and unknown coding sequence space in microbiome analyses

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], Rob 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

Bridging the gap between the known and the unknown coding sequence space is one of the biggest challenges in molecular biology today. This challenge is especially extreme in microbiome analyses where between 40% and 60% of the coding sequences detected are of unknown function, and ignoring this fraction limits our understanding of microbial systems. Discarding the uncharacterized fraction is not an option anymore. Here, we present an in-depth exploration of the microbial unknown fraction through the lenses of a conceptual framework and a computational workflow we developed to unify the microbial known and unknown coding sequence space. Our approach partitions the coding sequence space in gene clusters and contextualizes them with genomic and environmental information. We analyzed 415,971,742 genes predicted from 1,749 metagenomes and 28,941 bacterial and archaeal genomes, putting into perspective the extent of the unknown fraction, its diversity, and its relevance in a genomic and environmental context. With the identification of a target gene of unknown function for antibiotic resistance, we demonstrate how a contextualized unknown coding sequence space provides a robust framework for the generation of hypotheses that can be used to augment experimental data.

Introduction

Thousands of isolate, single-cell, and metagenome-assembled genomes are guiding us towards a better understanding of how microbes shape life on Earth[1–7], thus bringing about a golden age of microbial genomics. An ever-increasing number of genomes and metagenomes are unlocking uncharted regions of microbial diversity[1,8,9], providing new perspectives on the evolution of life[10,11]. However, our rapidly growing inventories of new genes have a glaring issue: between 40% and 60% cannot be assigned to a known function[12–15]. Current analytical approaches for genomic and metagenomic data[16–20] generally do not include this uncharacterized fraction in downstream analyses, constraining their results to conserved pathways and housekeeping functions[17]. This inability to handle shades of the unknown is an immense impediment to realizing the potential for discovery of microbial genomics and microbiology at large[12,21].

Predicting function from traditional single sequence similarity appears to have yielded all it can[22–24], thus several groups have attempted to resolve gene function by other means. Such efforts include combining biochemistry and crystallography[25]; using environmental co-occurrence[26]; by grouping those genes into evolutionarily related families[27–30]; using remote homologies[31,32]; or more recently using deep learning approaches[33,34]. In 2018, Price et al.[13]

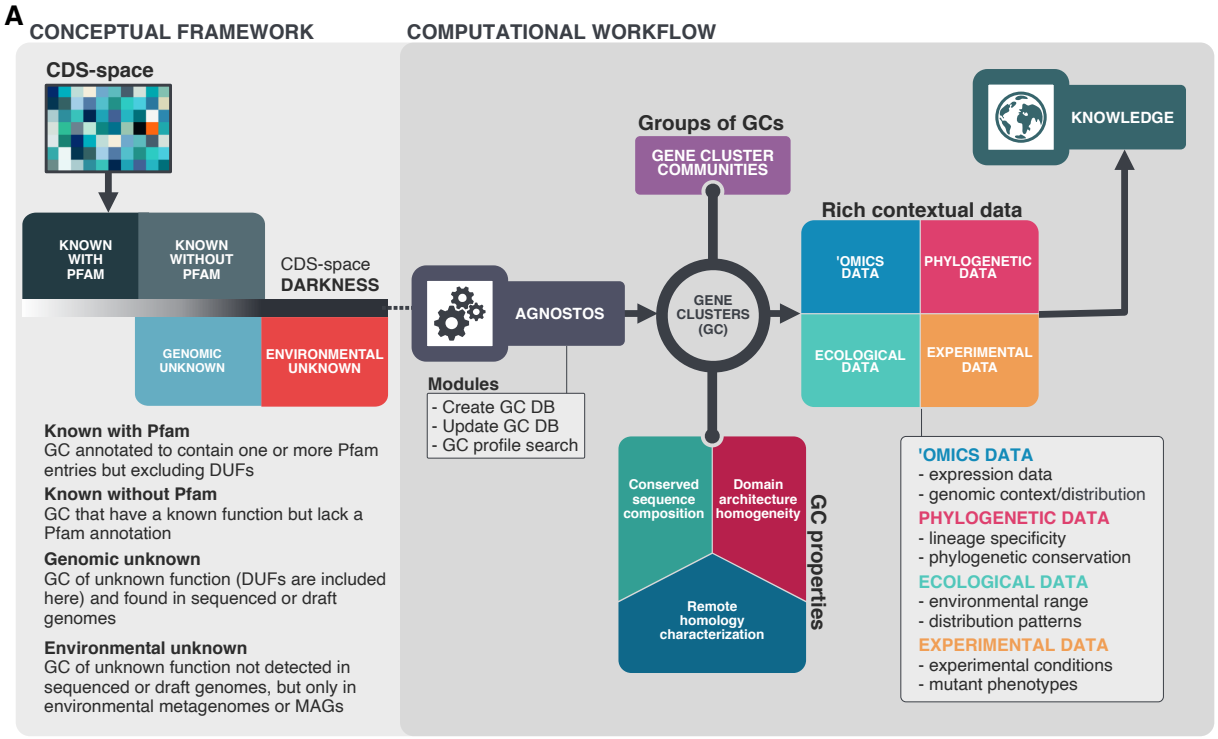
75 developed a high-throughput experimental pipeline that provides mutant phenotypes for
76 thousands of bacterial genes of unknown function being one of the most promising methods to
77 tackle the unknown. Despite their promise, experimental methods are labor-intensive and
78 require novel computational methods that could bridge the existing gap between the known and
79 unknown coding sequence space (CDS-space).
80 Here we present a conceptual framework and a computational workflow that closes the gap
81 between the known and unknown CDS-space by connecting genomic and metagenomic gene
82 clusters. Our approach adds context to vast amounts of unknown biology, providing an
83 invaluable resource to get a better understanding of the unknown functional fraction and boost
84 the current methods for its experimental characterization. The application of our approach to
85 415,971,742 genes predicted from 1,749 metagenomes and 28,941 bacterial and archaeal
86 genomes shows that (1) the extent of the unknown fraction is smaller than expected, (2) that the
87 diversity of gene clusters in the unknown fraction is higher than in the known fraction, and (3)
88 that the unknown fraction is phylogenetically more conserved and is predominantly lineage-
89 specific at the species level. Finally, we show how we can connect all the outputs produced by
90 our approach to augment the results from experimental data and add context to genes of
91 unknown function through hypothesis-driven molecular investigations.

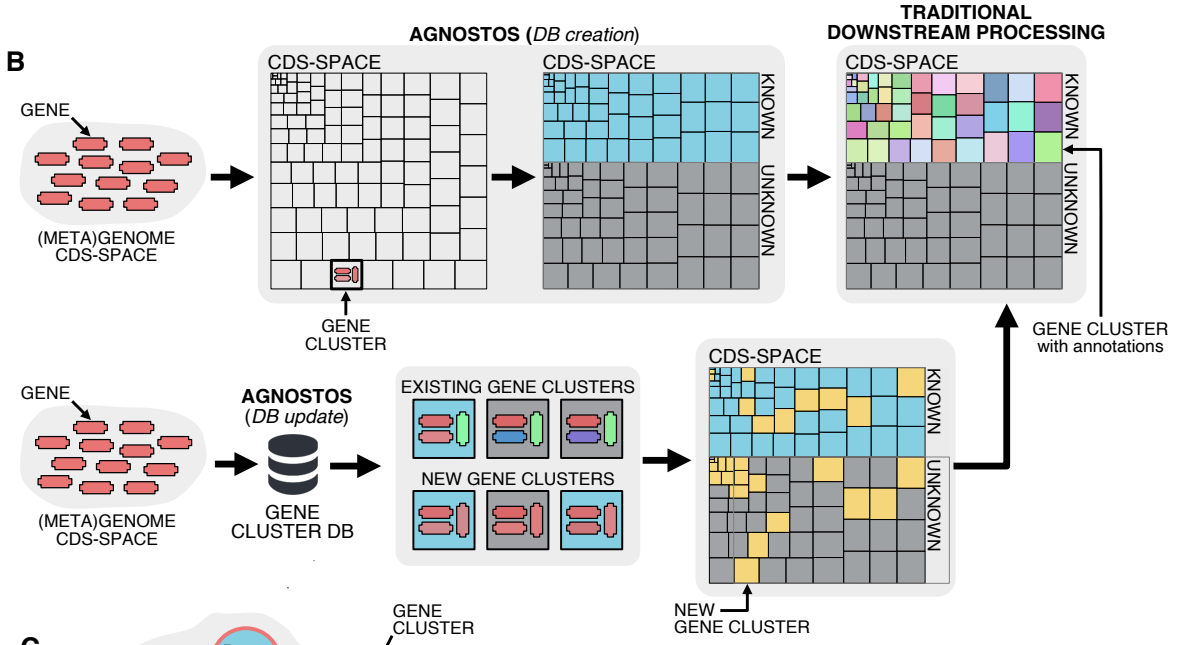
92

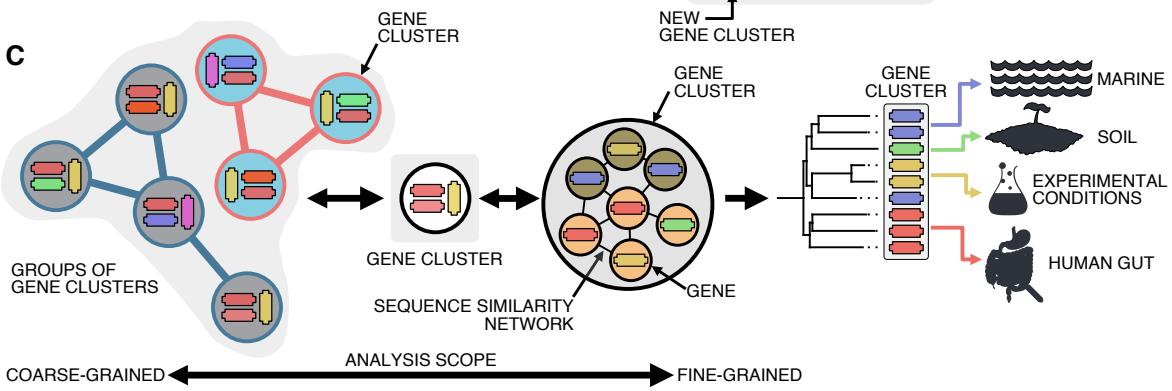
Results

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

We created the conceptual and technical foundations to unify the known and unknown CDS-space and provide a practical solution to one of the most significant ongoing challenges in microbiome analyses. 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[35]. 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). GCs produce a structured CDS-space reducing its complexity (Fig. 1B), are independent of the known and unknown fraction, are conserved across environments and organisms, and 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
GENE
(META)GENOME CDS-SPACE
AGNOSTOS (DB creation)
CDS-SPACE
GENE CLUSTER
KNOWN
UNKNOWN
TRADITIONAL DOWNSTREAM PROCESSING
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

114

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

Driven by the concepts defined in the conceptual framework, we developed AGNOSTOS, a computational workflow that infers, validates, refines, and classifies GCs in the four proposed categories (Fig. 1A; Fig. 1B; Supp. Fig 1). AGNOSTOS provides two operational modules (*DB creation* and *DB update*) to produce GCs with a highly conserved intra-homogeneous structure (Fig. 1B), both in terms of sequence similarity and domain architecture homogeneity; it exhausts any existing homology to known genes and provides a proper delimitation of the unknown CDS-space before classifying each GC in one of the four categories. In the last step, we decorate each GC with a rich collection of contextual data that we compile from different sources, or that we generate by analyzing the GC contents in different contexts (Fig. 1A). For each GC, we also offer several products that can be used for analytical purposes like improved representative sequences, consensus sequences, sequence profiles for MMseqs2[36] and HHblits[37], or the GC members as a sequence similarity network (see Online Methods). To complement the collection, we also provide a subset of what we define as *high-quality* GCs. The defining criteria are (1) the representative is a complete gene and (2) more than one-third of genes within a GC 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[38] 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).

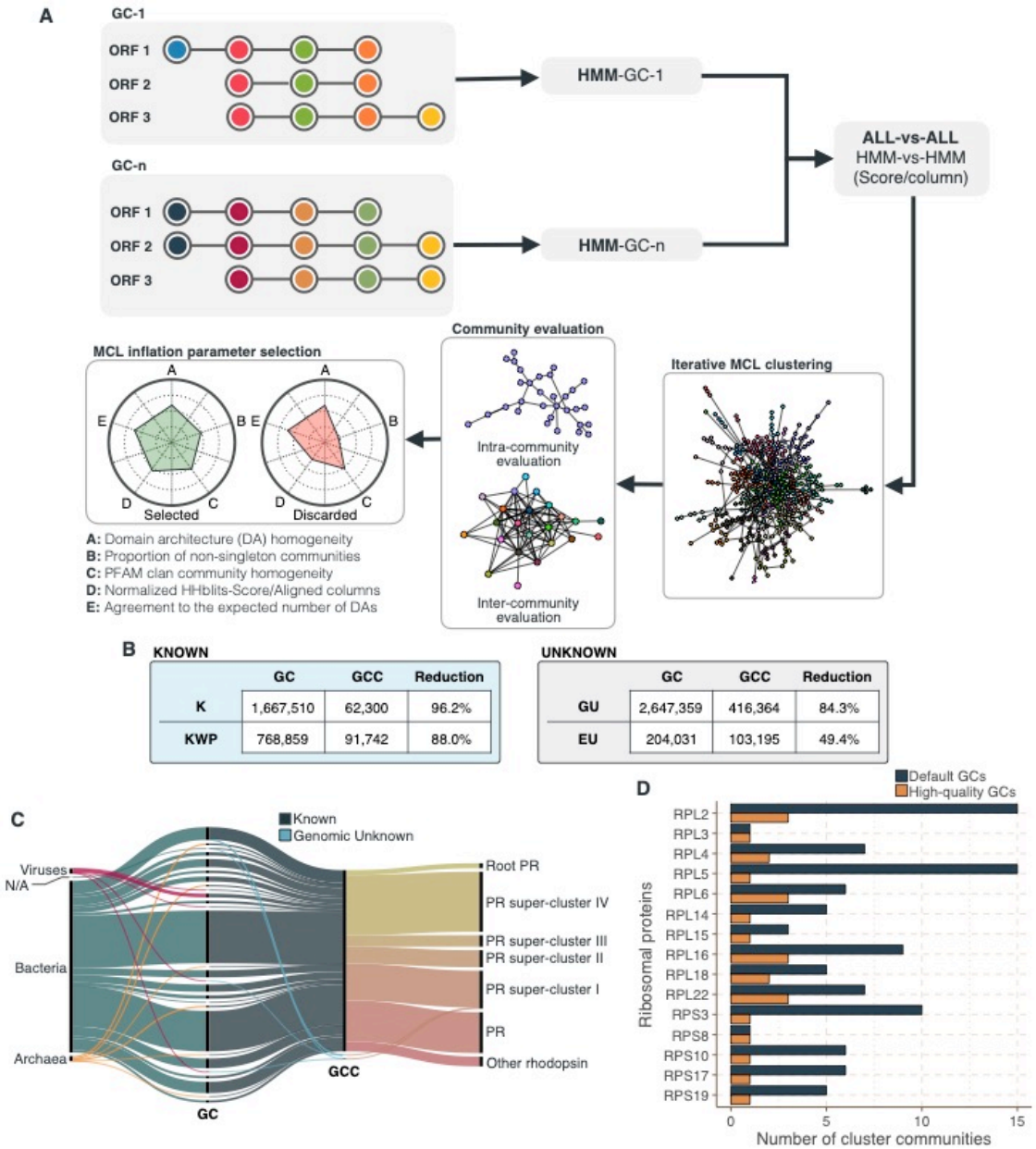
We build the link between the environmental and genomic CDS-space by expanding the final collection of GCs with the genes predicted from GTDB_r86 (Supp Fig 2A). Our environmental GCs already included 72% of the genes from GTDB_r86; 22% of them created 2,400,037 new GCs, and the rest 6% resulted in singleton 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[39].

The KWP category contains the largest proportion of incomplete ORFs (Supp. Table 3), impeding 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 disorder in the DPD database[40] and that have characteristic functions of the intrinsically disordered proteins[41] (IDP) like cellular processes and signaling as predicted by eggNOG annotations (Supp. Table 4).

As part of the workflow, each GC is complemented with a rich set of information, as shown in Fig 1A (Supp. Table 5; Supp Note 6).

181 # Beyond the twilight zone, communities of gene clusters

182 The method we developed to group GCs in gene cluster communities (GCCs) (Fig. 2A) reduced
183 the final collection of GCs by 87%, producing 673,601 GCCs (Fig. 2B; Supp. Note 7). We
184 validated the ability of our approach in capturing remote homologies between related GCs using
185 two well-known gene families present in our environmental datasets, proteorhodopsins[42] and
186 bacterial ribosomal proteins[43]. In our dataset, 64 GCs (12,184 genes) and 3 GCCs (Supp Note
187 8) contained sequences classified as proteorhodopsin (PR). One *Known* GCC contained 99% of
188 the PR annotated genes (Fig. 2C), with the only exception of 85 genes taxonomically annotated
189 as viral and assigned to the *PR Supercluster I*[44] enclosed in two GU communities (five GU gene
190 clusters; Supp Note 8). For the ribosomal proteins, the results were not so satisfactory. We
191 identified 1,843 GCs (781,579 genes) and 98 GCCs. The number of GCCs is larger compared
192 to the expected number of ribosomal proteins families (16) used for validation. When we use
193 *high-quality* GCs (Supp. Note 8), we get closer to the expected number of GCCs (Fig. 2D). With
194 this subset, we identified 26 GCCs and 145 GCs (1,687 genes). The cross-validation of our
195 method against the approach used in Méheust et al.[43] (Supp. Note 9) confirms the intrinsic
196 complexity of analyzing metagenomic data. Both approaches showed a high agreement in the
197 GCCs identified (Supp. Table 9-1). Still, our method inferred fewer GCCs for each of the
198 ribosomal protein families (Supplementary Figure 9-3), coping better with the nuisances of a
199 metagenomic setup, like incomplete genes (Supp. Table 6).
200



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

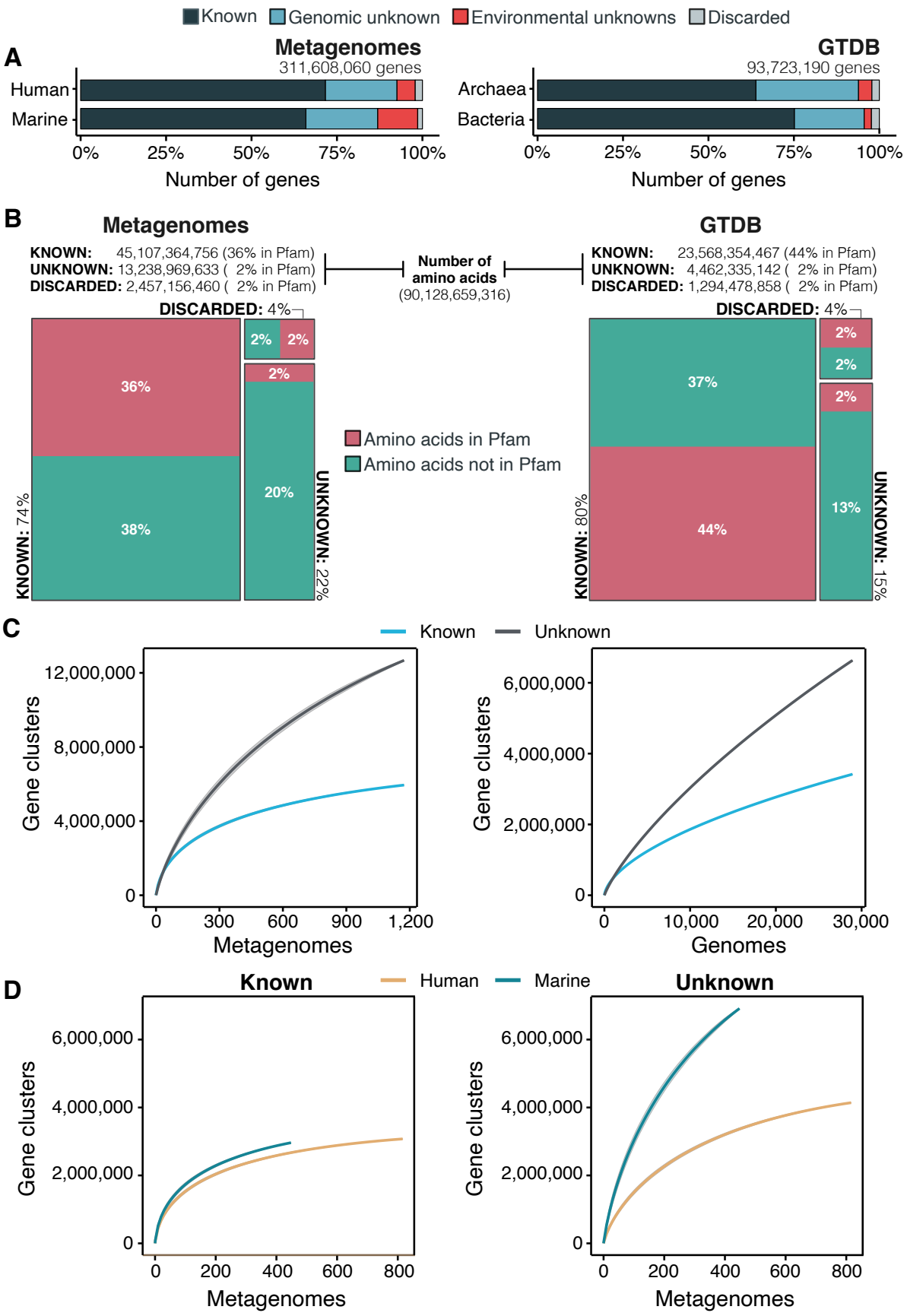
212 # A smaller but highly diverse unknown coding sequence space

213 Combining clustering and remote homology searches reduce the extent of the unknown CDS-
214 space compared to the traditional genomic and metagenomic analysis approaches (Fig. 3A).
215 Our workflow recruited as much as 71% of genes in human-related metagenomic samples and
216 65% of the genes in marine metagenomes into the known CDS-space. In both human and
217 marine microbiomes, the genomic unknown fraction showed a similar proportion of genes (21%,
218 Fig. 3A). The number of genes corresponding to EU gene clusters is higher in marine
219 metagenomes; in total, 12% of the genes are part of this GC category. We obtained a
220 comparable result when we evaluated the genes from the GTDB_r86, 75% of bacterial and 64%
221 of archaeal genes were part of the known CDS-space. Archaeal genomes contained more
222 unknowns than those from Bacteria, where 30% of the genes are classified as genomic
223 unknowns in Archaea, and only 20% in Bacteria (Fig. 3A; Supp. Table 7). We observed a similar
224 trend when we evaluated the number of amino acids belonging to the known and unknown
225 CDS-space. From the 90,128,659,316 amino acids analyzed, the majority of the amino acids in
226 metagenomes (74%) and in GTDB_r86 (80%) are in the known CDS-space (Fig. 3B; Supp.
227 Table 7). In both cases, approximately 40% of the amino acids in the known CDS-space were
228 part of a Pfam domain (Fig. 3B; Supp. Table 7). The proportion of amino acids in the unknown
229 CDS-space ranged from the 22% in metagenomes and 15% in GTDB_r86. In both cases, only
230 2% of the amino acids in the unknown CDS-space were covered by a Pfam domain.

231 To evaluate the coverage of our dataset, we calculated the accumulation rates of GCs and
232 GCCs. For the metagenomic dataset we used 1,264 metagenomes (18,566,675 GCs and
233 282,580 GCCs) and for the genomic dataset 28,941 genomes (9,586,109 GCs and 496,930
234 GCCs). The rate of accumulation of unknown GCs was three times higher than the known (2
235 times for the genomic), and both cases were far from reaching a plateau (Fig. 3C). This is not
236 the case for the GCC accumulation curves (Supp Fig 4B), where they reached a plateau. The
237 rate of accumulation is largely determined by the number of singletons, and especially
238 singletons from EUs (Supp note 11 and Supp Fig 5). While the accumulation rate of known GCs
239 between marine and human metagenomes is almost identical, there are striking differences for
240 the unknown GCs (Fig. 3D). These differences are maintained even when we remove the virus-
241 enriched samples from the marine metagenomes (Supp Fig 4A). Although the marine
242 metagenomes include a large variety of environments, from coastal to the deep sea, the known
243 space remains quite constrained.

244 Despite only including marine and human metagenomes in our database, our coverage to other
245 databases and environments is quite comprehensive, with an overall coverage of 76% (Supp.

246 Note 12). The lowest covered biomes are freshwater, soil and human non-digestive as revealed
247 by the screening of MGnify[16] (release 2018_09; 11 biomes; 843,535,6116 proteins) where we
248 assigned 74% of the MGnify proteins into one of our categories (Supplementary Fig. 6).
249


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

250

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

258 Revealing the importance of the unknown coding sequence space 259 in marine and human environments

260 Although the role of the unknown fraction in the environment is still a mystery, the large number
261 of gene counts and abundance observed underlines its inherent ecological relevance (Fig. 4A).
262 In some samples, the genomic unknown fraction can account for more than 40% of the total
263 gene abundance observed (Fig. 4A). The environmental unknown fraction is also relevant in
264 several samples, where singleton GCs are the majority (Fig. 4A). We identified two
265 metagenomes with an unusual composition in terms of environmental unknown singletons. The
266 marine metagenome corresponds to a sample from Lake Faro (OSD42), a meromictic saline
267 with a unique extreme environment where Archaea plays an important role[45]. The HMP
268 metagenome (SRS143565) corresponds to a human sample from the right cubital fossa from a
269 healthy female subject. To understand the unusual composition of this metagenome, we should
270 perform further analyses to discard potential technical artifacts like sample contamination.
271 The ratio between the unknown and known GCs revealed that the metagenomes located at the
272 upper left quadrant in Fig. 4B-C are enriched in GCs of unknown function. In human
273 metagenomes, we can distinguish between body sites, with the gastrointestinal tract, where
274 microbial communities are expected to be more diverse and complex, especially enriched with
275 genomic unknowns. The HMP metagenomes with the largest ratio of unknowns are those
276 samples identified to contain crAssphages[46,47] and HPV viruses[48] (Supp. Table 8; Supp. Fig. 7).
277 Consistently, in marine metagenomes (Fig. 4D) we can separate between size fractions, where
278 the highest ratio in genomic and environmental unknowns corresponds to the ones enriched
279 with viruses and giant viruses.
280 To complement the previous findings, we performed a large-scale analysis to investigate the GC
281 occurrence patterns in the environment. The narrow distribution of the unknown fraction (Fig.
282 4D) suggests that these GCs might provide a selective advantage and be necessary for the
283 adaptation to specific environmental conditions. But the pool of broadly distributed

environmental unknowns is the most interesting result. We identified traces of potential ubiquitous organisms left uncharacterized by traditional approaches, as more than 80% of these GCs cannot be associated with a metagenome-assembled genome (MAG) (Supp Table 9, Supp. Note 10).

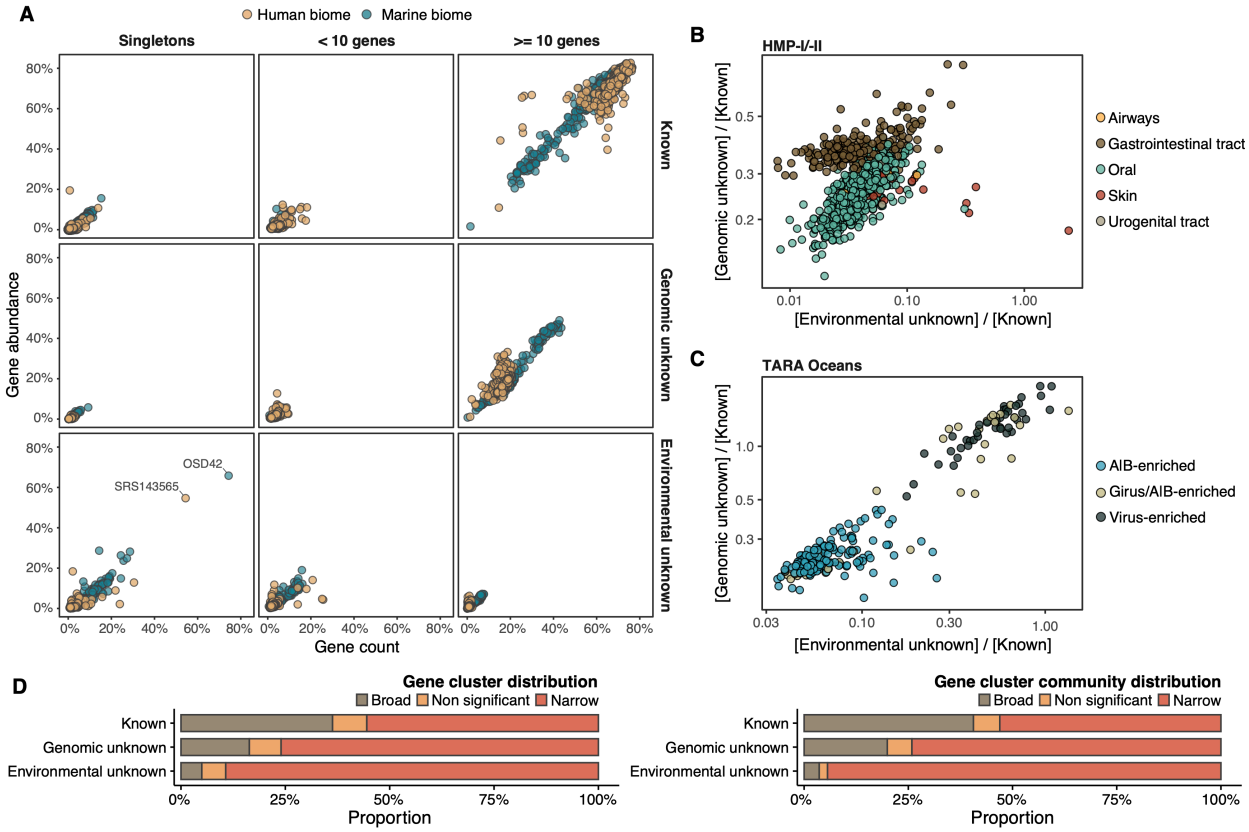


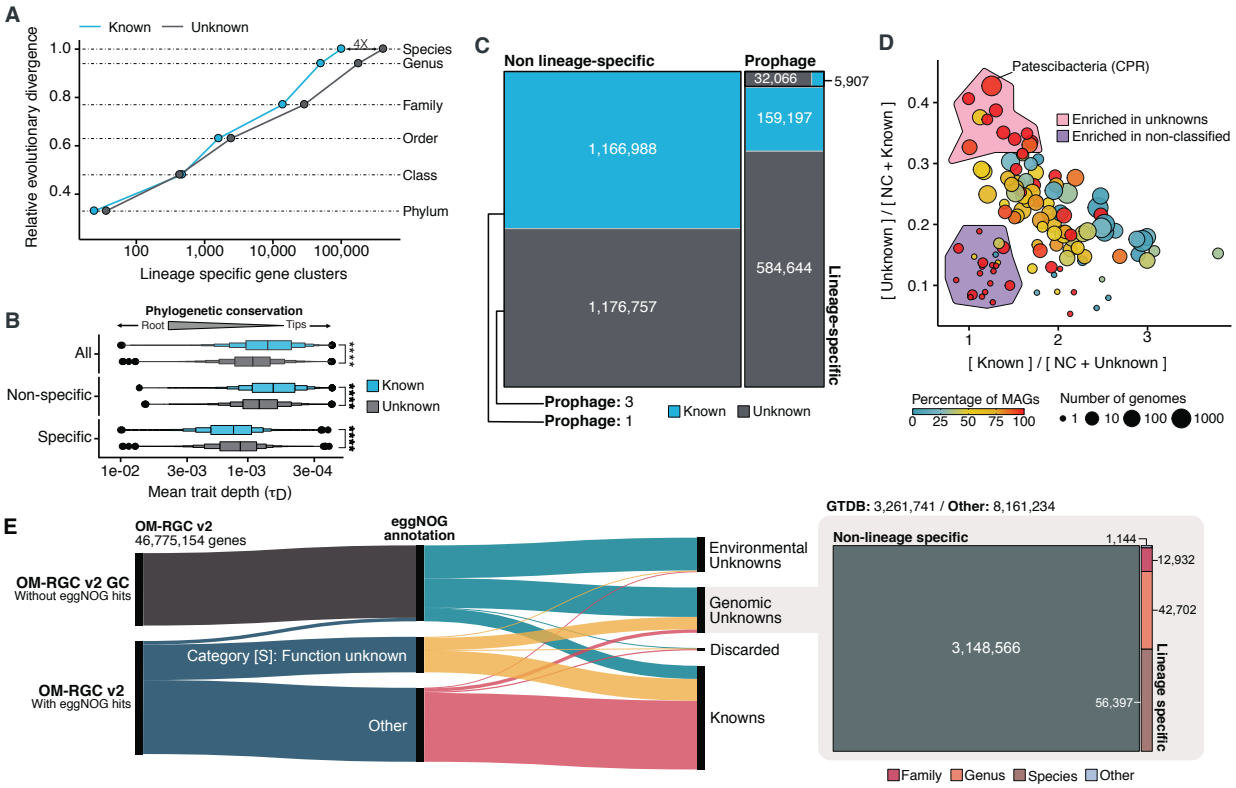
Figure 4: Distribution of the unknown coding sequence space in the human and marine metagenomes (A) Ratio between the proportion of the number of genes and their estimated abundances per cluster category and biome. Columns represented in the facet depicts three cluster categories based on the size of 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.

The genomic unknown coding sequence space is lineage-specific

We already showed that the unknown CDS-space is habitat-specific and might be relevant for organism adaptation. With the inclusion of the genomes from GTDB_r86, we have accessed a phylogenomic framework to assess how conserved and exclusive is a GC within a lineage (lineage-specifity[49]) and the clade depth where organisms share a GC (phylogenetic conservation[50]). We identified 781,814 lineage-specific GCs and 464,923 phylogenetically conserved ($P < 0.05$) GCs in Bacteria (Supp. Table 10; Supp. Note 13 for Archaea). The number of lineage-specific GCs increases with the Relative Evolutionary Distance[11] (Fig. 5A) and differences between the known and the unknown fraction start to be evident at the Family level. The unknown GCs are more phylogenetically conserved than the known (Fig. 5B, $p < 0.0001$), revealing the importance of the genome's uncharacterized fraction. However, this is not the case for the lineage-specific and phylogenetically conserved GCs, where the unknown GCs are less phylogenetically conserved (Fig. 5B), agreeing with the large number of lineage-specific GCs at Genus and Species level. To discard the possibility that the lineage-specific GCs of unknown function have a viral origin, we screened all GTDB_r86 genomes for prophages. We only found 37,163 lineage-specific GCs in prophage genomic regions, being 86% of them GCs of unknown function. After unveiling the potential relevance of the GCs of unknown function in bacterial genomes, we identified phyla in GTDB_r86 enriched with these types of clusters. A clear pattern emerged when we partitioned the phyla based on the ratio of known to unknown GCs and vice versa (Fig. 5D), the 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[9,51,52], both with the largest unknown to known ratio.

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[53] (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[54] and those without known homologs in the

336 eggNOG database[53]) that can lead to potential experimental targets at the organism level to
337 complement the metatranscriptomic approach proposed by Salazar et al[53].
338



339

340 **Figure 5:** Phylogenomic exploration of the unknown coding sequence space. (A) Distribution of the
341 lineage-specific GCs by taxonomic level. Lineage-specific unknown GCs are more abundant in the lower
342 taxonomic levels (genus, species). (B) Phylogenetic conservation of the known and unknown coding
343 sequence space in 27,372 bacterial genomes from GTDB_r86. We observe differences in the
344 conservation between the known and the unknown coding sequence space for lineage- and non-lineage
345 specific GCs (paired Wilcoxon rank-sum test; all p-values < 0.0001). (C) The majority of the lineage-
346 specific clusters are part of the unknown coding sequence space, being a small proportion found in
347 prophages present in the GTDB_r86 genomes. (D) Known and unknown coding sequence space of the
348 27,732 GTDB_r86 bacterial genomes grouped by bacterial phyla. Phyla are partitioned based on the ratio
349 of known to unknown GCs and vice versa. Phyla enriched in MAGs have higher proportions in GCs of
350 unknown function. Phyla with a high proportion of non-classified clusters (NC; discarded during the
351 validation steps) tend to contain a small number of genomes. (E) The left side of the alluvial plot shows
352 the uncharacterized (OM-RGC v2 GC) and characterized (OM-RGC v2) fraction of the gene catalog. The
353 functional annotation is based on the eggNOG annotations provided by Salazar et al.[53]. The right side of
354 the alluvial plot shows the new organization of the OM-RGC v2 coding sequence space based on the
355 approach described in this study. The treemap in the right links the metagenomic and genomic space
356 adding context to the unknown fraction of the OM-RGC v2
357

358 # Augmenting experimental data through a structured coding
359 sequence space

360 We selected one of the experimental conditions tested in Price et al.[13] to demonstrate the
361 potential of our approach to augment experimental data. We compared the fitness values in
362 plain rich medium with added Spectinomycin dihydrochloride pentahydrate to the fitness in plain
363 rich medium (LB) in *Pseudomonas fluorescens FW300-N2C3* (Fig. 6A). This antibiotic inhibits
364 protein synthesis and elongation by binding to the bacterial 30S ribosomal subunit and
365 interferes with the peptidyl tRNA translocation. We identified the gene with locus id
366 AO356_08590 that presents a strong phenotype (fitness = -3.1; t = -9.1) and has no known
367 function. This gene belongs to the genomic unknown GC GU_19737823. We can track this GC
368 into the environment and explore the occurrence in the different samples we have in our
369 database. As expected, the GC is mostly found in non-human metagenomes (Fig. 6B) as
370 *Pseudomonas* are common inhabitants of soil and water environments[55]. However, finding this
371 GC also in human-related samples is very interesting, due to the potential association of *P.*
372 *fluorescens* and human disease where Crohn's disease patients develop serum antibodies to
373 this microbe[56]. We can add another layer of information to the selected GC by looking at the
374 associated remote homologs in the GCC GU_c_21103 (Fig. 6C). We identified all the genes in
375 the GTDB_r86 genomes that belong to the GCC GU_c_21103 (Supp. Table 11) and explored
376 their genomic neighborhoods. All members from GU_c_21103 are constrained to the class
377 *Gammaproteobacteria*, and interestingly GU_19737823 is mostly exclusive to the order
378 *Pseudomonadales*. The gene order in the different genomes analyzed is highly conserved,
379 finding GU_19737823 after the *rpsF*::*rpsR* operon and before *rplI*. *rpsF* and *rpsR* encode for
380 30S ribosomal proteins, the prime target of spectinomycin. The combination of the experimental
381 evidence and the associated data inferred by our approach provides strong support to generate
382 the hypothesis that the gene AO356_08590 might be involved in the resistance to
383 spectinomycin.
384

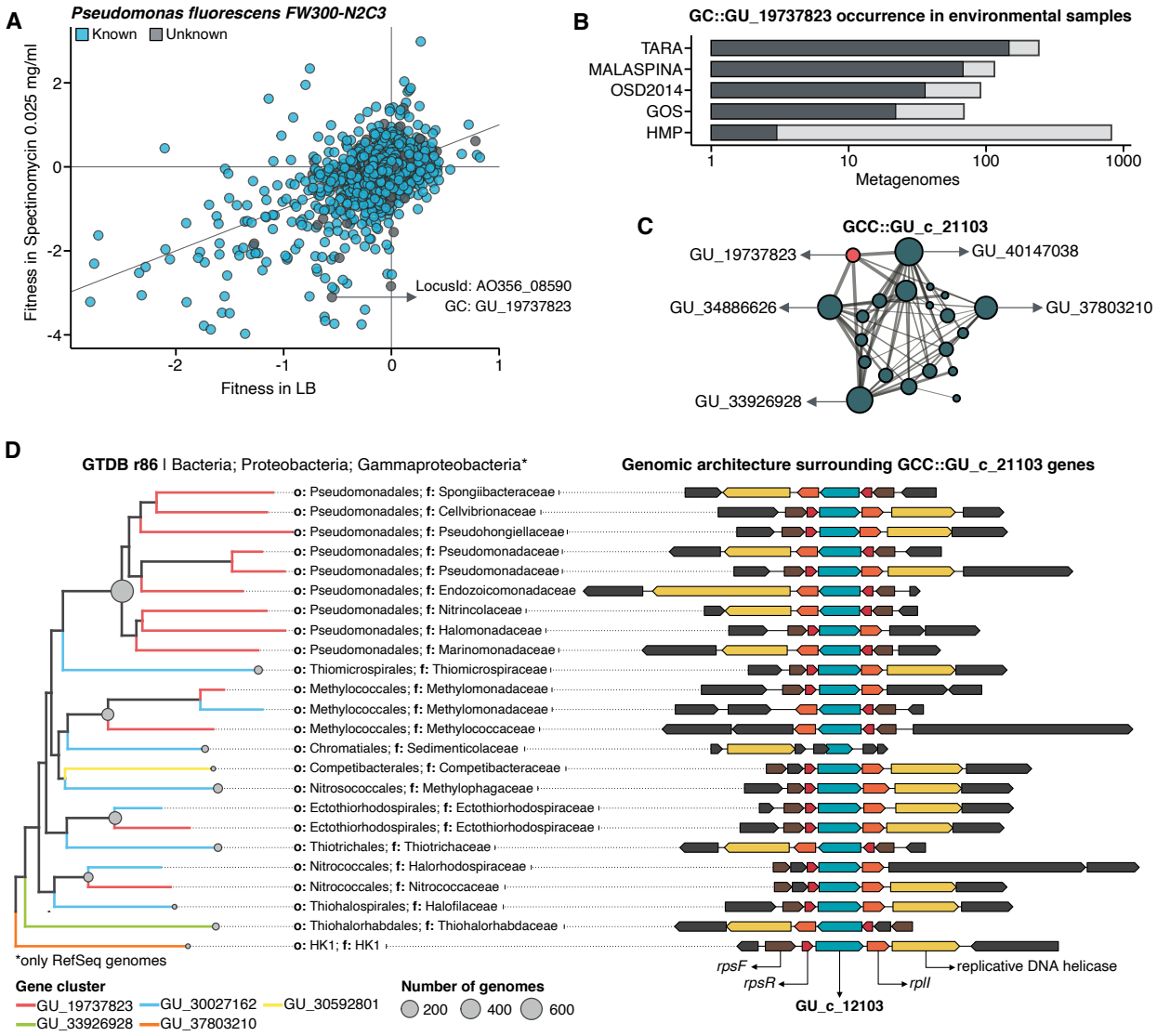


Figure 6: Augmenting experimental data with GCs of unknown function. (A) We used the fitness values from the experiments from Price et al.[13] to identify genes of unknown function that are important for fitness under certain experimental conditions. The selected gene belongs to the genomic unknown GC 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 genomes. Branch colors correspond to the different GCs found in GU_c_21103. Bubble plot depicts the number of genomes with a gene that belongs to GU_c_21103.

Discussion

We present a new conceptual framework and computational workflow to unify the known and unknown CDS-space in microbial analyses. Using this framework, we performed an in-depth exploration of the microbial unknown CDS-space. We demonstrated that we could link the unknown fraction of metagenomic studies to specific genomes and provide a powerful tool for hypothesis generation. During the last years, the microbiome community has established a standard operating procedure[17] for analyzing metagenomes that we can briefly summarize into (1) assembly, (2) gene prediction, (3) gene catalog inference, (4) binning, and (5) characterization. Thanks to recent computational developments[36,57], we envisioned an alternative to this workflow where we can maximize the information used when analyzing genomic and metagenomic data. In addition, we provide a mechanism to reconcile top-down and bottom-up approaches, thanks to the well-structured CDS-space proposed by our framework. AGNOSTOS can create environmental- and organism-specific variations of a seed GC database. Then, it integrates the predicted genes from new genomes and metagenomes and dynamically creates and classifies new GCs with those genes not integrated during the initial step (Fig. 1B). Afterward, the potential functions of the known GCs can be carefully characterized by incorporating them into the traditional workflows.

One of the most appealing characteristics of our approach is that the GCs provide unified groups of homologous genes across environments and organisms indifferently if they belong to the known or unknown CDS-space, and we can contextualize the unknown fraction using this genomic and environmental information. Our combination of partitioning and contextualization features a smaller unknown CDS-space than we expected. On average, for our genomic and metagenomic data, only 30% of the genes fall in the unknown fraction. One hypothesis to reconcile this surprising finding is that until recently, the methodologies to identify remotely homologous sequences in large datasets were computationally prohibitive. New methods[36,37], like the ones used in AGNOSTOS, are enabling large scale distant homology searches. Still, one has to apply conservative measures to control the trade-off between specificity and sensitivity to avoid overclassification.

We found that the majority of the coding sequence space at gene and amino acid is known, both in genomes and metagenomes. However, it presents a high diversity as shown in the GC accumulation curves highlighting the vast remaining untapped microbial fraction and its potential importance for niche adaptation owing to its narrow ecological distribution. In a genomic context, the unknown fraction is predominantly species' lineage-specific and phylogenetically more

435 conserved than the known fraction, supporting the signal observed in the environmental data
436 and emphasizing that the unknown fraction should not be ignored. We also ruled out the effect
437 of prophages, strengthening the hypothesis that the lineage-specific GCs of unknown function
438 might be associated with the mechanisms of microbial diversification and niche adaptation as a
439 result of the constant diversification of gene families and the survival of new gene lineages[58,59].
440 It is worth noting that we need to explore further the unknown fraction to identify new potential
441 protein domains. Only 10% of the unknown CDS-space amino acids are part of a Pfam domain
442 (DUF and others); this contrasts with the numbers observed in the known CDS-space, where
443 Pfam domains include 50% of the amino acids.
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 their unusual biology[9]. 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 advancement of knowledge in the CPR research
451 efforts.
452 Our effort to tackle the unknown provides a pathway to unlock a large pool of likely relevant data
453 that remains untapped to analysis and discovery. With the identification of a potential target
454 gene of unknown function for antibiotic resistance, we demonstrate the value of our approach
455 and how it can boost insights from model organism experiments. But severe challenges remain,
456 such as the dependence on the quality of the assemblies and their gene predictions, as shown
457 by the analysis of the ribosomal protein GCCs where many of the recovered genes are
458 incomplete. While sequence assembly has been an active area of research[60], this has not been
459 the case for gene prediction methods[60], which are becoming outdated[61] and cannot cope with
460 the current amount of data. Alternatives like protein-level assembly[62] combined with the
461 exploration of the assembly graphs' neighborhoods[63] become very attractive for our purposes.
462 In any case, we still face the challenge of discriminating between real and artifactual
463 singletons[64]. At the moment, there are no methods available to provide a plausible solution and,
464 at the same time, being scalable. We devise a potential solution in the recent developments in
465 unsupervised deep learning methods where they use large corpora of proteins to define a
466 language model *embedding* for protein sequences[65]. These models could be applied to predict
467 *embeddings* in singletons, which could be clustered or used to determine their coding potential.
468 Another issue is that we might be creating more GCs than expected. We follow a conservative

469 approach to avoid mixing multidomain proteins in GCs owing to the fragmented nature of the
470 metagenome assemblies that could result in the split of a GC. However, not only splitting can be
471 a problem, but also lumping unrelated genes or GCs owing to the use of remote homologies.
472 Although the inference of GCCs is using very sensitive methods to compare profile HMMs, low
473 sequence diversity in GCs can limit its effectiveness. Our approach is affected by the presence
474 and propagation of contamination in reference databases, a significant problem in 'omics[66,67]. In
475 our case, we only use Pfam as a source for annotation owing to its high-quality and manual
476 curation process. The categorization process of our GCs depends on the information from other
477 databases, and to minimize the potential impact of contamination, we apply methods that weight
478 the annotations of the identified homologs to discriminate if a GC belongs to the known or
479 unknown CDS-space. We foresee the integration of our approach to assist in the manual
480 curation process and increase the quality of the recovered MAGs[68].
481 The work presented here should incentivize the scientific community to build a collective effort
482 to define the different levels of unknown[69] where clear guidelines and protocols should be
483 established. Our work proves that the integration of the unknown fraction is possible and aims to
484 provide a new brighter future for microbiome analyses.
485

486 Material and methods

487 Genomic and metagenomic dataset

488 We used a set of 583 marine metagenomes from four of the major metagenomic surveys of the
489 ocean microbiome: Tara Oceans expedition (TARA)[2], Malaspina expedition[70], Ocean Sampling
490 Day (OSD)[3], and Global Ocean Sampling Expedition (GOS)[71]. We complemented this set with
491 1,246 metagenomes obtained from the Human Microbiome Project (HMP) phase I and II[72]. We
492 used the assemblies provided by TARA, Malaspina, OSD and HMP projects and the long
493 Sanger reads from GOS[73]. A total of 156M (156,422,969) contigs and 12.8M long-reads were
494 collected (Supp. Table 6).
495 For the genomic dataset, we used the 28,941 prokaryotic genomes (27,372 bacterial and 1,569
496 archaeal) from the Genome Taxonomy Database[11] (GTDB) Release 03-RS86 (19th August
497 2018).

498 ## Computational workflow development

499 We implemented a computation workflow based on Snakemake[74] for the easy processing of
500 large datasets in a reproducible manner. The workflow provides three different strategies to
501 analyze the data. The module *DB-creation* creates the gene cluster database, validates and
502 partitions the gene clusters (GCs) in the main functional categories. The module *DB-update*
503 allows the integration of new sequences (either at the contig or predicted gene level) in the
504 existing gene cluster database. In addition, the workflow has a *profile-search* function to quickly
505 screen samples using the gene cluster PSSM profiles in the database.

506 ## Metagenomic and genomic gene prediction

507 We used Prodigal (v2.6.3)[75] in metagenomic mode to predict the genes from the metagenomic
508 dataset. For the genomic dataset, we used the gene predictions provided by Annotree[76], since
509 they were obtained, consistently, with Prodigal v2.6.3. We identified potential spurious genes
510 using the *AntiFam* database[77]. Furthermore, we screened for '*shadow' genes* using the
511 procedure described in Yooseph et al.[78].

512 ## PFAM annotation

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

518 ## Determination of the gene clusters

519 We clustered the metagenomic predicted genes using the cascaded-clustering workflow of the
520 MMseqs2 software[57] ("*--cov-mode 0 -c 0.8 --min-seq-id 0.3"*). We discarded from downstream
521 analyses the singletons and clusters with a size below a threshold identified after applying a
522 broken-stick model[81]. We integrated the genomic data into the metagenomic cluster database
523 using the "DB-update" module of the workflow. This module uses the *clusterupdate* module of
524 MMseqs2[36], with the same parameters used for the metagenomic clustering.

525 # Quality-screening of gene clusters

526 We examined the GCs to ensure their high intra-cluster homogeneity. We applied two
527 methodologies to validate their cluster sequence composition and functional annotation
528 homogeneity. We identified non-homologous sequences inside each cluster combining the
529 identification of a new cluster representative sequence via a sequence similarity network (SSN)
530 analysis, and the investigation of intra-cluster multiple sequence alignments (MSAs), given the
531 new representative. Initially, we generated an SSN for each cluster, using the semi-global
532 alignment methods implemented in *PARASAIL*[82] (version 2.1.5). We trimmed the SSN using a
533 custom algorithm[83,84] that removes edges while maintaining the network structural integrity and
534 obtaining the smallest connected graph formed by a single component. Finally, the new cluster
535 representative was identified as the most central node of the trimmed SSN by the eigenvector
536 centrality algorithm, as implemented in igraph[85]. After this step, we built a multiple sequence
537 alignment for each cluster using *FAMSA*[86] (version 1.1). Then, we screened each cluster-MSA
538 for non-homologous sequences to the new cluster representative. Owing to computational
539 limitations, we used two different approaches to evaluate the cluster-MSAs. We used *LEON-*
540 *BIS*[87] for the clusters with a size ranging from 10 to 1,000 genes and OD-SEQ[88] for the clusters
541 with more than 1,000 genes. In the end, we applied a broken-stick model[81] to determine the
542 threshold to discard a cluster.

543 The predicted genes can have multi-domain annotations in different orders, therefore to validate
544 the consistency of intra-cluster Pfam annotations, we applied a combination of w-shingling[89] and
545 Jaccard similarity. We used w-shingling (k-shingle = 2) to group consecutive domain
546 annotations as a single object. We measured the homogeneity of the *shingle sets* (sets of
547 domains) between genes using the Jaccard similarity and reported the median similarity value
548 for each cluster. Moreover, we took into consideration the Clan membership of the Pfam
549 domains and that a gene might contain N-, C- and M-terminal domains for the functional
550 homogeneity validation. We discarded clusters with a median similarity < 1.

551 After the validation, we refined the gene cluster database removing the clusters identified to be
552 discarded and the clusters containing ≥ 30% *shadow genes*. Lastly, we removed the single
553 shadow, spurious and non-homologous genes from the remaining clusters (Supplementary Note
554 2).

555 # Remote homology classification of gene clusters

556 To partition the validated GCs into the four main categories, we processed the set of GCs
557 containing Pfam annotated genes and the set of not annotated GCs separately. For the
558 annotated GCs, we inferred a consensus protein domain architecture (DA) (an ordered
559 combination of protein domains) for each annotated gene cluster. To identify each gene cluster
560 consensus DA, we created directed acyclic graphs connecting the Pfam domains based on their
561 topological order on the genes using *igraph*[85]. We collapsed the repetitions of the same domain.
562 Then we used the gene completeness as a positive-weighting value for the selection of the
563 cluster consensus DA. Within this step, we divided the GCs into "Knowns" (Known) if annotated
564 to at least one Pfam domains of known function (DKFs) and "Genomic unknowns" (GU) if
565 annotated entirely to Pfam domains of unknown function (DUFs).
566 We aligned the sequences of the non-annotated GCs with FAMSA[86] and obtained cluster
567 consensus sequences with the *hhconsensus* program from *HH-SUITE*[37]. We used the cluster
568 consensus sequences to perform a nested search against the UniRef90 database (release
569 2017_11)[90] and NCBI *nr* database (release 2017_12)[91] to retrieve non-Pfam annotations with
570 *MMSeqs2*[36] (“*-e 1e-05 --cov-mode 2 -c 0.6”*). We kept the hits within 60% of the Log(best-e-
571 value) and searched the annotations for any of the terms commonly used to define proteins of
572 unknown function (Supp. Table 12). We used a quorum majority voting approach to decide if a
573 gene cluster would be classified as *Genomic Unknown* or *Known without Pfams* based on the
574 annotations retrieved. We searched the consensus sequences without any homologs in the
575 UniRef90 database against NCBI *nr*. We applied the same approach and criteria described for
576 the first search. Ultimately, we classified as *Environmental Unknown* those GCs whose
577 consensus sequences did not align with any of the NCBI *nr* entries.
578 In addition, we developed some conservative measures to control the trade-off between
579 specificity and sensitivity for the remote homology searches such as (1) a modification of the
580 algorithm described in Hingamp et al.[92] to get a confident group of homologs to determine if a
581 query protein is known or unknown by a quorum majority voting approach (Supp Note 3); (2)
582 strict parameters in terms of iterations, bidirectional coverage and probability thresholds for the
583 HHblits alignments to minimize the inclusion of non-homologous sequences; and (3) avoid
584 providing annotations for our gene clusters, as we believe that annotation should be a careful
585 process done on a smaller scale and with experimental context.

Gene cluster remote homology refinement

We refined the *Environmental Unknown* GCs to ensure the lack of any characterization by searching for remote homologies in the Uniclust database (release 30_2017_10) using the HMM/HMM alignment method *HHblits*[93]. We created the HMM profiles with the *hhmake* program from the *HH-SUITE*[37]. We only accepted those hits with an *HHblits-probability* ≥ 90% and we re-classified them following the same majority vote approach as previously described. The clusters with no hits remained as the refined set of EUs. We applied a similar refinement approach to the KWP clusters to identify GCs with remote homologies to Pfam protein domains. The KWP HMM profiles were searched against the Pfam *HH-SUITE* database (version 31), using *HHblits*. We accepted hits with a probability ≥ 90% and a target coverage > 60% and removed overlapping domains as described earlier. We moved the KWP with remote homologies to known Pfams to the Known set, and those showing remote homologies to Pfam DUFs to the GUs. The clusters with no hits remained as the refined set of KWP.

Gene cluster characterization

To retrieve the taxonomic composition of our clusters we applied the *MMseqs2 taxonomy* program (version: b43de8b7559a3b45c8e5e9e02cb3023dd339231a), which allows computing the lowest common ancestor through the implementation of the 2bLCA protocol [92]. We searched all cluster genes against UniProtKB (release of January 2018) [94] using the following parameters "*-e 1e-05 --cov-mode 0 -c 0.6"*. We parsed the results to keep only the hits within 60% of the log10(best-e-value). To retrieve the taxonomic lineages, we used the R package *CHNOSZ*[95]. We measured the intra-cluster taxonomic admixture by applying the *entropy.empirical()* function from the *entropy* R package[96]. This function estimates the Shannon entropy based on the different taxonomic annotation frequencies. For each cluster, we also retrieved the cluster consensus taxonomic annotation, which we defined as the taxonomic annotation of the majority of the genes in the cluster.

In addition to the taxonomy, we evaluated the clusters' level of darkness and disorder using the Dark Proteome Database (DPD)[40] as reference. We searched the cluster genes against the DPD, applying the MMseqs2 search program[36] with "*-e 1e-20 --cov-mode 0 -c 0.6*". For each cluster, we then retrieved the mean and the median level of darkness, based on the gene DPD annotations.

616 ## High-quality clusters

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

621 ## A set of non-redundant domain architectures

622 We estimated the number of potential domain architectures present in the *Known* GCs taking
623 into account the large proportion of fragmented genes in the metagenomic dataset and that
624 could inflate the number of potential domain architectures. To identify fragments of larger
625 domain architecture, we took into account their topological order in the genes. To reduce the
626 number of comparisons, we calculated the pairwise string cosine distance (q-gram = 3) between
627 domain architectures and discarded the pairs that were too divergent (cosine distance ≥ 0.9).
628 We collapsed a fragmented domain architecture to the larger one when it contained less than
629 75% of complete genes.

630 ## Inference of gene cluster communities

631 We aggregated distant homologous GCs into GCCs. The community inference approach
632 combined an all-vs-all HMM gene cluster comparison with Markov Cluster Algorithm (MCL)[97]
633 community identification. We started performing the inference on the Known GCs to use the
634 Pfam DAs as constraints. We aligned the gene cluster HMMs using HHblits[93] (-n 2 -Z 10000000
635 -B 10000000 -e 1) and we built a homology graph using the cluster pairs with probability ≥ 50%
636 and bidirectional coverage > 60%. We used the ratio between HHblits-bitscore and aligned-
637 columns as the edge weights (Supp. Note 9). We used MCL[97] (v. 12-068) to identify the
638 communities present in the graph. We developed an iterative method to determine the optimal
639 MCL inflation parameter that tries to maximize the relationship of five intra-/inter-community
640 properties: (1) the proportion of MCL communities with one single DA, based on the consensus
641 DAs of the cluster members; (2) the ratio of MCL communities with more than one cluster; (3)
642 the proportion of MCL communities with a PFAM clan entropy equal to 0; (4) the intra-
643 community HHblits-score/Aligned-columns score (normalized by the maximum value); and (5)
644 the number of MCL communities, which should, in the end, reflect the number of non-redundant
645 DAs. We iterated through values ranging from 1.2 to 3.0, with incremental steps of 0.1. During
646 the inference process, some of the GCs became orphans in the graph. We applied a three-step

647 approach to assigning a community membership to these GCs. First, we used less stringent
648 conditions (probability ≥ 50% and coverage >= 40%) to find homologs in the already existing
649 GCCs. Then, we ran a second iteration to find secondary relationships between the newly
650 assigned GCs and the missing ones. Lastly, we created new communities with the remaining
651 GCs. We repeated the whole process with the other categories (KWP, GU and EU), applying
652 the optimal inflation value found for the Known (2.2 for metagenomic and 2.5 for genomic data).

653 Gene cluster communities validation

654 We tested the biological significance of the GCCs using the phylogeny of proteorhodopsin[44]
655 (PR). We used the proteorhodopsin HMM profiles[42] to screen the marine metagenomic datasets
656 using *hmmsearch* (version 3.1b2)[79]. We kept the hits with a coverage > 0.4 and e-value <= 1e-
657 5. We removed identical duplicates from the sequences assigned to PR with CD-HIT[98] (v4.6)
658 and cleaned from sequences with less than 100 amino acids. To place the identified PR
659 sequences into the MicRhode[44] PR tree first, we optimized the initial tree parameters and
660 branch lengths with RAxML (v8.2.12)[99]. We used PaPaRA (v2.5)[100] to incrementally align the
661 query PR sequences against the MicRhode PR reference alignment and *pplacer*[101]
662 (v1.1.alpha19-0-g807f6f3) to place the sequences into the tree. Finally, we assigned the query
663 PR sequences to the MicRhode PR Superclusters based on the phylogenetic placement. We
664 further investigated the GCs annotated as viral (196 genes, 14 GC) comparing them to the six
665 newly discovered viral PRs[102] using Parasail[82] (-a sg_stats_scan_sse2_128_16 -t 8 -c 1 -x). As
666 an additional evaluation, we investigated the distributions of standard GCCs and HQ GCCs
667 within ribosomal protein families. We obtained the ribosomal proteins used for the analysis
668 combining the set of 16 ribosomal proteins from Méheust et al.[43] and those contained in the
669 collection of bacterial single-copy genes of Anvi'o[103]. Also, for the ribosomal proteins, we
670 compared the outcome of our method to the one proposed by Méheust et al.[43] (Supp. Note 9).

671 Metagenomic sample selection for downstream analyses

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

675 ## Gene cluster abundance profiles in genomes and metagenomes

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

686 ## Rate of genomic and metagenomic gene clusters accumulation

687 We calculated the cumulative number of known and unknown GCs as a function of the number
688 of metagenomes and genomes. For each metagenome count, we generated 1000 random sets,
689 and we calculated the number of GCs and GCCs recovered. For this analysis, we used 1,246
690 HMP metagenomes and 358 marine metagenomes (242 from TARA and 116 from Malaspina).
691 We repeated the same procedure for the genomic dataset. We removed the singletons from the
692 metagenomic dataset with an abundance smaller than the mode abundance of the singletons
693 that got reclassified as good-quality clusters after integrating the GTDB data to minimize the
694 impact of potential spurious singletons. To complement those analyses, we evaluated the
695 coverage of our dataset by searching seven different state-of-the-art databases against our set
696 of metagenomic GC HMM profiles (Supp. Note 12).
697

698 ## Occurrence of gene clusters in the environment

699 We used 1,264 metagenomes from the TARA Oceans, MALASPINA Expedition, OSD2014 and
700 HMP-I/II to explore the properties of the unknown CDS-space in the environment. We applied
701 the Levins Niche Breadth (NB) index[106] to investigate the GCs and GCCs environmental
702 distributions. We removed the GCs and cluster communities with a mean relative abundance <
703 1e-5. We followed a divide-and-conquer strategy to avoid the computational burden of
704 generating the null-models to test the significance of the distributions owing to the large number
705 of metagenomes and GCs. First, we grouped similar samples based on the gene cluster content

706 using the Bray-Curtis dissimilarity[107] in combination with the *Dynamic Tree Cut*[108] R package.
707 We created 100 random datasets picking up one random sample from each group. For each of
708 the 100 random datasets, we created 100 random abundance matrices using the *nullmodel*
709 function of the *quasiswap* count method[109]. Then we calculated the *observed* NB and obtained
710 the 2.5% and 97.5% quantiles based on the randomized sets. We compared the observed and
711 quantile values for each gene cluster and defined it to have a *Narrow distribution* when the
712 *observed* was smaller than the 2.5% quantile and to have a *Broad distribution* when it was
713 larger than the 97.5% quantile. Otherwise, we classified the cluster as *Non-significant*[110]. We
714 used a majority voting approach to get a consensus distribution classification based on the ten
715 random datasets.

716 Identification of prophages in genomic sequences

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

725 Lineage-specific gene clusters

726 We used the F1-score developed for AnnoTree[76] to identify the lineage-specific GCs and to
727 which rank they are specific. Following similar criteria to the ones used in Mendler et al.[76], we
728 considered a gene cluster to be lineage-specific if it is present in less than half of all genomes
729 and at least 2 with F1-score > 0.95.

730 Phylogenetic conservation of gene clusters

731 We calculated the phylogenetic conservation (τD) of each gene cluster using the
732 *consenTRAIT*[50] function implemented in the R package *castor*[50]. We used a paired Wilcoxon
733 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)[53] 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[54] category S within the context of our database.

Augmenting experimental data

We searched the 37,684 genes of unknown function associated with mutant phenotypes from Price et al.[13] 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 the different methods and results of this manuscript are available at https://dark.metagenomics.eu. The workflow AGNOSTOS is available at https://github.com/functional-dark-side/agnostos-wf, and its database can be downloaded from https://doi.org/10.6084/m9.figshare.12459056. Supplementary information is available at https://doi.org/10.6084/m9.figshare.12588263.

Acknowledgments

The authors thankfully acknowledge the computer resources at MareNostrum and the technical support provided by Barcelona Supercomputing Center (RES-AECT-2014-2-0085), the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B),

the University of Oxford Advanced Research Computing (http://dx.doi.org/10.5281/zenodo.22558) and the MARBITS bioinformatics core at ICM-CSIC. CV was supported by the Max Planck Society. AFG received funding from the European Union's Horizon 2020 research and innovation program Blue Growth: Unlocking the potential of Seas and Oceans under grant agreement no. 634486 (project acronym INMARE). AM was supported by the Biotechnology and Biological Sciences Research Council [BB/M011755/1, BB/R015228/1] and RDF by the European Molecular Biology Laboratory core funds. EOC was supported by project INTERACTOMA RTI2018-101205-B-I00 from the Spanish Agency of Science MICIU/AEI. SGA and PS received additional funding by the project MAGGY (CTM2017-87736-R) from the Spanish Ministry of Economy and Competitiveness. The Malaspina 2010 Expedition was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) through the Consolider-Ingenio program (ref. CSD2008-00077). 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. RF 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. Eloe-Fadrosh, E. A. *et al.* Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. *Nat. Commun.* **7**, 10476 (2016).
9. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
10. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
11. 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).
12. Bernard, G., Pathmanathan, J. S., Lannes, R., Lopez, P. & Bapteste, E. Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and

819 Empirically Sketch a Logic of Scientific Discovery. *Genome Biol. Evol.* **10**, 707–715 (2018).
820 13. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function.
821 *Nature* **557**, 503–509 (2018).
822 14. Carradec, Q. *et al.* A global ocean atlas of eukaryotic genes. *Nat. Commun.* **9**, 373 (2018).
823 15. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut
824 microbiome. *Nat. Biotechnol.* (2020) doi:10.1038/s41587-020-0603-3.
825 16. Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*
826 **48**, D570–D578 (2020).
827 17. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun
828 metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
829 18. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and
830 metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
831 19. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology
832 Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
833 20. Chen, I.-M. A. *et al.* IMG/M v.5.0: an integrated data management and comparative
834 analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–
835 D677 (2019).
836 21. Hanson, A. D., Pribat, A., Waller, J. C. & Crécy-Lagard, V. de. ‘Unknown’proteins and
837 ‘orphan’enzymes: the missing half of the engineering parts list--and how to find it. *Biochem.*
838 *J* **425**, 1–11 (2010).
839 22. Arnold, F. H. Design by Directed Evolution. *Acc. Chem. Res.* **31**, 125–131 (1998).
840 23. Brandenberg, O. F., Fasan, R. & Arnold, F. H. Exploiting and engineering hemoproteins for
841 abiological carbene and nitrene transfer reactions. *Curr. Opin. Biotechnol.* **47**, 102–111
842 (2017).
843 24. Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem. Int. Ed*
844 *Engl.* **57**, 4143–4148 (2018).

845 25. Jaroszewski, L. *et al.* Exploration of uncharted regions of the protein universe. *PLoS Biol.* **7**,
846 (2009).
847 26. Buttigieg, L. P. *et al.* Ecogenomic Perspectives on Domains of Unknown Function:
848 Correlation-Based Exploration of Marine Metagenomes. *PLoS One* **8**, (2013).
849 27. Yooseph, S. *et al.* The Sorcerer II global ocean sampling expedition: Expanding the
850 universe of protein families. *PLoS Biol.* **5**, 0432–0466 (2007).
851 28. Wyman, S. K., Avila-Herrera, A., Nayfach, S. & Pollard, K. S. A most wanted list of
852 conserved microbial protein families with no known domains. *PLoS One* **13**, e0205749
853 (2018).
854 29. Brum, J. R. *et al.* Illuminating structural proteins in viral “dark matter” with metaproteomics.
855 *Proc. Natl. Acad. Sci. U. S. A.* **113**, 2436–2441 (2016).
856 30. Bateman, A., Coggill, P. & Finn, D. R. DUFs: Families in search of function. *Acta*
857 *Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **66**, 1148–1152 (2010).
858 31. Lobb, B., Kurtz, D. A., Moreno-Hagelsieb, G. & Doxey, A. C. Remote homology and the
859 functions of metagenomic dark matter. *Front. Genet.* **6**, 1–12 (2015).
860 32. Bitard-Feildel, T. & Callebaut, I. Exploring the dark foldable proteome by considering
861 hydrophobic amino acids topology. *Sci. Rep.* **7**, 41425 (2017).
862 33. Bileschi, M. L. *et al.* Using Deep Learning to Annotate the Protein Universe. *bioRxiv* 626507
863 (2019) doi:10.1101/626507.
864 34. Liu, X. L. Deep Recurrent Neural Network for Protein Function Prediction from Sequence.
865 *bioRxiv* 103994 (2017) doi:10.1101/103994.
866 35. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng. Des. Sel.* **12**, 85–94
867 (1999).
868 36. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the
869 analysis of massive data sets. *Nat. Biotechnol.* **advance on**, (2017).
870 37. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein

871 annotation. *BMC Bioinformatics* **20**, 473 (2019).

872 38. Skewes-Cox, P., Sharpton, T. J., Pollard, K. S. & DeRisi, J. L. Profile hidden Markov
873 models for the detection of viruses within metagenomic sequence data. *PLoS One* **9**,
874 e105067 (2014).

875 39. Sberro, H. *et al.* Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small,
876 Novel Genes. *Cell* **178**, 1245-1259.e14 (2019).

877 40. Perdigão, N., Rosa, A. C. & O'Donoghue, S. I. The Dark Proteome Database. *BioData Min.*
878 **10**, 1–11 (2017).

879 41. Habchi, J., Tompa, P., Longhi, S. & Uversky, V. N. Introducing protein intrinsic disorder.
880 *Chem. Rev.* **114**, 6561–6588 (2014).

881 42. Olson, D. K., Yoshizawa, S., Boeuf, D., Iwasaki, W. & DeLong, E. F. Proteorhodopsin
882 variability and distribution in the North Pacific Subtropical Gyre. *ISME J.* **12**, 1047–1060
883 (2018).

884 43. Méheust, R., Burstein, D., Castelle, C. J. & Banfield, J. F. The distinction of CPR bacteria
885 from other bacteria based on protein family content. *Nat. Commun.* **10**, 4173 (2019).

886 44. Boeuf, D., Audic, S., Brillet-Guéguen, L., Caron, C. & Jeanthon, C. MicRhoDE: a curated
887 database for the analysis of microbial rhodopsin diversity and evolution. *Database* **2015**,
888 (2015).

889 45. La Cono, V. *et al.* Partaking of Archaea to biogeochemical cycling in oxygen-deficient
890 zones of meromictic saline Lake Faro (Messina, Italy). *Environ. Microbiol.* **15**, 1717–1733
891 (2013).

892 46. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread
893 human gut virus crAssphage. *Nat Microbiol* **4**, 1727–1736 (2019).

894 47. Dubinkina, V. B., Ischenko, D. S., Ulyantsev, V. I., Tyakht, A. V. & Alexeev, D. G.
895 Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC*
896 *Bioinformatics* vol. 17 (2016).

897 48. Ma, Y. *et al.* Human papillomavirus community in healthy persons, defined by
898 metagenomics analysis of human microbiome project shotgun sequencing data sets. *J.*
899 *Virol.* **88**, 4786–4797 (2014).
900 49. Mendler, K. *et al.* AnnoTree: visualization and exploration of a functionally annotated
901 microbial tree of life. *Nucleic Acids Res.* **47**, 4442–4448 (2019).
902 50. Martiny, A. C., Treseder, K. & Pusch, G. Phylogenetic conservatism of functional traits in
903 microorganisms. *ISME J.* **7**, 830–838 (2013).
904 51. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter.
905 *Nature* **499**, 431–437 (2013).
906 52. Anantharaman, K. *et al.* Expanded diversity of microbial groups that shape the dissimilatory
907 sulfur cycle. *ISME J.* **12**, 1715–1728 (2018).
908 53. Salazar, G. *et al.* Gene Expression Changes and Community Turnover Differentially Shape
909 the Global Ocean Metatranscriptome. *Cell* **179**, 1068-1083.e21 (2019).
910 54. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically
911 annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids*
912 *Res.* **47**, D309–D314 (2019).
913 55. Heffernan, B., Murphy, C. D. & Casey, E. Comparison of planktonic and biofilm cultures of
914 Pseudomonas fluorescens DSM 8341 cells grown on fluoroacetate. *Appl. Environ.*
915 *Microbiol.* **75**, 2899–2907 (2009).
916 56. Scales, B. S., Dickson, R. P., LiPuma, J. J. & Huffnagle, G. B. Microbiology, genomics, and
917 clinical significance of the Pseudomonas fluorescens species complex, an unappreciated
918 colonizer of humans. *Clin. Microbiol. Rev.* **27**, 927–948 (2014).
919 57. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat.*
920 *Commun.* **9**, 2542 (2018).
921 58. Francino, M. P. The ecology of bacterial genes and the survival of the new. *Int. J. Evol.*
922 *Biol.* **2012**, 394026 (2012).

923 59. Muller, E. E. L. Determining Microbial Niche Breadth in the Environment for Better
924 Ecosystem Fate Predictions. *mSystems* **4**, (2019).
925 60. Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I. & Watson, M. A Review
926 of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. *Front.*
927 *Genet.* **8**, 23 (2017).
928 61. Ivanova, N. N. *et al.* Stop codon reassignments in the wild. *Science* **344**, 909–913 (2014).
929 62. Steinegger, M., Mirdita, M. & Söding, J. Protein-level assembly increases protein sequence
930 recovery from metagenomic samples manyfold. *Nat. Methods* **16**, 603–606 (2019).
931 63. Titus Brown, C. *et al.* Exploring neighborhoods in large metagenome assembly graphs
932 reveals hidden sequence diversity. *bioRxiv* 462788 (2018) doi:10.1101/462788.
933 64. Höps, W., Jeffryes, M. & Bateman, A. Gene Unprediction with Spurio: A tool to identify
934 spurious protein sequences. *F1000Res.* **7**, 261 (2018).
935 65. Heinzinger, M. *et al.* Modeling aspects of the language of life through transfer-learning
936 protein sequences. *BMC Bioinformatics* **20**, 723 (2019).
937 66. Breitwieser, F. P., Pertea, M., Zimin, A. & Salzberg, S. L. Human contamination in bacterial
938 genomes has created thousands of spurious proteins. *Genome Res.* (2019)
939 doi:10.1101/gr.245373.118.
940 67. Steinegger, M. & Salzberg, S. L. Terminating contamination: large-scale search identifies
941 more than 2,000,000 contaminated entries in GenBank. *Genome Biol.* **21**, 115 (2020).
942 68. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and
943 complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
944 69. Thomas, A. M. & Segata, N. Multiple levels of the unknown in microbiome research. *BMC*
945 *Biol.* **17**, 48 (2019).
946 70. Duarte, C. M. Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation
947 Expedition. *Limnol. Oceanog. Bull.* **24**, 11–14 (2015).
948 71. 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 72. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome
951 Project. *Nature* **550**, 61–66 (2017).
952 73. 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 74. Köster, J. Reproducible data analysis with Snakemake. *F1000Res.* **7**, (2018).
955 75. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
956 identification. *BMC Bioinformatics* **11**, 119–119 (2010).
957 76. Mendler, K. *et al.* AnnoTree: visualization and exploration of a functionally annotated
958 microbial tree of life. *Nucleic Acids Res.* **47**, 4442–4448 (2019).
959 77. Eberhardt, R. Y. *et al.* AntiFam: a tool to help identify spurious ORFs in protein annotation.
960 *Database* **2012**, bas003–bas003 (2012).
961 78. Yooseph, S., Li, W. & Sutton, G. Gene identification and protein classification in microbial
962 metagenomic sequence data via incremental clustering. *BMC Bioinformatics* **9**, 1–13
963 (2008).
964 79. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity
965 searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
966 80. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future.
967 *Nucleic Acids Res.* **44**, D279–D285 (2016).
968 81. Bennett, K. D. Determination of the number of zones in a biostratigraphical sequence. *New*
969 *Phytol.* **132**, 155–170 (1996).
970 82. Daily, J. Parasail: SIMD C library for global, semi-global, and local pairwise sequence
971 alignments. *BMC Bioinformatics* **17**, 81–81 (2016).
972 83. Žure, M., Fernandez-Guerra, A., Munn, C. B. & Harder, J. Geographic distribution at
973 subspecies resolution level: closely related Rhodopirellula species in European coastal
974 sediments. *ISME J.* **11**, 478–489 (2017).

975 84. Chafee, M. *et al.* Recurrent patterns of microdiversity in a temperate coastal marine
976 environment. *ISME J.* **12**, 237–252 (2018).
977 85. Csardi, G. & Nepusz, T. The igraph software package for complex network research.
978 *InterJournal* vol. Complex Systems 1695 (2006).
979 86. Deorowicz, S., Debudaj-Grabysz, A. & Gudyś, A. FAMSA: Fast and accurate multiple
980 sequence alignment of huge protein families. *Sci. Rep.* **6**, 33964–33964 (2016).
981 87. Vanhoutreve, R. *et al.* LEON-BIS: multiple alignment evaluation of sequence neighbours
982 using a Bayesian inference system. *BMC Bioinformatics* **17**, 271–271 (2016).
983 88. Jehl, P., Sievers, F. & Higgins, D. G. OD-seq: outlier detection in multiple sequence
984 alignments. *BMC Bioinformatics* **16**, 269–269 (2015).
985 89. Broder, A. Z. On the resemblance and containment of documents. in *Proceedings.*
986 *Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)* 21–29 (IEEE,
987 1997).
988 90. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*
989 **45**, D158–D169 (2017).
990 91. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology
991 Information. *Nucleic Acids Res.* **46**, D8–D13 (2018).
992 92. Hingamp, P. *et al.* Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans
993 microbial metagenomes. *ISME J.* **7**, 1678–1695 (2013).
994 93. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: Lightning-fast iterative protein
995 sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2012).
996 94. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*
997 **46**, 2699 (2018).
998 95. Dick, J. M. Calculation of the relative metastabilities of proteins using the CHNOSZ
999 software package. *Geochem. Trans.* **9**, 10 (2008).
1000 96. Hausser, J. & Strimmer, K. Entropy inference and the James-Stein estimator, with

application to nonlinear gene association networks. *arXiv [stat.ML]* (2008).

97. van Dongen, S. & Abreu-Goodger, C. Using MCL to Extract Clusters from Networks. in *Bacterial Molecular Networks: Methods and Protocols* (eds. van Helden, J., Toussaint, A. & Thieffry, D.) 281–295 (Springer New York, 2012).

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

99. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

100. Berger, S. A. & Stamatakis, A. PaPaRa 2.0: a vectorized algorithm for probabilistic phylogeny-aware alignment extension. *Heidelberg Institute for Theoretical Studies, http://sco. h-its. org/exelixis/publica tions. html. Exelixis-RRDR-2012-2015* (2012).

101. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**, 538 (2010).

102. Needham, D. M. *et al.* A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 20574–20583 (2019).

103. Murat Eren, A. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).

104. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

105. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

106. Levins, R. THE STRATEGY OF MODEL BUILDING IN POPULATION BIOLOGY. *Am. Sci.* **54**, 421–431 (1966).

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).

1027 108. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree:
1028 the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).
1029 109. Miklós, I. & Podani, J. RANDOMIZATION OF PRESENCE–ABSENCE MATRICES:
1030 COMMENTS AND NEW ALGORITHMS. *Ecology* vol. 85 86–92 (2004).
1031 110. Salazar, G. *et al.* Particle-association lifestyle is a phylogenetically conserved trait in
1032 bathypelagic prokaryotes. *Mol. Ecol.* **24**, 5692–5706 (2015).