Unifying the known and unknown microbial coding sequence space

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

Genes of unknown function are among the biggest challenges in molecular biology, especially in microbial systems, where 40%-60% of the predicted genes are unknown. Despite previous attempts, systematic approaches to include the unknown fraction into analytical workflows are still lacking. Here, we propose a conceptual framework and a computational workflow that bridge the known-unknown gap in genomes and metagenomes. We showcase our approach by exploring 415,971,742 genes predicted from 1,749 metagenomes and 28,941 bacterial and archaeal genomes. We quantify the extent of the unknown fraction, its diversity, and its relevance across multiple biomes. Furthermore, we provide a collection of 283,874 lineage-specific genes of unknown function for *Cand*. Patescibacteria, being a significant resource to expand our understanding of their unusual biology. Finally, by identifying a target gene of unknown function for antibiotic resistance, we demonstrate how we can enable the generation of hypotheses that can be used to augment experimental data.

Introduction

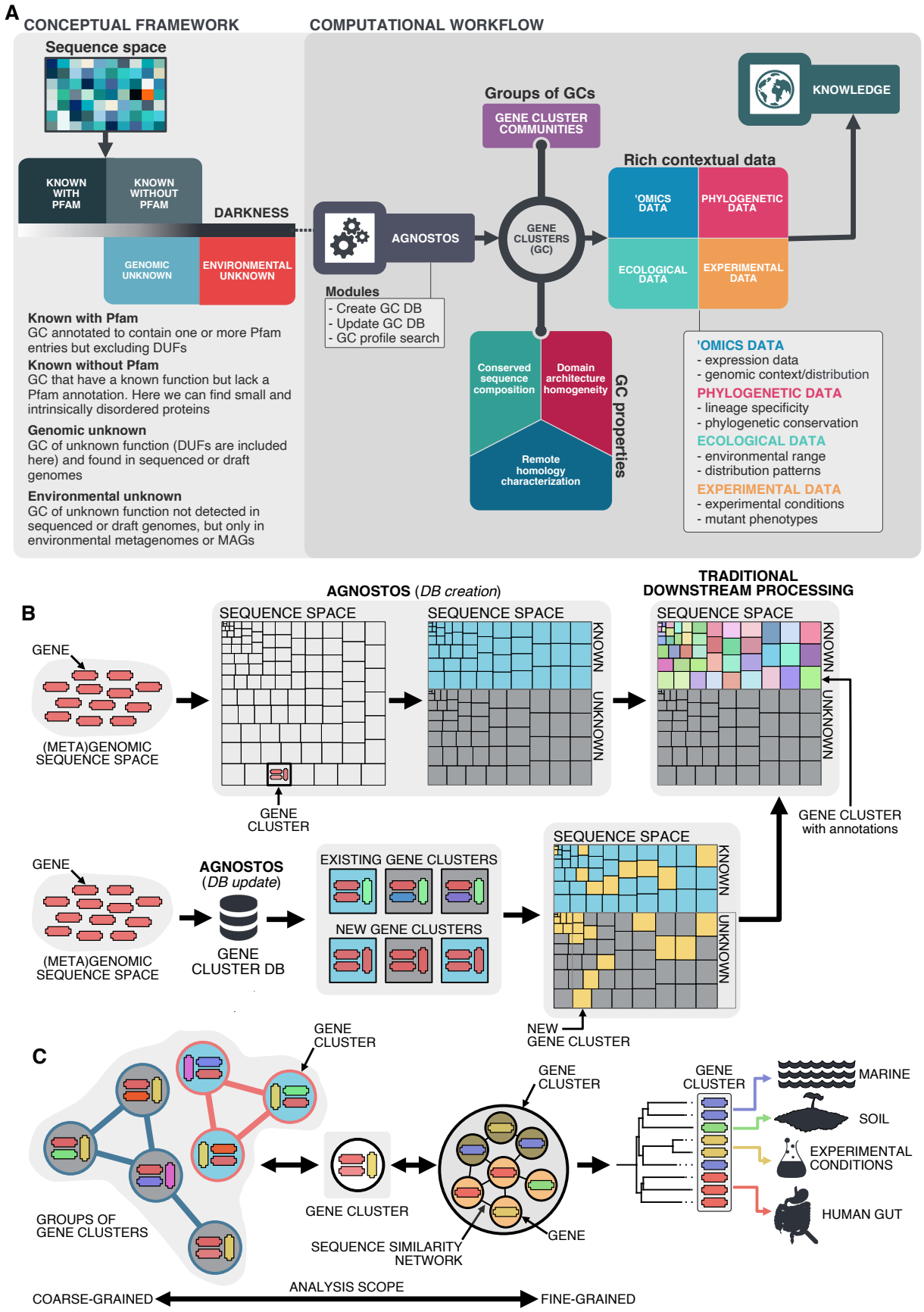
Thousands of isolate, single-cell, and metagenome-assembled genomes are guiding us towards a better understanding of life on Earth (Almeida et al., 2019; Cross et al., 2019; Delmont et al., 2020; Hug et al., 2016; Kopf et al., 2015; Pachiadaki et al., 2019; Pasolli et al., 2019; Sunagawa et al., 2015). At the same time, the ever-increasing number of genomes and metagenomes, unlocking uncharted regions of life's diversity, (Brown et al., 2015; Eloe-Fadrosh et al., 2016; Hug et al., 2016) are providing new perspectives on the evolution of life (Parks et al., 2018; Spang et al., 2015). However, our rapidly growing inventories of new genes have a glaring issue: between 40% and 60% cannot be assigned to a known function (Almeida et al., 2020; Bernard et al., 2018; Carradec et al., 2018; Price et al., 2018). Current analytical approaches for genomic and metagenomic data (Chen et al., 2019; Franzosa et al., 2018; Huerta-Cepas et al., 2017; Mitchell et al., 2020; Quince et al., 2017) generally do not include this uncharacterized fraction in downstream analyses, constraining their results to conserved pathways and housekeeping functions (Quince et al., 2017). This inability to handle the unknown is an immense impediment to realizing the potential for discovery of microbiology and molecular biology at large (Bernard et al., 2018; Hanson et al., 2010). Predicting function from traditional single sequence similarity appears to have yielded all it can (Arnold, 2018, 1998; Brandenberg et al., 2017), thus several groups have attempted to resolve gene function by other means. Such efforts include combining biochemistry and crystallography (Jaroszewski et al., 2009); using environmental co-occurrence (Buttigieg et al., 2013); by grouping those genes into evolutionarily related families (Bateman et al., 2010; Brum et al., 2016; Wyman et al., 2018; Yooseph et al., 2007); using remote homologies (Bitard-Feildel and Callebaut, 2017; Lobb et al., 2015); or more recently using deep learning approaches (Bileschi et al., 2019; Liu, 2017). In 2018, Price et al. (Price et al., 2018) developed a high-throughput experimental pipeline that provides mutant phenotypes for thousands of bacterial genes of unknown function being one of

80 the most promising methods to tackle the unknown. Despite their promise, experimental
81 methods are labor-intensive and require novel computational methods that could bridge the
82 existing gap between the genes with known and unknown function.
83 Here we present a conceptual framework and a computational workflow that closes the gap by
84 connecting genomic and metagenomic data by the exploitation of groups of homologous genes,
85 and facilitates the integration of genes of unknown function into ecological, evolutionary, and
86 biotechnological investigations. The conceptual framework is based on the partitioning of the
87 known and unknown fractions into four different categories that reflects the level of darkness
88 (Figure 1A). The category “Known” (K) contains these sequences predicted to harbor domains
89 of known function described by Pfam (Finn et al., 2016). Some sequences without Pfam
90 domains of unknown function, like the ones encoding for small and intrinsically disordered
91 proteins, can have homology to characterized proteins, hence we classify them as "Known
92 without Pfam" (KWP). Finally, sequences that cannot be classified by either of these criteria, are
93 classified as "Genomic Unknown" (GU) when they may be found in the genomes in reference
94 databases; and "Environmental Unknown" (EU) when only have been observed in
95 environmental samples. By contextualizing the different categories with information from several
96 sources (Figure 1C), we provide an invaluable resource for the investigation of the genes with
97 unknown function and boost the current methods for its experimental characterization.
98 The application of our approach to 415,971,742 genes predicted from 1,749 metagenomes and
99 28,941 bacterial and archaeal genomes revealed that the unknown fraction (1) is smaller than
100 has been previously reported, (2) is exceptionally diverse, and (3) is phylogenetically more
101 conserved than the known fraction and predominantly lineage-specific at the Species level.
102 Finally, we show how we can connect all the outputs produced by our approach to augment the
103 results from experimental data and add context to genes of unknown function through
104 hypothesis-driven molecular investigations.

Results

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

Driven by the concepts defined in the conceptual framework, we developed AGNOSTOS, a computational workflow that infers, validates, refines, and classifies groups of homologous genes or gene clusters (GCs) in the four proposed categories (Fig. 1A; Fig. 1B; Supp. Fig 1). AGNOSTOS produces 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 genes before classifying each GC in one of the four categories (Methods). In the last step, we decorate each GC with a rich collection of contextual data compiled from different sources or generated 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 (Steinegger and Soding, 2017) and HHblits (Steinegger et al., 2019a), or the GC members as a sequence similarity network (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.


A
CONCEPTUAL FRAMEWORK
COMPUTATIONAL WORKFLOW
Sequence space
KNOWN WITH PFAM
KNOWN WITHOUT PFAM
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. Here we can find small and intrinsically disordered proteins
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)GENOMIC SEQUENCE SPACE
SEQUENCE 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
ANALYSIS SCOPE
COARSE-GRAINED
FINE-GRAINED

124

Figure 1: Conceptual framework to unify the known and unknown sequence space and integration of the framework in the current analytical workflows (A) Link between the conceptual framework and the computational workflow to partition the sequence 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 generating hypotheses that can be used to augment experimental data. (B) The computational workflow provides two mechanisms to structure sequence space using GCs, de novo creation of the GCs (*DB creation*), or integrating the dataset in an existing GC database (*DB update*). The structured sequence 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 integrate the unknown fraction into microbiome analyses easily. 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, given that GCs are conserved across environments, organisms and experimental conditions give us access to an unprecedented amount of information to design and interpret experimental data.

Partitioning and contextualizing the coding sequence space of genomes and metagenomes

First, we used AGNOSTOS in a metagenomic context to show its ability to process complex and noisy datasets. We explored the unknown sequence space of 1,749 human and marine metagenomes. In total, we predicted 322,248,552 genes from the environmental dataset and assigned Pfam annotations to 44% of them (Supp. Fig. 2A). Next, AGNOSTOS clustered the predicted genes in 32,465,074 GCs and flagged those gene clusters that contain less than ten genes (Skewes-Cox et al., 2014). Flagged gene clusters are not processed through the validation workflow because they don't have enough sequences to provide reliable results. We flagged 29,461,177 gene clusters, 19,911,324 of them being singletons (Supp. Fig. 2A; Supp. Note 1). A total of 3,003,897 GCs (83% of the original genes) will go through the validation steps. 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 unknowns after the classification and remote homology refinement steps (Supp. Fig. 2A, Supp. Note 3).

158 Lastly, we demonstrate how AGNOSTOS can integrate a new dataset into the already existing
159 metagenomic database by integrating 28,941 genomes from the GTDB_r86 (Supp. Fig. 2A).
160 With this integration we can build links between the environmental and genomic sequence
161 space by expanding the final collection of GCs with the genes predicted from GTDB_r86.
162 Surprisingly, the integration showed that the environmental GCs (human and marine) not only
163 included a large proportion of the genes predicted from GTDB_r86 (72%) but also that most of
164 these genes (90%) are part of the known sequence space (Supp. Fig. 2A; Supp. Note 4; Supp.
165 Note 5). Only 22% of the GTDB_r86 genes created new GCs (2,400,037), being most of these
166 genes (84%) classified as unknowns by AGNOSTOS (Supp. Fig. 2A; Supp. Note 4; Supp. Note
167 5). Finally, only a small proportion of the genes from GTDB_r86 (6%) resulted in singletons
168 (Supp. Fig. 2A; Supp. Note 4; Supp. Note 5).
169 The final dataset after being validated and categorized resulted in 5,287,759 GCs (Supp. Fig.
170 2A), with both datasets sharing only 922,599 GCs (Supp. Fig. 2B). The integration of GTDB_r86
171 into the metagenomic database increased the proportion of GCs in the unknown sequence
172 space from 43% to the 54%.
173 Additionally, AGNOSTOS identified a subset of 203,217 *high-quality* GCs (Supp. Table 2). In
174 these *high-quality* GCs, we identified 12,313 clusters potentially encoding for small proteins (<=
175 50 amino acids). Most of these GCs are unknown (66%), which agrees with recent findings on
176 novel small proteins from metagenomes (Sberro et al., 2019). We also observed that the KWP
177 category contains the largest proportion of incomplete genes (Supp. Table 3), disrupting the
178 detection and assignment of Pfam domains. But it also incorporates sequences with an unusual
179 amino acid composition that have homology to proteins with high levels of disorder in the DPD
180 database (Perdigão et al., 2017) and has characteristic functions of intrinsically disordered
181 proteins (Habchi et al., 2014) (IDP) like cellular processes and signaling as predicted by
182 eggNOG annotations (Supp. Table 4).

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

185 Beyond the twilight zone, communities of gene clusters

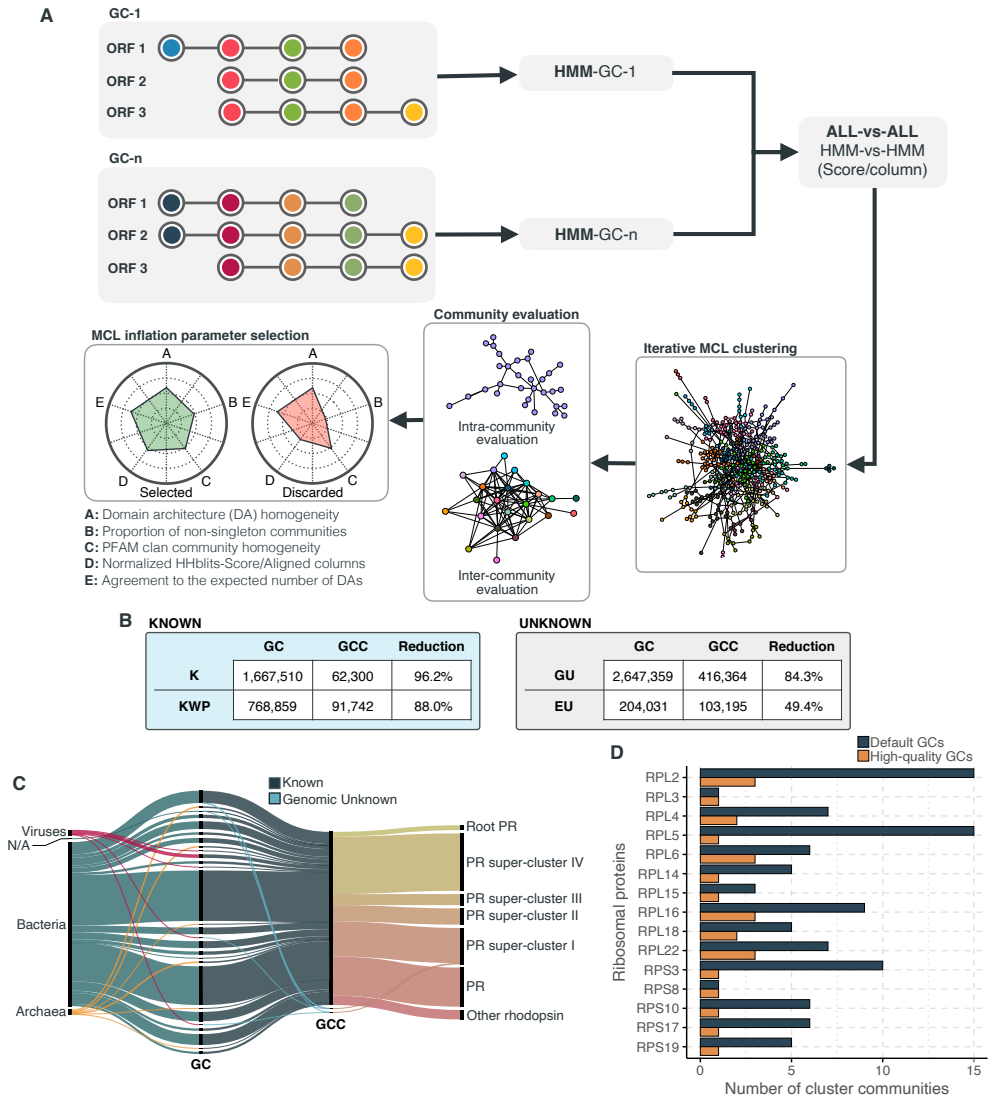
186 To find relationships between gene clusters, we grouped them in gene cluster communities
187 (GCCs) (Fig. 2A) using remote homologies. As we are dealing with the unknown, we identified
188 GCCs independently in each category, using the gene clusters from category K as a reference
189 to identify the best parameters for the clustering of the gene cluster homology network
190 (Methods; Fig. 2B; Supp. Note 7) and applying the learnt parameters to the other categories.
191 The grouping reduced the final collection of GCs by 87%, producing 673,601 GCCs (Methods;
192 Fig. 2B; Supp. Note 7).
193 We validated our approach to capture remote homologies between related GCs using two well-
194 known gene families present in our environmental datasets, proteorhodopsins (Olson et al.,
195 2018) and bacterial ribosomal proteins (Méheust et al., 2019). Theoretically, we would expect to
196 have one or a very low number of GCCs for each of these gene families.
197 We identified 64 GCs (12,184 genes) and 3 GCCs (Supp. Note 8) containing genes classified
198 as proteorhodopsin (PR). One community from the category K contained 99% of the PR
199 annotated genes (Fig. 2C), except 85 genes taxonomically annotated as viral and assigned to
200 the *PR Supercluster I* (Boeuf et al., 2015) within two GU communities (five GU gene clusters;
201 Supp. Note 8).
202 For the ribosomal proteins, the results were not so satisfactory. We identified 1,843 GCs
203 (781,579 genes) and 98 GCCs. The number of GCCs is larger than the expected number of
204 ribosomal protein families (16) used for validation. When we used *high-quality* GCs (Supp. Note
205 8), we got closer to the expected number of GCCs (Fig. 2D). With this subset, we identified 26
206 GCCs and 145 GCs (1,687 genes). The cross-validation of our method against the approach
207 used in Méheust et al. (Méheust et al., 2019) (Supp. Note 9) confirms the intrinsic complexity of

208 analyzing metagenomic data. Both approaches showed a high agreement in the GCCs

209 identified (Supp. Table 9-1). Still, our method inferred fewer GCCs for each of the ribosomal

210 protein families (Supp. Fig. 9-3), coping better with the complexities of a metagenomic setup,

211 such as incomplete genes (Supp. Table 6).



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

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

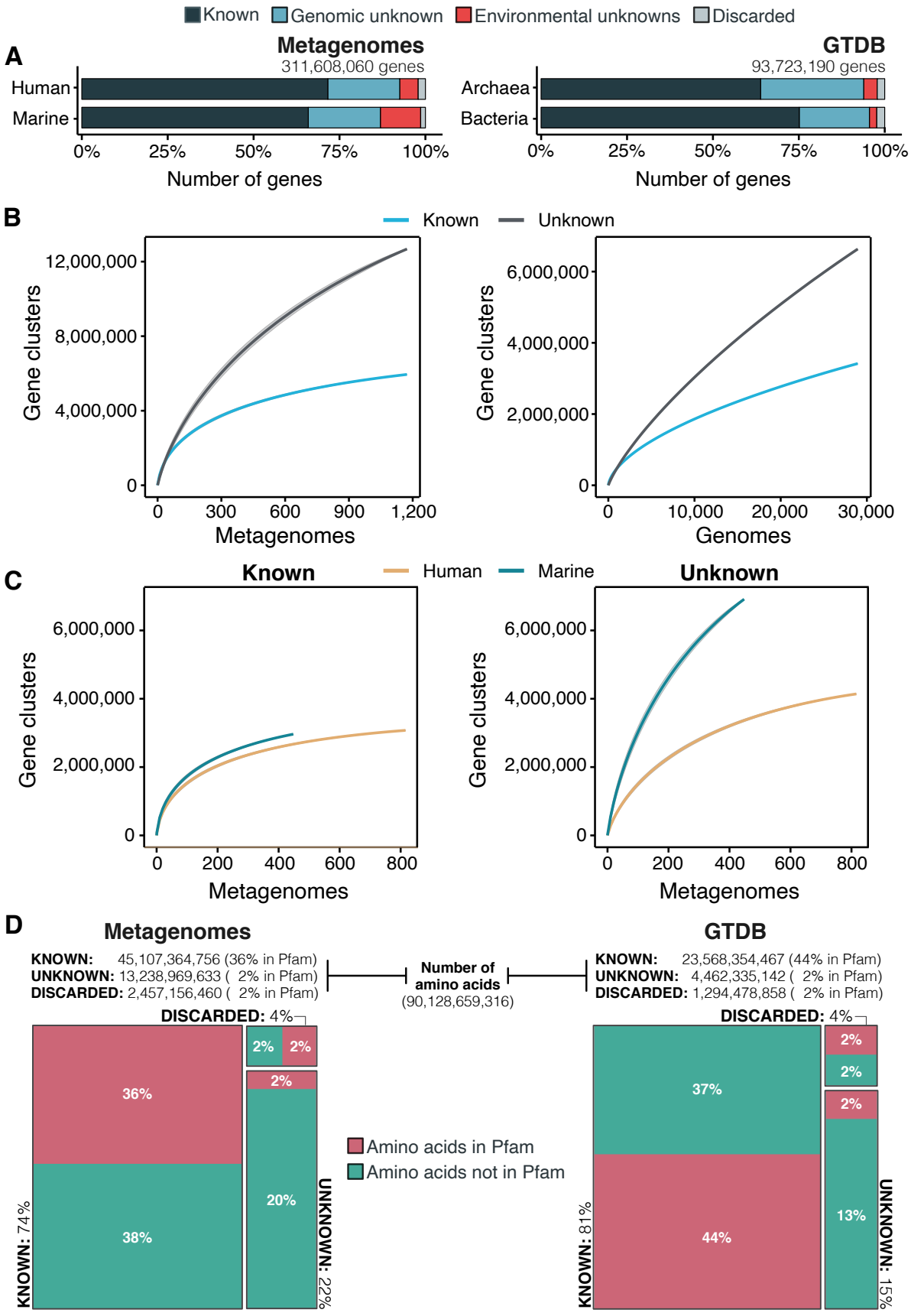
212

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

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

223 A smaller but highly diverse unknown coding sequence space

224 By combining clustering and remote homology searches we reduced the extent of the unknown
225 sequence space compared to what is reported by the traditional genomic and metagenomic
226 analysis approaches where the unknown fraction can add up to 60% in marine metagenomes
227 (Salazar et al., 2019) or up to 40% in human metagenomes (Thomas and Segata, 2019) (Fig.
228 3A). Our workflow recruited as much as 71% of genes in human-related metagenomic samples
229 and 65% of the genes in marine metagenomes into the known sequence space. In both human
230 and marine microbiomes, the genomic unknown fraction showed a similar proportion of genes
231 (21%, Fig. 3A). The number of genes corresponding to EU gene clusters is higher in marine
232 metagenomes; 12% of the genes are part of this GC category. We obtained a comparable result
233 when we evaluated the genes from the GTDB_r86, 75% of bacterial and 64% of archaeal genes
234 were part of the known sequence space. Archaeal genomes contained more unknowns than
235 those from Bacteria, where 30% of the genes are classified as genomic unknowns in Archaea,
236 and only 20% in Bacteria (Fig. 3A; Supp. Table 7).
237 To further evaluate the differences between the known and unknown sequence space, we
238 calculated the accumulation rates of GCs and GCCs, using them as operational protein families
239 (Schloss and Handelsman, 2008). For the metagenomic dataset we used 1,264 metagenomes
240 (18,566,675 GCs and 282,580 GCCs) and for the genomic dataset 28,941 genomes (9,586,109
241 GCs and 496,930 GCCs). The rate of accumulation of unknown GCs was three times higher
242 than the known (2X for the genomic), and in both cases the curves were far from reaching a
243 plateau (Fig. 3B). This is not the case for the GCC accumulation curves (Supp. Fig. 4B), which
244 reached a plateau.


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
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
C
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
D
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%
2%
2%
2%
36%
38%
20%
KNOWN: 74%
UNKNOWN: 22%
Amino acids in Pfam
Amino acids not in Pfam
DISCARDED: 4%
2%
2%
2%
37%
44%
13%
KNOWN: 81%
UNKNOWN: 15%

245

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

254 The accumulation rate is largely determined by the number of singletons, especially singletons
255 from EUs (Supp. note 10 and Supp. Fig. 5). While the accumulation rate of known GCs between
256 marine and human metagenomes is almost identical, there are striking differences for the
257 unknown GCs (Fig. 3C). These differences are maintained even when we remove the virus-
258 enriched samples from the marine metagenomes (Supp. Fig. 4A). Although the marine
259 metagenomes include a large variety of environments, from coastal to the deep sea, the known
260 space remains quite constrained.
261 Next, we wanted to know how much of the sequence space we integrated with AGNOSTOS
262 was found in other databases (Supp. Note 11). Despite only including marine and human
263 metagenomes in our database, we already cover, in average, 76% of the sequence space of
264 seven datasets spanning different databases and environments (Supp. Fig. 11-1). By screening
265 MGnify (Mitchell et al., 2020) (release 2018_09; 11 biomes; 843,535,6116 proteins) we
266 identified freshwater, soil and human non-digestive as the biomes less covered by our data
267 (Supp. Fig. 6). Two of the seven analyzed datasets are designed to study genes of unknown
268 function (Supp. Table 11-1). On (Wyman et al., 2018), where they defined Function Unknown
269 Families of homologous proteins (FunkFams), we identified 20% of their FunkFams to be
270 members of the known sequence space. On (Price et al., 2018), we classified as known, 44% of
271 the genes of unknown function used in their experimental conditions.
272 One indirect consequence of our approach is that we can provide a detailed view of the
273 sequence space at the amino acid level. We estimated the number of amino acids belonging to
274 the known or unknown sequence space, and how many of these amino acids are contained

275 within Pfam domain boundaries. From the 90,128,659,316 amino acids analyzed, most of the
276 amino acids in metagenomes (74%) and genomes (80%) are in the known sequence space
277 (Fig. 3D; Supp. Table 7) while only 22% in metagenomes and 15% in genomes are part of the
278 unknowns. In both cases, approximately 40% of the amino acids in the known sequence space
279 were part of a Pfam domain (Fig. 3D; Supp. Table 7). While this result is expected based on the
280 large number of genes present in the known space (Fig. 3A), what is surprising is the low
281 proportion of amino acids (2%) corresponding to DUF Pfam domains in genomes and
282 metagenomes (Fig. 3D). If we use as reference the proportions of amino acids observed in the
283 known sequence space, we can hypothesize that there are still many DUFs to be unearthed.
284 With AGNOSTOS and its throughout validation and characterization of the genomic and
285 environmental unknowns, we provide the basic building blocks (gene clusters' multiple
286 sequence alignments) to identify conserved regions that might become new potential DUFs.

287 The unknown sequence space has a limited ecological 288 distribution in human and marine environments

289 Although the role of the unknown fraction in the environment is still a mystery, the large number
290 of gene counts and abundance observed underlines its inherent ecological relevance (Fig. 4A).
291 In some metagenomes, the genomic unknown fraction can account for more than 40% of the
292 total gene abundance observed (Fig. 4A). The environmental unknown fraction is also relevant
293 in several samples, where singleton GCs are the majority (Fig. 4A). We identified two
294 metagenomes with an unusual composition in terms of environmental unknown singletons. The
295 marine metagenome corresponds to a sample from Lake Faro (OSD42), a meromictic saline
296 with a unique extreme environment where Archaea plays an important role (La Cono et al.,
297 2013). And the HMP metagenome (SRS143565) that corresponds to a human sample from the
298 right cubital fossa from a healthy female subject. To understand this unusual composition, we

should perform further analyses to discard potential technical artifacts like sample contamination.

The ratio between the unknown and known GCs is useful to reveal which metagenomes are enriched in GCs of unknown function (upper left quadrant in Fig. 4B-C) and it can be used as a proxy to assess the sequence contained in a metagenome. In human metagenomes, this ratio can distinguish between body sites, with the gastrointestinal tract, an ecologically complex environment (Qin et al., 2010), significantly enriched with genomic unknowns. Furthermore, it is not surprising that the human and marine metagenomes with the largest ratio of unknowns are those samples enriched with viral sequences. Specifically, in the HMP metagenomes are those samples identified to contain crAssphages (Dubinkina et al., 2016; Edwards et al., 2019) and HPV viruses (Ma et al., 2014) (Supp. Table 8; Supp. Fig. 7). While, in marine metagenomes (Fig. 4C) the highest ratio in genomic and environmental unknowns correspond to the ones enriched with viruses and giant viruses.

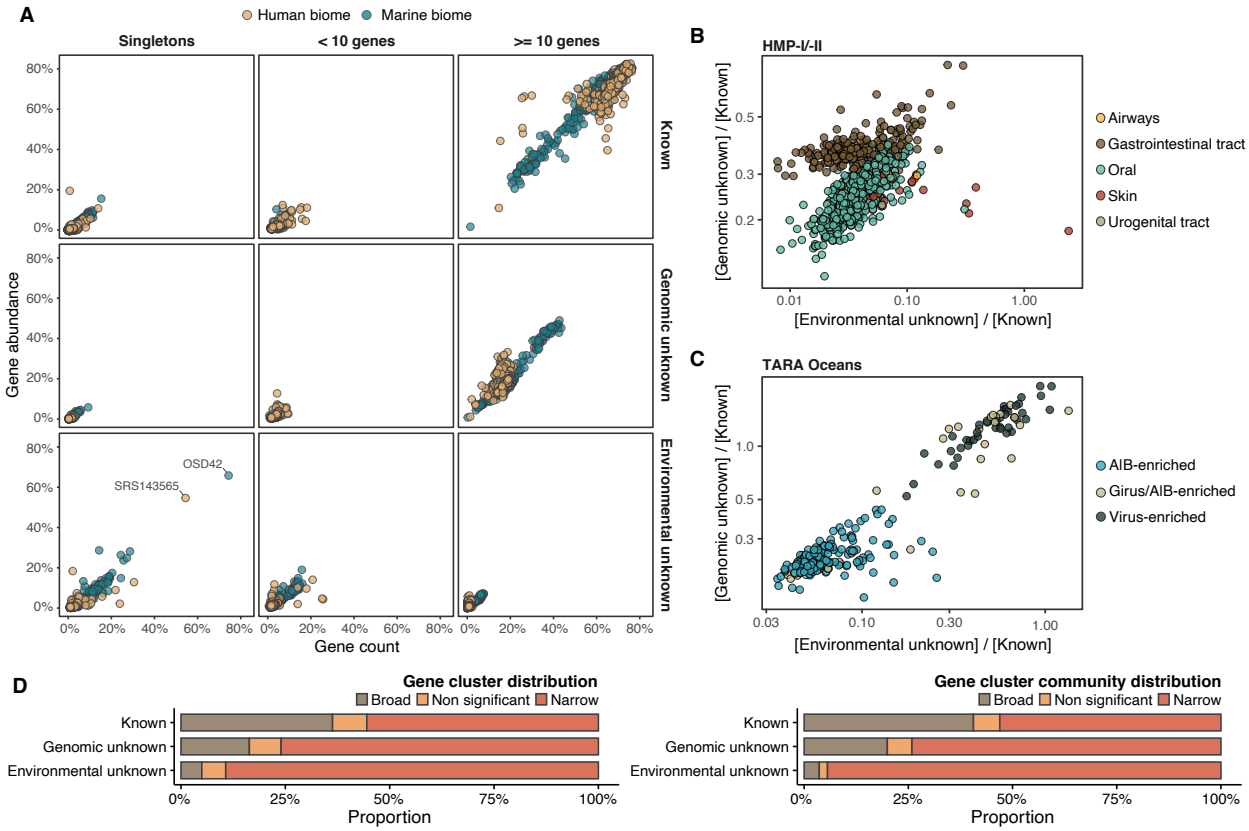


Figure 4: Distribution of the unknown sequence space in the human and marine metagenomes (A) Ratio between the proportion of the number of genes and their estimated abundances per cluster category and biome. Columns represented in the facet depicts three cluster categories based on the size of the clusters.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

We performed a large-scale analysis to investigate the occurrence patterns of the GCs in the environment by analyzing their abundance and distribution breadth. The narrow distribution of the unknown fraction (Fig. 4D) suggests that these GCs might provide a selective advantage and be necessary to adapt to specific environmental conditions. But the pool of broadly distributed environmental unknowns is the most exciting 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 12).

The genomic unknown coding sequence space is lineage-specific

With the inclusion of the genomes from GTDB_r86, we have access to a phylogenomic framework that can be used to assess how exclusive is a GC within a lineage (lineage-specifity (Mendler et al., 2019)) and how conserved (phylogenetic conservation) is a GC across the different clades in the GTDB_r86 phylogenomic tree (Martiny et al., 2013). 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 (Parks et al., 2018) (Fig. 5A) and differences between the known and the unknown fraction start to be evident at the Family level resulting in 4X more unknown lineage-specific GCs at the Species level. In general terms, the unknown

343 GCs are more phylogenetically conserved (GCs shared among members of deep clades) than
344 the known (Fig. 5B, $p < 0.0001$), revealing the importance of the genome's uncharacterized
345 fraction. However, the lineage-specific unknown GCs are less phylogenetically conserved (Fig.
346 5B) than the known, agreeing with the large number of lineage-specific GCs observed at Genus
347 and Species level (Fig. 5A).
348 One potential confounding factor that might contribute to inflate the number of lineage-specific
349 GCs in the unknown fraction, is the presence of prophages owing to their potential host
350 specificity (Ross et al., 2016). To discard the possibility that the lineage-specific GCs of
351 unknown function have a viral origin, we screened all GTDB_r86 genomes for prophages. We
352 only found 37,163 lineage-specific GCs in prophage genomic regions, being 86% of these GCs
353 of unknown function.
354 After unveiling the potential relevance of the GCs of unknown function in bacterial genomes, we
355 identified phyla in GTDB_r86 enriched with these types of clusters. A clear pattern emerged
356 when we partitioned the phyla based on the ratio of known to unknown GCs and vice versa (Fig.
357 5D), the phyla with a larger number of MAGs are enriched in GCs of unknown function (Fig.
358 5D). Phyla with a high proportion of non-classified GCs (those discarded during the validation
359 steps) contain a small number of genomes and are primarily composed of MAGs. These groups
360 of phyla highly enriched in unknowns and represented mainly by MAGs include newly described
361 phyla such as *Cand.* Riflebacteria and *Cand.* Patescibacteria (Anantharaman et al., 2018;
362 Brown et al., 2015; Rinke et al., 2013), both with the largest unknown to known ratio.
363

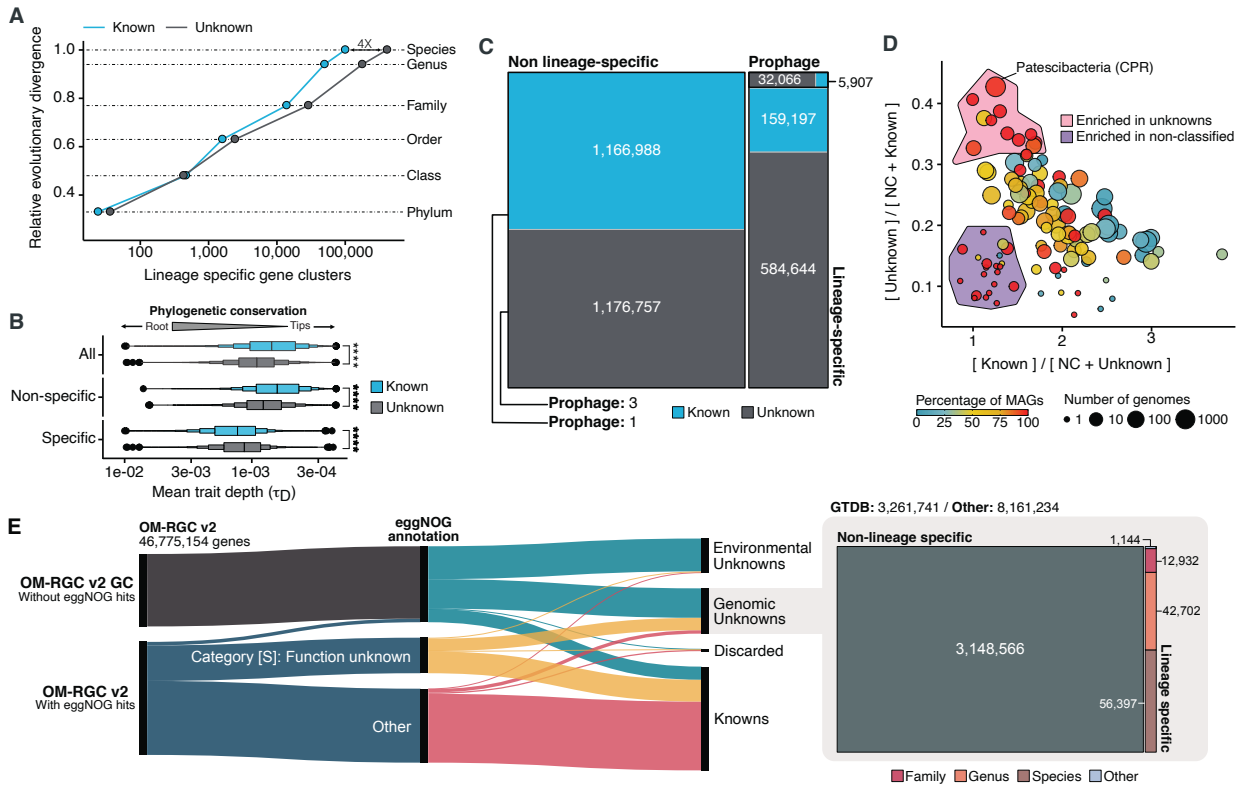


Figure 5: Phylogenomic exploration of the unknown sequence space. (A) Distribution of the lineage-specific GCs by taxonomic level. Lineage-specific unknown GCs are more abundant in the lower taxonomic levels (Genus, Species). (B) Phylogenetic conservation of the known and unknown sequence space in 27,372 bacterial genomes from GTDB_r86. We observe differences in the conservation between the known and the unknown 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 sequence space, being a small proportion found in prophages present in the GTDB_r86 genomes. (D) Known and unknown 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 alluvial plot's left side 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.(Salazar et al., 2019). The right side of the alluvial plot shows the new organization of the OM-RGC v2 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

One of the strengths of AGNOSTOS is the possibility of bridging genomic and metagenomic data and simultaneously unifying the known and unknown sequence space by integrating the new Ocean Microbial Reference Gene Catalog (Salazar et al., 2019) (OM-RGC v2) into our

387 database. We assigned 26,170,875 genes to known GCs, 11,422,975 to genomic unknowns,
388 8,661,221 to environmental unknown and 520,083 were discarded. From the 11,422,975 genes
389 classified as genomic unknowns, we could associate 3,261,741 to a GTDB_r86 genome and we
390 identified 113,175 as lineage-specific. The alluvial plot in Fig. 5E depicts the new organization of
391 the OM-RGC v2 after being integrated into our framework and how we can provide context to
392 the two original types of unknowns in the OM-RGC (those annotated as category S in eggNOG
393 (Huerta-Cepas et al., 2019) and those without known homologs in the eggNOG database
394 (Salazar et al., 2019)) that can lead to potential experimental targets at the organism level to
395 complement the metatranscriptomic approach proposed by Salazar et al. (Salazar et al., 2019).

396 A structured sequence space augments the interpretation of 397 experimental data

398 We selected one of the experimental conditions tested in Price et al. (Price et al., 2018) to
399 demonstrate the potential of our approach to augment experimental data. We compared the
400 fitness values in plain rich medium with added Spectinomycin dihydrochloride pentahydrate to
401 the fitness in plain rich medium (LB) in *Pseudomonas fluorescens FW300-N2C3* (Fig. 6A). This
402 antibiotic inhibits protein synthesis and elongation by binding to the bacterial 30S ribosomal
403 subunit and interferes with the peptidyl tRNA translocation. We identified the gene with locus id
404 AO356_08590 that presents a strong phenotype (fitness = -3.1; t = -9.1) and has no known
405 function. This gene belongs to the genomic unknown GC GU_19737823. We can track this GC
406 into the environment and explore the occurrence in the different samples we have in our
407 database. As expected, the GC is mostly found in non-human metagenomes (Fig. 6B) as
408 *Pseudomonas* are common inhabitants of soil and water environments (Heffernan et al., 2009).
409 However, finding this GC also in human-related samples is very interesting due to the potential
410 association of *P. fluorescens* and human disease where Crohn's disease patients develop

411 serum antibodies to this microbe (Scales et al., 2014). We can add another layer of information
412 to the selected GC by looking at the associated remote homologs in the GCC GU_c_21103
413 (Fig. 6C). We identified all the genes in the GTDB_r86 genomes that belong to the GCC
414 GU_c_21103 (Supp. Table 11) and explored their genomic neighborhoods. All members from
415 GU_c_21103 are constrained to the class *Gammaproteobacteria*, and interestingly
416 GU_19737823 is mostly exclusive to the order *Pseudomonadales*. The gene order in the
417 different genomes analyzed is highly conserved, finding GU_19737823 after the *rpsF*::*rpsR*
418 operon and before *rplI*. *rpsF* and *rpsR* encode for 30S ribosomal proteins, the prime target of
419 spectinomycin. The combination of the experimental evidence and the associated data inferred
420 by our approach provides strong support to generate the hypothesis that the gene
421 AO356_08590 might be involved in the resistance to Spectinomycin.
422

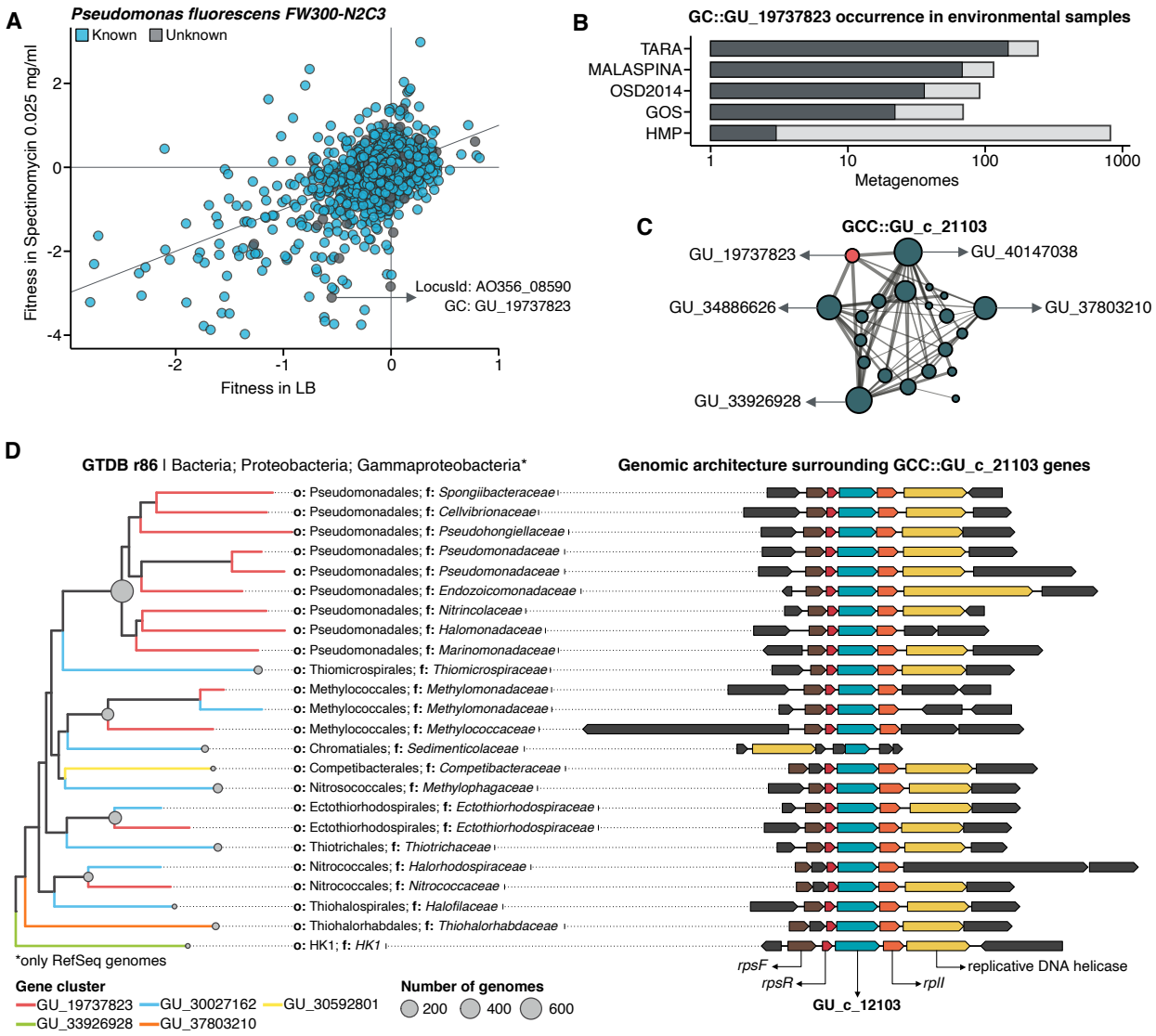


Figure 6: Augmenting experimental data with GCs of unknown function. (A) We used the fitness values from the experiments from Price et al. (Price et al., 2018) to identify genes of unknown function that are important for fitness under certain experimental conditions. The selected gene belongs to the 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. The 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 sequence space. Using this framework, we performed an in-depth exploration of the microbial unknown sequence space, demonstrating that we can link the unknown fraction of metagenomic studies to specific genomes and provide a powerful tool for hypothesis generation. 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 (Rost, 1999). With this objective in mind, we use gene clusters instead of genes as the fundamental unit to compartmentalize the sequence space owing to their unique properties (Fig. 1B). Gene clusters (1) provide a structured sequence space that helps to reduce its complexity, (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, 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).

AGNOSTOS integrates transparently into the standard operating procedure for analyzing metagenomes (Quince et al., 2017) adopted by the microbiome community. It can briefly be summarized into (1) assembly, (2) gene prediction, (3) gene catalog inference, (4) binning, and (5) characterization. AGNOSTOS exploits recent computational developments (Steinegger and Söding, 2018; Steinegger and Soding, 2017) to maximize the information used when analyzing genomic and metagenomic data. In addition, we provide a mechanism to reconcile top-down

466 and bottom-up approaches, thanks to the well-structured sequence space proposed by our
467 framework. AGNOSTOS can create environmental- and organism-specific variations of a seed
468 database based on gene clusters. Then, it integrates the predicted genes from new genomes
469 and metagenomes and dynamically creates and classifies new GCs with those genes not
470 integrated during the initial step (Fig. 1B). Afterward, the potential functions of the known GCs
471 can be carefully characterized by incorporating them into the traditional standard operating
472 procedure described previously.
473
474 One of the most appealing characteristics of our approach is that the GCs provide unified
475 groups of homologous genes across environments and organisms indifferently if they belong to
476 the known or unknown sequence space, and we can contextualize the unknown fraction using
477 this genomic and environmental information. Our combination of partitioning and
478 contextualization features a smaller unknown sequence space than previously reported
479 (Steinegger et al., 2019b). On average, only 30% of the genes fall in the unknown fraction for
480 our genomic and metagenomic data. One hypothesis to reconcile this surprising finding is that
481 the methodologies to identify remotely homologous sequences in large datasets were
482 computationally prohibitive until recently. New methods (Steinegger et al., 2019a; Steinegger
483 and Soding, 2017), like the ones used in AGNOSTOS, enable large-scale remote homology
484 searches. Still, one must apply conservative measures to control the trade-off between
485 specificity and sensitivity to avoid overclassification.
486
487 We found that most of the sequence space at the gene and amino acid level is known, both in
488 genomes and metagenomes. However, the GC accumulation curves showed that the unknown
489 fraction is far more diverse than the known. When we combine the high diversity and its narrow
490 ecological distribution, we can unveil the magnitude of the untapped unknown functional fraction
491 and its potential importance for niche adaptation. In a genomic context and after ruling out the

492 effect of prophages, the unknown fraction is predominantly Species' lineage-specific and
493 phylogenetically more conserved than the known fraction, supporting the signal observed in the
494 environmental data emphasizing that we should not ignore the unknown fraction. It is worth
495 noting that the high diversity observed in the unknowns only represents the 20% of the amino
496 acids in the sequence space we analyzed, and only 10% of this unknown amino acid space is
497 part of a Pfam domain (DUF and others). This contrasts with the numbers observed in the
498 known sequence space, where Pfam domains include 50% of the amino acids. All this evidence
499 combined strengthens the hypothesis that the genes of unknown function, especially the
500 lineage-specific ones, might be associated with the mechanisms of microbial diversification and
501 niche adaptation due to the constant diversification of gene families and the survival of new
502 gene lineages (Francino, 2012; Muller, 2019).
503
504 Metagenome-assembled genomes are not only unveiling new regions of the microbial universe
505 (42% of the genomes in GTDB_r86), but they are also enriching the tree of life with genes of
506 unknown function. One excellent example is *Cand*. Patescibacteria, more commonly known as
507 Candidate Phyla Radiation (CPR), a phylum that has raised considerable interest due to its
508 unusual biology (Brown et al., 2015). We performed an in-depth exploration of this phylum, and
509 we provide a collection of 54,343 lineage-specific GCs of unknown function at different
510 taxonomic level resolutions (Supp. Table 12; Supp. Note 14), which will be a valuable resource
511 for the CPR advancement research efforts.
512
513 One of the ultimate goals of our approach is to provide a mechanism to unlock the large pool of
514 likely relevant data that remains untapped to analysis and discovery and boost insights from
515 model organism experiments. We demonstrated the value of our approach by identifying a
516 potential target gene of unknown function for antibiotic resistance. But severe challenges
517 remain, such as the dependence on the quality of the assemblies and their gene predictions, as

518 shown by the analysis of the ribosomal protein GCCs where many of the recovered genes are
519 incomplete. While sequence assembly has been an active area of research (Roumpeka et al.,
520 2017), this has not been the case for gene prediction methods (Roumpeka et al., 2017), which
521 are becoming outdated (Ivanova et al., 2014) and cannot cope with the current amount of data.
522 Alternatives like protein-level assembly (Steinegger et al., 2019b) combined with exploring the
523 assembly graphs' neighborhoods (Brown et al., 2020) become very attractive for our purposes.
524 In any case, we still face the challenge of discriminating between genuine and artifactual
525 singletons (Höps et al., 2018). There are currently no methods available to provide a plausible
526 solution and, at the same time, being scalable. We devise a potential solution in the recent
527 developments in unsupervised deep learning methods where they use large corpora of proteins
528 to define a language model *embedding* for protein sequences (Heinzinger et al., 2019). These
529 models could be applied to predict *embeddings* in singletons, which could be clustered or used
530 to determine their coding potential. Another issue is that we might be creating more GCs than
531 expected. We follow a conservative approach to avoid mixing multi-domain proteins in GCs
532 owing to the fragmented nature of the metagenome assemblies that could result in the split of a
533 GC. However, not only splitting can be a problem, but also lumping unrelated genes or GCs
534 owing to the use of remote homologies. Although the inference of GCCs is using very sensitive
535 methods to compare profile HMMs, low sequence diversity in GCs can limit its effectiveness.
536
537 Moreover, our approach is affected by the presence and propagation of contamination in
538 reference databases, a significant problem in 'omics (Breitwieser et al., 2019; Steinegger and
539 Salzberg, 2020). In our case, we only use Pfam (Finn et al., 2016) as a source for annotation
540 owing to its high-quality and manual curation process. The categorization process of our GCs
541 depends on the information from other databases, and to minimize the potential impact of
542 contamination, we apply methods that weight the annotations of the identified homologs to
543 discriminate if a GC belongs to the known or unknown sequence space.

544 The results presented here prove that the integration and the analysis of the unknown fraction
545 are possible. We are unveiling a brighter future, not only for microbiome analyses but also for
546 boosting eukaryotic-related studies, thanks to the increasing number of projects, including
547 metatranscriptomic data (Delmont et al., 2020; Vorobev et al., 2020). Furthermore, our work
548 lays the foundations for further developments of clear guidelines and protocols to define the
549 different levels of unknown (Thomas and Segata, 2019) and should encourage the scientific
550 community for a collaborative effort to tackle this challenge.

551 Material and methods

552 Genomic and metagenomic dataset

553 We used a set of 583 marine metagenomes from four of the major metagenomic surveys of the
554 ocean microbiome: Tara Oceans expedition (TARA) (Sunagawa et al., 2015), Malaspina
555 expedition (Duarte, 2015), Ocean Sampling Day (OSD) (Kopf et al., 2015), and Global Ocean
556 Sampling Expedition (GOS) (Rusch et al., 2007). We complemented this set with 1,246
557 metagenomes obtained from the Human Microbiome Project (HMP) phase I and II (Lloyd-Price
558 et al., 2017). We used the assemblies provided by TARA, Malaspina, OSD and HMP projects
559 and the long Sanger reads from GOS (Sanger et al., 1977). A total of 156M (156,422,969)
560 contigs and 12.8M long-reads were collected (Supp. Table 6).
561 For the genomic dataset, we used the 28,941 prokaryotic genomes (27,372 bacterial and 1,569
562 archaeal) from the Genome Taxonomy Database (Parks et al., 2018) (GTDB) Release 03-RS86
563 (19th August 2018).

564 ## Computational workflow development

565 We implemented a computation workflow based on Snakemake (Köster, 2018) for the easy
566 processing of large datasets in a reproducible manner. The workflow provides three different
567 strategies to analyze the data. The module *DB-creation* creates the gene cluster database,
568 validates and partitions the gene clusters (GCs) in the main functional categories. The module
569 *DB-update* allows the integration of new sequences (either at the contig or predicted gene level)
570 in the existing gene cluster database. In addition, the workflow has a *profile-search* function to
571 quickly screen samples using the gene cluster PSSM profiles in the database.

572 ## Metagenomic and genomic gene prediction

573 We used Prodigal (v2.6.3) (Hyatt et al., 2010) in metagenomic mode to predict the genes from
574 the metagenomic dataset. For the genomic dataset, we used the gene predictions provided by
575 Annotree (Mendler et al., 2019), since they were obtained, consistently, with Prodigal v2.6.3.
576 We identified potential spurious genes using the *AntiFam* database (Eberhardt et al., 2012).
577 Furthermore, we screened for '*shadow*' genes using the procedure described in Yooseph et al.
578 (Yooseph et al., 2008).

579 ## PFAM annotation

580 We annotated the predicted genes using the *hmmsearch* program from the *HMMER* package
581 (version: 3.1b2) (Finn et al., 2011) in combination with the Pfam database v31 (Finn et al.,
582 2016). We kept the matches exceeding the internal gathering threshold and presenting an
583 independent e-value < 1e-5 and coverage > 0.4. In addition, we considered multi-domain
584 annotations, and we removed overlapping annotations when the overlap is larger than 50%,
585 keeping the ones with the smaller e-value.

586 ## Determination of the gene clusters

587 We clustered the metagenomic predicted genes using the cascaded-clustering workflow of the
588 MMseqs2 software (Steinegger and Söding, 2018) ("*--cov-mode 0 -c 0.8 --min-seq-id 0.3"*). We
589 discarded from downstream analyses the singletons and clusters with a size below a threshold
590 identified after applying a modification of the broken-stick model (Macarthur, 1957). We
591 randomly split the number of gene clusters into p subsets, where p is defined by the proportion
592 of outlier genes per gene cluster. The subsets are then sorted by decreasing size. We iterated
593 over all subsets averaging the results over all iterations. The broken stick model generates the
594 outlier gene proportions, which would occur by chance alone, that is, the distribution of outlier
595 gene proportions if there were no structure in the data.
596 We integrated the genomic data into the metagenomic cluster database using the "DB-update"
597 module of the workflow. This module uses the *clusterupdate* module of MMseqs2 (Steinegger
598 and Soding, 2017), with the same parameters used for the metagenomic clustering.

599 ## Quality-screening of gene clusters

600 We examined the GCs to ensure their high intra-cluster homogeneity. We applied two
601 methodologies to validate their cluster sequence composition and functional annotation
602 homogeneity. We identified non-homologous sequences inside each cluster combining the
603 identification of a new cluster representative sequence via a sequence similarity network (SSN)
604 analysis, and the investigation of intra-cluster multiple sequence alignments (MSAs), given the
605 new representative. Initially, we generated an SSN for each cluster, using the semi-global
606 alignment methods implemented in *PARASAIL* (Daily, 2016) (version 2.1.5). We trimmed the
607 SSN using a filtering algorithm (Chafee et al., 2018; Žure et al., 2017) that removes edges while
608 maintaining the network structural integrity and obtaining the smallest connected graph formed
609 by a single component. Finally, the new cluster representative was identified as the most central

610 node of the trimmed SSN by the eigenvector centrality algorithm, as implemented in igraph
611 (Csardi and Nepusz, 2006). After this step, we built a multiple sequence alignment for each
612 cluster using *FAMSA* (Deorowicz et al., 2016) (version 1.1). Then, we screened each cluster-
613 MSA for non-homologous sequences to the new cluster representative. Owing to computational
614 limitations, we used two different approaches to evaluate the cluster-MSAs. We used *LEON-BIS*
615 (Vanhoutreve et al., 2016) for the clusters with a size ranging from 10 to 1,000 genes and OD-
616 SEQ (Jehl et al., 2015) for the clusters with more than 1,000 genes. In the end, we applied a
617 broken-stick model (Macarthur, 1957) to determine the threshold to discard a cluster.
618 The predicted genes can have multi-domain annotations in different orders, therefore to validate
619 the consistency of intra-cluster Pfam annotations, we applied a combination of w-shingling
620 (Broder, 1997) and Jaccard similarity. We used w-shingling (k-shingle = 2) to group consecutive
621 domain annotations as a single object. We measured the homogeneity of the *shingle sets* (sets
622 of domains) between genes using the Jaccard similarity and reported the median similarity
623 value for each cluster. Moreover, we took into consideration the Clan membership of the Pfam
624 domains and that a gene might contain N-, C- and M-terminal domains for the functional
625 homogeneity validation. We discarded clusters with a median similarity < 1.
626 After the validation, we refined the gene cluster database removing the clusters identified to be
627 discarded and the clusters containing ≥ 30% *shadow genes*. Lastly, we removed the single
628 shadow, spurious and non-homologous genes from the remaining clusters (Supp. Note 2).

629 Remote homology classification of gene clusters

630 To partition the validated GCs into the four main categories, we processed the set of GCs
631 containing Pfam annotated genes and the set of not annotated GCs separately. For the
632 annotated GCs, we inferred a consensus protein domain architecture (DA) (an ordered
633 combination of protein domains) for each annotated gene cluster. To identify each gene cluster
634 consensus DA, we created directed acyclic graphs connecting the Pfam domains based on their

635 topological order on the genes using *igraph* (Csardi and Nepusz, 2006). We collapsed the
636 repetitions of the same domain. Then we used the gene completeness as a positive-weighting
637 value for the selection of the cluster consensus DA. Within this step, we divided the GCs into
638 "Knowns" (Known) if annotated to at least one Pfam domains of known function (DKFs) and
639 "Genomic unknowns" (GU) if annotated entirely to Pfam domains of unknown function (DUFs).
640 We aligned the sequences of the non-annotated GCs with FAMSA (Deorowicz et al., 2016) and
641 obtained cluster consensus sequences with the *hhconsensus* program from *HH-SUITE*
642 (Steinegger et al., 2019a). We used the cluster consensus sequences to perform a nested
643 search against the UniRef90 database (release 2017_11) (The UniProt Consortium, 2017) and
644 NCBI *nr* database (release 2017_12) (NCBI Resource Coordinators, 2018) to retrieve non-Pfam
645 annotations with *MMSeqs2* (Steinegger and Soding, 2017) ("*-e 1e-05 --cov-mode 2 -c 0.6"*). We
646 kept the hits within 60% of the Log(best-e-value) and searched the annotations for any of the
647 terms commonly used to define proteins of unknown function (Supp. Table 12). We used a
648 quorum majority voting approach to decide if a gene cluster would be classified as *Genomic*
649 *Unknown* or *Known without Pfams* based on the annotations retrieved. We searched the
650 consensus sequences without any homologs in the UniRef90 database against NCBI *nr*. We
651 applied the same approach and criteria described for the first search. Ultimately, we classified
652 as *Environmental Unknown* those GCs whose consensus sequences did not align with any of
653 the NCBI *nr* entries.
654 In addition, we developed some conservative measures to control the trade-off between
655 specificity and sensitivity for the remote homology searches such as (1) a modification of the
656 algorithm described in Hingamp et al. (Hingamp et al., 2013) to get a confident group of
657 homologs to determine if a query protein is known or unknown by a quorum majority voting
658 approach (Supp. Note 3); (2) strict parameters in terms of iterations, bidirectional coverage and
659 probability thresholds for the HHblits alignments to minimize the inclusion of non-homologous

660 sequences; and (3) avoid providing annotations for our gene clusters, as we believe that
661 annotation should be a careful process done on a smaller scale and with experimental context.

662 Gene cluster remote homology refinement

663 We refined the *Environmental Unknown* GCs to ensure the lack of any characterization by
664 searching for remote homologies in the Uniclust database (release 30_2017_10) using the
665 HMM/HMM alignment method *HHblits* (Remmert et al., 2012). We created the HMM profiles
666 with the *hhmake* program from the *HH-SUITE* (Steinegger et al., 2019a). We only accepted
667 those hits with an *HHblits-probability* ≥ 90% and we re-classified them following the same
668 majority vote approach as previously described. The clusters with no hits remained as the
669 refined set of EUs. We applied a similar refinement approach to the KWP clusters to identify
670 GCs with remote homologies to Pfam protein domains. The KWP HMM profiles were searched
671 against the Pfam *HH-SUITE* database (version 31), using *HHblits*. We accepted hits with a
672 probability ≥ 90% and a target coverage > 60% and removed overlapping domains as described
673 earlier. We moved the KWP with remote homologies to known Pfams to the Known set, and
674 those showing remote homologies to Pfam DUFs to the GUs. The clusters with no hits remained
675 as the refined set of KWP.

676 Gene cluster characterization

677 We used the *MMseqs2 taxonomy* module (commit:
678 b43de8b7559a3b45c8e5e9e02cb3023dd339231a) in combination with the UniProtKB (release
679 of January 2018) (UniProt Consortium, 2018) to retrieve the taxonomic ids of all genes in a gene
680 cluster. The *taxonomy* module implements the 2bLCA (Hingamp et al., 2013) to compute the
681 lowest common ancestor of query sequence. We used the following parameters "*-e 1e-05 –cov-*
682 *mode 0 -c 0.6"* for the search. To retrieve the taxonomic lineages, we used the R package
683 *CHNOSZ* (Dick, 2008).

684 We used eggNOG-mapper (Huerta-Cepas et al., 2017) and the EggNog5 database (Huerta-
685 Cepas et al., 2019) to provide functional annotations for each gene in a gene cluster. We refined
686 the functional annotations by selecting the orthologous group within the lowest taxonomic level
687 predicted by EggNog-mapper.
688 We measured the intra-cluster taxonomic and functional admixture by applying the
689 *entropy.empirical()* function from the *entropy* R package (Hausser and Strimmer, 2008). This
690 function estimates the Shannon entropy based on the different taxonomic and functional
691 annotation frequencies. For each cluster, we also retrieved the cluster consensus taxonomic
692 and functional annotation using a quorum majority voting approach..
693 In addition to the taxonomic and functional annotations, we evaluated the clusters' level of
694 darkness and disorder using the Dark Proteome Database (DPD) (Perdigão et al., 2017) as
695 reference. We searched the cluster genes against the DPD, applying the MMseqs2 search
696 program (Steinegger and Soding, 2017) with "*-e 1e-20 --cov-mode 0 -c 0.6*". For each cluster,
697 we then retrieved the mean and the median level of darkness, based on the gene DPD
698 annotations.

699 High-quality clusters

700 We defined a subset of high-quality clusters based on the completeness of the cluster genes
701 and their representatives. We identified the minimum required percentage of complete genes
702 per cluster by a broken-stick model (Macarthur, 1957) applied to the percentage distribution.
703 Then, we selected the GCs found above the threshold and with a complete representative.

704 A set of non-redundant domain architectures

705 We estimated the number of potential domain architectures present in the *Known* GCs
706 considering the large proportion of fragmented genes in the metagenomic dataset and that
707 could inflate the number of potential domain architectures. To identify fragments of larger

708 domain architecture, we considered their topological order in the genes. To reduce the number
709 of comparisons, we calculated the pairwise string cosine distance (q-gram = 3) between domain
710 architectures and discarded the pairs that were too divergent (cosine distance ≥ 0.9). We
711 collapsed a fragmented domain architecture to the larger one when it contained less than 75%
712 of complete genes.

713 Inference of gene cluster communities

714 We aggregated distant homologous GCs into GCCs. The community inference approach
715 combined an all-vs-all HMM gene cluster comparison with Markov Cluster Algorithm (MCL) (van
716 Dongen and Abreu-Goodger, 2012) community identification. We started performing the
717 inference on the Known GCs to use the Pfam DAs as constraints. We aligned the gene cluster
718 HMMs using HHblits (Remmert et al., 2012) (-n 2 -Z 10000000 -B 10000000 -e 1) and we built a
719 homology graph using the cluster pairs with probability ≥ 50% and bidirectional coverage > 60%.
720 We used the ratio between HHblits-bitscore and aligned-columns as the edge weights (Supp.
721 Note 9). We used MCL (van Dongen and Abreu-Goodger, 2012) (v. 12-068) to identify the
722 communities present in the graph. We developed an iterative method to determine the optimal
723 MCL inflation parameter that tries to maximize the relationship of five intra-/inter-community
724 properties: (1) the proportion of MCL communities with one single DA, based on the consensus
725 DAs of the cluster members; (2) the ratio of MCL communities with more than one cluster; (3)
726 the proportion of MCL communities with a PFAM clan entropy equal to 0; (4) the intra-
727 community HHblits-score/Aligned-columns score (normalized by the maximum value); and (5)
728 the number of MCL communities, which should, in the end, reflect the number of non-redundant
729 DAs. We iterated through values ranging from 1.2 to 3.0, with incremental steps of 0.1. During
730 the inference process, some of the GCs became orphans in the graph. We applied a three-step
731 approach to assigning a community membership to these GCs. First, we used less stringent
732 conditions (probability ≥ 50% and coverage >= 40%) to find homologs in the already existing

733 GCCs. Then, we ran a second iteration to find secondary relationships between the newly
734 assigned GCs and the missing ones. Lastly, we created new communities with the remaining
735 GCs. We repeated the whole process with the other categories (KWP, GU and EU), applying
736 the optimal inflation value found for the Known (2.2 for metagenomic and 2.5 for genomic data).

737 Validation of gene cluster communities

738 We tested the biological significance of the GCCs using the phylogeny of proteorhodopsin
739 (Boeuf et al., 2015) (PR). We used the proteorhodopsin HMM profiles (Olson et al., 2018) to
740 screen the marine metagenomic datasets using *hmmsearch* (version 3.1b2) (Finn et al., 2011).
741 We kept the hits with a coverage > 0.4 and e-value <= 1e-5. We removed identical duplicates
742 from the sequences assigned to PR with CD-HIT (Li and Godzik, 2006) (v4.6) and cleaned from
743 sequences with less than 100 amino acids. To place the identified PR sequences into the
744 MicRhode (Boeuf et al., 2015) PR tree first, we optimized the initial tree parameters and branch
745 lengths with RAxML (v8.2.12) (Stamatakis, 2014). We used PaPaRA (v2.5) (Berger and
746 Stamatakis, 2012) to incrementally align the query PR sequences against the MicRhode PR
747 reference alignment and *pplacer* (Matsen et al., 2010) (v1.1.alpha19-0-g807f6f3) to place the
748 sequences into the tree. Finally, we assigned the query PR sequences to the MicRhode PR
749 Superclusters based on the phylogenetic placement. We further investigated the GCs annotated
750 as viral (196 genes, 14 GC) comparing them to the six newly discovered viral PRs (Needham et
751 al., 2019) using Parasail (Daily, 2016) (-a sg_stats_scan_sse2_128_16 -t 8 -c 1 -x). As an
752 additional evaluation, we investigated the distributions of standard GCCs and HQ GCCs within
753 ribosomal protein families. We obtained the ribosomal proteins used for the analysis combining
754 the set of 16 ribosomal proteins from Méheust et al. (Méheust et al., 2019) and those contained
755 in the collection of bacterial single-copy genes of anvi'o (Eren et al., 2021). Also, for the
756 ribosomal proteins, we compared the outcome of our method to the one proposed by Méheust
757 et al. (Méheust et al., 2019) (Supp. Note 9).

758 ## Metagenomic sample selection for downstream analyses

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

762 ## Gene cluster abundance profiles in genomes and metagenomes

763 We estimated abundance profiles for the metagenomic cluster categories using the read
764 coverage to each predicted gene as a proxy for abundance. We calculated the coverage by
765 mapping the reads against the assembly contigs using the *bwa-mem* algorithm from *BWA*
766 *mapper* (Li and Durbin, 2010). Then, we used *BEDTOOLS* (Quinlan and Hall, 2010), to find the
767 intersection of the gene coordinates to the assemblies, and normalize the per-base coverage by
768 the length of the gene. We calculated the cluster abundance in a sample as the sum of the
769 cluster gene abundances in that sample, and the cluster category abundance in a sample as the
770 sum of the cluster abundances. We obtained the proportions of the different gene cluster
771 categories applying a total-sum-scaling normalization. For the genomic abundance profiles, we
772 used the number of genes in the genomes and normalized by the total gene counts per
773 genome.

774 ## Rate of genomic and metagenomic gene clusters accumulation

775 We calculated the cumulative number of known and unknown GCs as a function of the number
776 of metagenomes and genomes. For each metagenome count, we generated 1000 random sets,
777 and we calculated the number of GCs and GCCs recovered. For this analysis, we used 1,246
778 HMP metagenomes and 358 marine metagenomes (242 from TARA and 116 from Malaspina).
779 We repeated the same procedure for the genomic dataset. We removed the singletons from the
780 metagenomic dataset with an abundance smaller than the mode abundance of the singletons

781 that got reclassified as good-quality clusters after integrating the GTDB data to minimize the
782 impact of potential spurious singletons. To complement those analyses, we evaluated the
783 coverage of our dataset by searching seven different state-of-the-art databases against our set
784 of metagenomic GC HMM profiles (Supp. Note 11).
785

786 Occurrence of gene clusters in the environment

787 We used 1,264 metagenomes from the TARA Oceans, MALASPINA Expedition, OSD2014 and
788 HMP-I/II to explore the properties of the unknown sequence space in the environment. We
789 applied the Levins Niche Breadth (NB) index (Levins, 1966) to investigate the GCs and GCCs
790 environmental distributions. We removed the GCs and cluster communities with a mean relative
791 abundance < 1e-5. We followed a divide-and-conquer strategy to avoid the computational
792 burden of generating the null-models to test the significance of the distributions owing to the
793 large number of metagenomes and GCs. First, we grouped similar samples based on the gene
794 cluster content using the Bray-Curtis dissimilarity (Bray et al., 1957) in combination with the
795 *Dynamic Tree Cut* (Langfelder et al., 2008) R package. We created 100 random datasets
796 picking up one random sample from each group. For each of the 100 random datasets, we
797 created 100 random abundance matrices using the *nullmodel* function of the *quasiswap* count
798 method (Miklós and Podani, 2004). Then we calculated the *observed* NB and obtained the 2.5%
799 and 97.5% quantiles based on the randomized sets. We compared the observed and quantile
800 values for each gene cluster and defined it to have a *Narrow distribution* when the *observed*
801 was smaller than the 2.5% quantile and to have a *Broad distribution* when it was larger than the
802 97.5% quantile. Otherwise, we classified the cluster as *Non-significant* (Salazar et al., 2015).
803 We used a majority voting approach to get a consensus distribution classification based on the
804 ten random datasets.

Identification of prophages in genomic sequences

We used PhageBoost (Sirén et al., 2021) to find gene regions in the microbial genomes that result in high viral signals against the overall genome signal. We set the following thresholds to consider a region prophage: minimum of 10 genes, maximum 5 gaps, single-gene probability threshold 0.9. We further smoothed the predictions using Parzen rolling windows of 20 periods and looked at the smoothed probability distribution across the genome. We disregarded regions that had a summed smoothed probability less than 0.5, and those regions that did differ from the overall population of the genes in a genome by using Kruskal–Wallis rank test (p-value 0.001).

Lineage-specific gene clusters

We used the F1-score developed for AnnoTree (Mendler et al., 2019) to identify the lineage-specific GCs and to which rank they are specific. Following similar criteria to the ones used in Mendler et al. (Mendler et al., 2019), we considered a gene cluster to be lineage-specific if it is present in less than half of all genomes 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* (Martiny et al., 2013) function implemented in the R package *castor* (Martiny et al., 2013). 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) (Salazar et al., 2019) into our cluster database using the same procedure as for the genomic data. We evaluated the uncharacterized fraction and the

827 genes classified into the eggNOG (Huerta-Cepas et al., 2019) category S within the context of
828 our database.

829 Augmenting RB-TnSeq experimental data

830 We searched the 37,684 genes of unknown function associated with mutant phenotypes from
831 Price et al. (Price et al., 2018) against our gene cluster profiles. We kept the hits with e-value ≤
832 1e-20 and a query coverage > 60%. Then we filtered the results to keep the hits within 90% of
833 the Log(best-e-value), and we used a majority vote function to retrieve the consensus category
834 for each hit. Lastly, we selected the best-hits based on the smallest e-value and the largest
835 query and target coverage values. We used the fitness values from the RB-TnSeq experiments
836 from Price et al. to identify genes of unknown function that are important for fitness under
837 certain experimental conditions.

838 Availability of data and materials

839 The code used for the analyses in the manuscript is available at https://github.com/functional-
840 dark-side/functional-dark-side.github.io/tree/master/scripts. The code to recreate the figures is
841 available at https://github.com/functional-dark-side/vanni_et_al-figures. Detailed descriptions of
842 the different methods and results of this manuscript are available at
843 https://dark.metagenomics.eu. The workflow AGNOSTOS is available at
844 https://github.com/functional-dark-side/agnostos-wf (the version used for the manuscript can be
845 found here https://doi.org/10.5281/zenodo.4557847) and its database can be downloaded from
846 https://doi.org/10.6084/m9.figshare.12459056.
847
848

Acknowledgements

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.

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

Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, Lawley TD, Finn RD. 2019. A new genomic blueprint of the human gut microbiota. *Nature* **568**:499–504.

Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E, Parks DH, Hugenholtz P, Segata N, Kyrpides NC, Finn RD. 2020. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol*. doi:10.1038/s41587-020-0603-3

Anantharaman K, Hausmann B, Jungbluth SP, Kantor RS, Lavy A, Warren LA, Rappé MS, Pester M, Loy A, Thomas BC, Banfield JF. 2018. Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. *ISME J* **12**:1715–1728.

Arnold FH. 2018. Directed Evolution: Bringing New Chemistry to Life. *Angew Chem Int Ed Engl* **57**:4143–4148.

Arnold FH. 1998. Design by Directed Evolution. *Acc Chem Res* **31**:125–131.

Bateman A, Coggill P, Finn DR. 2010. DUFs: Families in search of function. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **66**:1148–1152.

Berger SA, Stamatakis A. 2012. 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*.

Bernard G, Pathmanathan JS, Lannes R, Lopez P, Bapteste E. 2018. Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. *Genome Biol Evol* **10**:707–715.

Bileschi ML, Belanger D, Bryant D, Sanderson T, Carter B, Sculley D, DePristo MA, Colwell LJ. 2019. Using Deep Learning to Annotate the Protein Universe. *bioRxiv*. doi:10.1101/626507

Bitard-Feildel T, Callebaut I. 2017. Exploring the dark foldable proteome by considering hydrophobic amino acids topology. *Sci Rep* **7**:41425.

Boeuf D, Audic S, Brillet-Guéguen L, Caron C, Jeanthon C. 2015. MicRhoDE: a curated database for the analysis of microbial rhodopsin diversity and evolution. *Database* **2015**. doi:10.1093/database/bav080

Brandenberg OF, Fasan R, Arnold FH. 2017. Exploiting and engineering hemoproteins for abiological carbene and nitrene transfer reactions. *Curr Opin Biotechnol* **47**:102–111.

Bray JR, Roger Bray J, Curtis JT. 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*. doi:10.2307/1942268

Breitwieser FP, Pertea M, Zimin A, Salzberg SL. 2019. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res*. doi:10.1101/gr.245373.118

Broder AZ. 1997. On the resemblance and containment of documentsProceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171). IEEE. pp. 21–29.

Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**:208–211.

Brown CT, Moritz D, O'Brien MP, Reidl F, Reiter T, Sullivan BD. 2020. Exploring neighborhoods in large metagenome assembly graphs using spacegraphcats reveals hidden sequence diversity. *Genome Biol* **21**:164.

Brum JR, Cesar Ignacio-Espinoza J, Kim E-H, Trubl G, Jones RM, Roux S, VerBerkmoes NC, Rich VI, Sullivan MB. 2016. Illuminating structural proteins in viral "dark matter" with metaproteomics. *Proc Natl Acad Sci U S A* **113**:2436–2441.

Buttigieg LP, Hankeln W, Kostadinov I, Kottmann R, Yilmaz P, Duhaime BM, Gl??ckner OF.

2013. Ecogenomic Perspectives on Domains of Unknown Function: Correlation-Based Exploration of Marine Metagenomes. *PLoS One* **8**.

Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R, Lima-Mendez G, Rocha F, Tirichine L, Labadie K, Kirilovsky A, Bertrand A, Engelen S, Madoui M-A, Méheust R, Poulain J, Romac S, Richter DJ, Yoshikawa G, Dimier C, Kandels-Lewis S, Picheral M, Searson S, Tara Oceans Coordinators, Jaillon O, Aury J-M, Karsenti E, Sullivan MB, Sunagawa S, Bork P, Not F, Hingamp P, Raes J, Guidi L, Ogata H, de Vargas C, Iudicone D, Bowler C, Wincker P. 2018. A global ocean atlas of eukaryotic genes. *Nat Commun* **9**:373.

Chafee M, Fernàndez-Guerra A, Buttigieg PL, Gerdts G, Eren AM, Teeling H, Amann RI. 2018. Recurrent patterns of microdiversity in a temperate coastal marine environment. *ISME J* **12**:237–252.

Chen I-MA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, Huntemann M, Varghese N, White JR, Seshadri R, Smirnova T, Kirton E, Jungbluth SP, Woyke T, Eloe-Fadrosh EA, Ivanova NN, Kyrpides NC. 2019. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res* **47**:D666–D677.

Cross KL, Campbell JH, Balachandran M, Campbell AG, Cooper SJ, Griffen A, Heaton M, Joshi S, Klingeman D, Leys E, Yang Z, Parks JM, Podar M. 2019. Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. *Nat Biotechnol* **37**:1314–1321.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal*.

Daily J. 2016. Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. *BMC Bioinformatics* **17**:81–81.

Delmont TO, Gaia M, Hinsinger DD, Fremont P, Guerra AF, Murat Eren A, Vanni C, Kourlaiev A, d'Agata L, Clayssen Q, Villar E, Labadie K, Cruaud C, Poulain J, Da Silva C, Wessner M, Noel B, Aury J-M, Coordinators TO, de Vargas C, Bowler C, Karsenti E, Pelletier E, Wincker P, Jaillon O. 2020. Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. *Cold Spring Harbor Laboratory*. doi:10.1101/2020.10.15.341214

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

Dick JM. 2008. Calculation of the relative metastabilities of proteins using the CHNOSZ software package. *Geochem Trans* **9**:10.

Duarte CM. 2015. Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation Expedition. *Limnol Oceanog Bull* **24**:11–14.

Dubinkina VB, Ischenko DS, Ulyantsev VI, Tyakht AV, Alexeev DG. 2016. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC Bioinformatics*. doi:10.1186/s12859-015-0875-7

Eberhardt RY, Haft DH, Punta M, Martin M, O'Donovan C, Bateman A. 2012. AntiFam: a tool to help identify spurious ORFs in protein annotation. *Database* **2012**:bas003–bas003.

Edwards RA, Vega AA, Norman HM, Ohaeri M, Levi K, Dinsdale EA, Cinek O, Aziz RK, McNair K, Barr JJ, Bibby K, Brouns SJJ, Cazares A, de Jonge PA, Desnues C, Díaz Muñoz SL, Fineran PC, Kurilshikov A, Lavigne R, Mazankova K, McCarthy DT, Nobrega FL, Reyes Muñoz A, Tapia G, Trefault N, Tyakht AV, Vinuesa P, Wagemans J, Zhernakova A, Aarestrup FM, Ahmadov G, Alassaf A, Anton J, Asangba A, Billings EK, Cantu VA, Carlton JM, Cazares D, Cho G-S, Condeff T, Cortés P, Cranfield M, Cuevas DA, De la Iglesia R, Decewicz P, Doane MP, Dominy NJ, Dziewit L, Elwasila BM, Eren AM, Franz C, Fu J, Garcia-Aljaro C, Ghedin E, Gulino KM, Haggerty JM, Head SR, Hendriksen RS, Hill C, Hyöty H, Ilina EN, Irwin MT, Jeffries TC, Jofre J, Junge RE, Kelley ST, Khan Mirzaei M, Kowalewski M, Kumaresan D, Leigh SR, Lipson D, Lisitsyna ES, Llagostera

979 M, Maritz JM, Marr LC, McCann A, Molshanski-Mor S, Monteiro S, Moreira-Grez B,
980 Morris M, Mugisha L, Muniesa M, Neve H, Nguyen N-P, Nigro OD, Nilsson AS,
981 O'Connell T, Odeh R, Oliver A, Piuri M, Prussin AJ Ii, Qimron U, Quan Z-X, Rainetova P,
982 Ramírez-Rojas A, Raya R, Reasor K, Rice GAO, Rossi A, Santos R, Shimashita J,
983 Stachler EN, Stene LC, Strain R, Stumpf R, Torres PJ, Twaddle A, Ugochi Ibekwe M,
984 Villagra N, Wandro S, White B, Whiteley A, Whiteson KL, Wijmenga C, Zambrano MM,
985 Zschach H, Dutilh BE. 2019. Global phylogeography and ancient evolution of the
986 widespread human gut virus crAssphage. *Nat Microbiol* **4**:1727–1736.
987 Eloe-Fadrosh EA, Paez-Espino D, Jarett J, Dunfield PF, Hedlund BP, Dekas AE, Grasby SE,
988 Brady AL, Dong H, Briggs BR, Li W-J, Goudeau D, Malmstrom R, Pati A, Pett-Ridge J,
989 Rubin EM, Woyke T, Kyrpides NC, Ivanova NN. 2016. Global metagenomic survey
990 reveals a new bacterial candidate phylum in geothermal springs. *Nat Commun* **7**:10476.
991 Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, Fink I, Pan JN, Yousef M,
992 Fogarty EC, Trigodet F, Watson AR, Esen ÖC, Moore RM, Clayssen Q, Lee MD,
993 Kivenson V, Graham ED, Merrill BD, Karkman A, Blankenberg D, Eppley JM, Sjödin A,
994 Scott JJ, Vázquez-Campos X, McKay LJ, McDaniel EA, Stevens SLR, Anderson RE,
995 Fuessel J, Fernandez-Guerra A, Maignien L, Delmont TO, Willis AD. 2021. Community-
996 led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* **6**:3–6.
997 Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity
998 searching. *Nucleic Acids Res* **39**:W29–W37.
999 Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi
1000 M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein
1001 families database: towards a more sustainable future. *Nucleic Acids Res* **44**:D279–
1002 D285.
1003 Francino MP. 2012. The ecology of bacterial genes and the survival of the new. *Int J Evol Biol*
1004 **2012**:394026.
1005 Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Lipson KS,
1006 Knight R, Caporaso JG, Segata N, Huttenhower C. 2018. Species-level functional
1007 profiling of metagenomes and metatranscriptomes. *Nat Methods* **15**:962–968.
1008 Habchi J, Tompa P, Longhi S, Uversky VN. 2014. Introducing protein intrinsic disorder. *Chem*
1009 *Rev* **114**:6561–6588.
1010 Hanson AD, Pribat A, Waller JC, Crécy-Lagard V de. 2010. 'Unknown'proteins and
1011 'orphan'enzymes: the missing half of the engineering parts list--and how to find it.
1012 *Biochem J* **425**:1–11.
1013 Hausser J, Strimmer K. 2008. Entropy inference and the James-Stein estimator, with application
1014 to nonlinear gene association networks. *arXiv [statML]*.
1015 Heffernan B, Murphy CD, Casey E. 2009. Comparison of planktonic and biofilm cultures of
1016 Pseudomonas fluorescens DSM 8341 cells grown on fluoroacetate. *Appl Environ*
1017 *Microbiol* **75**:2899–2907.
1018 Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, Rost B. 2019. Modeling
1019 aspects of the language of life through transfer-learning protein sequences. *BMC*
1020 *Bioinformatics* **20**:723.
1021 Hingamp P, Grimsley N, Acinas SG, Clerissi C, Subirana L, Poulain J, Ferrera I, Sarmento H,
1022 Villar E, Lima-Mendez G, Faust K, Sunagawa S, Claverie J-M, Moreau H, Desdevises Y,
1023 Bork P, Raes J, de Vargas C, Karsenti E, Kandels-Lewis S, Jaillon O, Not F, Pesant S,
1024 Wincker P, Ogata H. 2013. Exploring nucleo-cytoplasmic large DNA viruses in Tara
1025 Oceans microbial metagenomes. *ISME J* **7**:1678–1695.
1026 Höps W, Jeffryes M, Bateman A. 2018. Gene Unprediction with Spurio: A tool to identify
1027 spurious protein sequences. *F1000Res* **7**:261.
1028 Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. 2017.
1029 Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-

1030 Mapper. *Mol Biol Evol* **34**:2115–2122.
1031 Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR,
1032 Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. 2019. eggNOG 5.0: a hierarchical,
1033 functionally and phylogenetically annotated orthology resource based on 5090
1034 organisms and 2502 viruses. *Nucleic Acids Res* **47**:D309–D314.
1035 Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN,
1036 Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM,
1037 Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. *Nat*
1038 *Microbiol* **1**:16048.
1039 Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic
1040 gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**:119–
1041 119.
1042 Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, Visel A, Woyke T,
1043 Kyrpides NC, Rubin EM. 2014. Stop codon reassignments in the wild. *Science* **344**:909–
1044 913.
1045 Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, Deacon MA, Wilson AI, Godzik A.
1046 2009. Exploration of uncharted regions of the protein universe. *PLoS Biol* **7**.
1047 Jehl P, Sievers F, Higgins DG. 2015. OD-seq: outlier detection in multiple sequence alignments.
1048 *BMC Bioinformatics* **16**:269–269.
1049 Kopf A, Bicak M, Kottmann R, Schnetzer J, Kostadinov I, Lehmann K, Fernandez-Guerra A,
1050 Jeanthon C, Rahav E, Ullrich M, Wichels A, Gerdts G, Polymenakou P, Kotoulas G,
1051 Siam R, Abdallah RZ, Sonnenschein EC, Cariou T, O'Gara F, Jackson S, Orlic S,
1052 Steinke M, Busch J, Duarte B, Caçador I, Canning-Clode J, Bobrova O, Marteinsson V,
1053 Reynisson E, Loureiro CM, Luna GM, Quero GM, Löscher CR, Kremp A, DeLorenzo
1054 ME, Øvreås L, Tolman J, LaRoche J, Penna A, Frischer M, Davis T, Katherine B, Meyer
1055 CP, Ramos S, Magalhães C, Jude-Lemeilleur F, Aguirre-Macedo ML, Wang S, Poulton
1056 N, Jones S, Collin R, Fuhrman JA, Conan P, Alonso C, Stambler N, Goodwin K,
1057 Yakimov MM, Baltar F, Bodrossy L, Van De Kamp J, Frampton DM, Ostrowski M, Van
1058 Ruth P, Malthouse P, Claus S, Deneudt K, Mortelmans J, Pitois S, Wallom D, Salter I,
1059 Costa R, Schroeder DC, Kandil MM, Amaral V, Biancalana F, Santana R, Pedrotti ML,
1060 Yoshida T, Ogata H, Ingleton T, Munnik K, Rodriguez-Ezpeleta N, Berteaux-Lecellier V,
1061 Wecker P, Cancio I, Vaulot D, Bienhold C, Ghazal H, Chaouni B, Essayeh S, Ettamimi
1062 S, Zaid EH, Boukhatem N, Bouali A, Chahboune R, Barrijal S, Timinouni M, El Otmani F,
1063 Bennani M, Mea M, Todorova N, Karamfilov V, Ten Hoopen P, Cochrane G, L'Haridon
1064 S, Bizsel KC, Vezzi A, Lauro FM, Martin P, Jensen RM, Hinks J, Gebbels S, Rosselli R,
1065 De Pascale F, Schiavon R, Dos Santos A, Villar E, Pesant S, Cataletto B, Malfatti F,
1066 Edirisinghe R, Silveira JAH, Barbier M, Turk V, Tinta T, Fuller WJ, Salihoglu I, Serakinci
1067 N, Ergoren MC, Bresnan E, Iriberri J, Nyhus PAF, Bente E, Karlsen HE, Golyshin PN,
1068 Gasol JM, Moncheva S, Dzhembekova N, Johnson Z, Sinigalliano CD, Gidley ML,
1069 Zingone A, Danovaro R, Tsiamis G, Clark MS, Costa AC, El Bour M, Martins AM, Collins
1070 RE, Ducluzeau A-L, Martinez J, Costello MJ, Amaral-Zettler LA, Gilbert JA, Davies N,
1071 Field D, Glöckner FO. 2015. The ocean sampling day consortium. *Gigascience* **4**:27.
1072 Köster J. 2018. Reproducible data analysis with Snakemake. *F1000Res* **7**.
1073 La Cono V, La Spada G, Arcadi E, Placenti F, Smedile F, Ruggeri G, Michaud L, Raffa C, De
1074 Domenico E, Sprovieri M, Mazzola S, Genovese L, Giuliano L, Slepak VZ, Yakimov MM.
1075 2013. Partaking of Archaea to biogeochemical cycling in oxygen-deficient zones of
1076 meromictic saline Lake Faro (Messina, Italy). *Environ Microbiol* **15**:1717–1733.
1077 Langfelder P, Zhang B, Horvath S. 2008. Defining clusters from a hierarchical cluster tree: the
1078 Dynamic Tree Cut package for R. *Bioinformatics* **24**:719–720.
1079 Levins R. 1966. THE STRATEGY OF MODEL BUILDING IN POPULATION BIOLOGY. *Am Sci*
1080 **54**:421–431.

1081 Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform.
1082 *Bioinformatics* **26**:589–595.
1083 Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein
1084 or nucleotide sequences. *Bioinformatics* **22**:1658–1659.
1085 Liu XL. 2017. Deep Recurrent Neural Network for Protein Function Prediction from Sequence.
1086 *bioRxiv*. doi:10.1101/103994
1087 Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, Brady A, Creasy HH,
1088 McCracken C, Giglio MG, McDonald D, Franzosa EA, Knight R, White O, Huttenhower
1089 C. 2017. Strains, functions and dynamics in the expanded Human Microbiome Project.
1090 *Nature* **550**:61–66.
1091 Lobb B, Kurtz DA, Moreno-Hagelsieb G, Doxey AC. 2015. Remote homology and the functions
1092 of metagenomic dark matter. *Front Genet* **6**:1–12.
1093 Ma Y, Madupu R, Karaoz U, Nossa CW, Yang L, Yooseph S, Yachimski PS, Brodie EL, Nelson
1094 KE, Pei Z. 2014. Human papillomavirus community in healthy persons, defined by
1095 metagenomics analysis of human microbiome project shotgun sequencing data sets. *J*
1096 *Virol* **88**:4786–4797.
1097 Macarthur RH. 1957. On the relative abundance of bird species. *Proc Natl Acad Sci U S A*
1098 **43**:293–295.
1099 Martiny AC, Treseder K, Pusch G. 2013. Phylogenetic conservatism of functional traits in
1100 microorganisms. *ISME J* **7**:830–838.
1101 Matsen FA, Kodner RB, Armbrust EV. 2010. pplacer: linear time maximum-likelihood and
1102 Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC*
1103 *Bioinformatics* **11**:538.
1104 Méheust R, Burstein D, Castelle CJ, Banfield JF. 2019. The distinction of CPR bacteria from
1105 other bacteria based on protein family content. *Nat Commun* **10**:4173.
1106 Mendler K, Chen H, Parks DH, Lobb B, Hug LA, Doxey AC. 2019. AnnoTree: visualization and
1107 exploration of a functionally annotated microbial tree of life. *Nucleic Acids Res* **47**:4442–
1108 4448.
1109 Miklós I, Podani J. 2004. RANDOMIZATION OF PRESENCE–ABSENCE MATRICES:
1110 COMMENTS AND NEW ALGORITHMS. *Ecology*. doi:10.1890/03-0101
1111 Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, Crusoe MR, Kale V,
1112 Potter SC, Richardson LJ, Sakharova E, Scheremetjew M, Korobeynikov A, Shlemov A,
1113 Kunyavskaya O, Lapidus A, Finn RD. 2020. MGnify: the microbiome analysis resource in
1114 2020. *Nucleic Acids Res* **48**:D570–D578.
1115 Muller EEL. 2019. Determining Microbial Niche Breadth in the Environment for Better
1116 Ecosystem Fate Predictions. *mSystems* **4**. doi:10.1128/mSystems.00080-19
1117 NCBI Resource Coordinators. 2018. Database resources of the National Center for
1118 Biotechnology Information. *Nucleic Acids Res* **46**:D8–D13.
1119 Needham DM, Yoshizawa S, Hosaka T, Poirier C, Choi CJ, Hehenberger E, Irwin NAT, Wilken
1120 S, Yung C-M, Bachy C, Kurihara R, Nakajima Y, Kojima K, Kimura-Someya T, Leonard
1121 G, Malmstrom RR, Mende DR, Olson DK, Sudo Y, Sudek S, Richards TA, DeLong EF,
1122 Keeling PJ, Santoro AE, Shirouzu M, Iwasaki W, Worden AZ. 2019. A distinct lineage of
1123 giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proc Natl*
1124 *Acad Sci U S A* **116**:20574–20583.
1125 Olson DK, Yoshizawa S, Boeuf D, Iwasaki W, DeLong EF. 2018. Proteorhodopsin variability
1126 and distribution in the North Pacific Subtropical Gyre. *ISME J* **12**:1047–1060.
1127 Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, Biller SJ, Poulton NJ, Burkart MD,
1128 La Clair JJ, Chisholm SW, Stepanauskas R. 2019. Charting the Complexity of the
1129 Marine Microbiome through Single-Cell Genomics. *Cell* **179**:1623-1635.e11.
1130 Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P.
1131 2018. A standardized bacterial taxonomy based on genome phylogeny substantially

1132 revises the tree of life. *Nat Biotechnol* **36**:996–1004.
1133 Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, Beghini F, Manghi P, Tett A,
1134 Ghensi P, Collado MC, Rice BL, DuLong C, Morgan XC, Golden CD, Quince C,
1135 Huttenhower C, Segata N. 2019. Extensive Unexplored Human Microbiome Diversity
1136 Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography,
1137 and Lifestyle. *Cell* **176**:649-662.e20.
1138 Perdigão N, Rosa AC, O'Donoghue SI. 2017. The Dark Proteome Database. *BioData Min* **10**:1–
1139 11.
1140 Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, Kuehl JV, Melnyk RA, Lamson
1141 JS, Suh Y, Carlson HK, Esquivel Z, Sadeeshkumar H, Chakraborty R, Zane GM, Rubin
1142 BE, Wall JD, Visel A, Bristow J, Blow MJ, Arkin AP, Deutschbauer AM. 2018. Mutant
1143 phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**:503–509.
1144 Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F,
1145 Yamada T, Mende DR, Li J, Xu J, Li Shaochuan, Li D, Cao J, Wang B, Liang H, Zheng
1146 H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J-M, Hansen T, Le Paslier D, Linneberg A,
1147 Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li
1148 Shengting, Jian M, Zhou Y, Li Y, Zhang X, Li Songgang, Qin N, Yang H, Wang Jian,
1149 Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J,
1150 MetaHIT Consortium, Bork P, Ehrlich SD, Wang Jun. 2010. A human gut microbial gene
1151 catalogue established by metagenomic sequencing. *Nature* **464**:59–65.
1152 Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. 2017. Shotgun metagenomics, from
1153 sampling to analysis. *Nat Biotechnol* **35**:833–844.
1154 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
1155 features. *Bioinformatics* **26**:841–842.
1156 Remmert M, Biegert A, Hauser A, Söding J. 2012. HHblits: Lightning-fast iterative protein
1157 sequence searching by HMM-HMM alignment. *Nat Methods* **9**:173–175.
1158 Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S,
1159 Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu W-T, Eisen
1160 JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T.
1161 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*
1162 **499**:431–437.
1163 Ross A, Ward S, Hyman P. 2016. More Is Better: Selecting for Broad Host Range
1164 Bacteriophages. *Front Microbiol* **7**:1352.
1165 Rost B. 1999. Twilight zone of protein sequence alignments. *Protein Eng Des Sel* **12**:85–94.
1166 Roumpeka DD, Wallace RJ, Escalettes F, Fotheringham I, Watson M. 2017. A Review of
1167 Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. *Front*
1168 *Genet* **8**:23.
1169 Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA,
1170 Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C,
1171 Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF,
1172 Utterback T, Rogers Y-H, Falc□n LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM,
1173 Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR,
1174 Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC. 2007. The Sorcerer II
1175 Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific.
1176 *PLoS Biol* **5**:1–34.
1177 Salazar G, Cornejo-Castillo FM, Borrull E, Díez-Vives C, Lara E, Vaqué D, Arrieta JM, Duarte
1178 CM, Gasol JM, Acinas SG. 2015. Particle-association lifestyle is a phylogenetically
1179 conserved trait in bathypelagic prokaryotes. *Mol Ecol* **24**:5692–5706.
1180 Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh H-J, Cuenca M, Field CM, Coelho
1181 LP, Cruaud C, Engelen S, Gregory AC, Labadie K, Marec C, Pelletier E, Royo-Llonch M,
1182 Roux S, Sánchez P, Uehara H, Zayed AA, Zeller G, Carmichael M, Dimier C, Ferland J,

1183 Kandels S, Picheral M, Pisarev S, Poulain J, Tara Oceans Coordinators, Acinas SG,
1184 Babin M, Bork P, Bowler C, de Vargas C, Guidi L, Hingamp P, Iudicone D, Karp-Boss L,
1185 Karsenti E, Ogata H, Pesant S, Speich S, Sullivan MB, Wincker P, Sunagawa S. 2019.
1186 Gene Expression Changes and Community Turnover Differentially Shape the Global
1187 Ocean Metatranscriptome. *Cell* **179**:1068-1083.e21.
1188 Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc*
1189 *Natl Acad Sci U S A* **74**:5463–5467.
1190 Sberro H, Fremin BJ, Zlitni S, Edfors F, Greenfield N, Snyder MP, Pavlopoulos GA, Kyrpides
1191 NC, Bhatt AS. 2019. Large-Scale Analyses of Human Microbiomes Reveal Thousands
1192 of Small, Novel Genes. *Cell* **178**:1245-1259.e14.
1193 Scales BS, Dickson RP, LiPuma JJ, Huffnagle GB. 2014. Microbiology, genomics, and clinical
1194 significance of the Pseudomonas fluorescens species complex, an unappreciated
1195 colonizer of humans. *Clin Microbiol Rev* **27**:927–948.
1196 Schloss PD, Handelsman J. 2008. A statistical toolbox for metagenomics: assessing functional
1197 diversity in microbial communities. *BMC Bioinformatics* **9**:34.
1198 Sirén K, Millard A, Petersen B, Gilbert MTP, Clokie MRJ, Sicheritz-Pontén T. 2021. Rapid
1199 discovery of novel prophages using biological feature engineering and machine learning.
1200 *NAR Genom Bioinform* **3**. doi:10.1093/nargab/lqaa109
1201 Skewes-Cox P, Sharpton TJ, Pollard KS, DeRisi JL. 2014. Profile hidden Markov models for the
1202 detection of viruses within metagenomic sequence data. *PLoS One* **9**:e105067.
1203 Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R,
1204 Schleper C, Guy L, Ettema TJG. 2015. Complex archaea that bridge the gap between
1205 prokaryotes and eukaryotes. *Nature* **521**:173–179.
1206 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
1207 large phylogenies. *Bioinformatics* **30**:1312–1313.
1208 Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. 2019a. HH-suite3
1209 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*
1210 **20**:473.
1211 Steinegger M, Mirdita M, Söding J. 2019b. Protein-level assembly increases protein sequence
1212 recovery from metagenomic samples manyfold. *Nat Methods* **16**:603–606.
1213 Steinegger M, Salzberg SL. 2020. Terminating contamination: large-scale search identifies
1214 more than 2,000,000 contaminated entries in GenBank. *Genome Biol* **21**:115.
1215 Steinegger M, Söding J. 2018. Clustering huge protein sequence sets in linear time. *Nat*
1216 *Commun* **9**:2542.
1217 Steinegger M, Soding J. 2017. MMseqs2 enables sensitive protein sequence searching for the
1218 analysis of massive data sets. *Nat Biotechnol* **advance on**. doi:10.1038/nbt.3988
1219 Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller
1220 G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen
1221 S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G,
1222 Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M,
1223 Searson S, Kandels-Lewis S, Tara Oceans coordinators, Bowler C, de Vargas C, Gorsky
1224 G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S,
1225 Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG,
1226 Bork P. 2015. Ocean plankton. Structure and function of the global ocean microbiome.
1227 *Science* **348**:1261359.
1228 The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids*
1229 *Res* **45**:D158–D169.
1230 Thomas AM, Segata N. 2019. Multiple levels of the unknown in microbiome research. *BMC Biol*
1231 **17**:48.
1232 UniProt Consortium T. 2018. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*
1233 **46**:2699.

van Dongen S, Abreu-Goodger C. 2012. Using MCL to Extract Clusters from Networks In: van Helden J, Toussaint A, Thieffry D, editors. Bacterial Molecular Networks: Methods and Protocols. New York, NY: Springer New York. pp. 281–295.

Vanhoutreve R, Kress A, Legrand B, Gass H, Poch O, Thompson JD. 2016. LEON-BIS: multiple alignment evaluation of sequence neighbours using a Bayesian inference system. *BMC Bioinformatics* **17**:271–271.

Vorobev A, Dupouy M, Carradec Q, Delmont T, Annamale A, Wincker P, Pelletier E. 2020. Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. *Genome Res*. doi:10.1101/gr.253070.119

Wyman SK, Avila-Herrera A, Nayfach S, Pollard KS. 2018. A most wanted list of conserved microbial protein families with no known domains. *PLoS One* **13**:e0205749.

Yooseph S, Li W, Sutton G. 2008. Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering. *BMC Bioinformatics* **9**:1–13.

Yooseph S, Sutton G, Rusch BD, Halpern LA, Williamson JS, Remington K, Eisen AJ, Heidelberg BK, Manning G, Li W, Jaroszewski L, Cieplak P, Miller SC, Li H, Mashiyama TS, Joachimiak PM, Van Belle C, Chandonia MJ, Soergel AD, Zhai Y, Natarajan K, Lee S, Raphael JB, Bafna V, Friedman R, Brenner ES, Godzik A, Eisenberg D, Dixon EJ, Taylor SS, Strausberg LR, Frazier M, Venter CJ. 2007. The Sorcerer II global ocean sampling expedition: Expanding the universe of protein families. *PLoS Biol* **5**:0432–0466.

Žure M, Fernandez-Guerra A, Munn CB, Harder J. 2017. Geographic distribution at subspecies resolution level: closely related Rhodopirellula species in European coastal sediments. *ISME J* **11**:478–489.