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

Chiara Vanni[1,2], Matthew Schechter[1,3], Silvia Acinas[4], Albert Barberán[5], Pier Luigi Buttigieg[6], Emilio O. Casamayor[7], Tom O. Delmont[8], Carlos M. Duarte[9], A. Murat Eren[3,10], Rob Finn[11], Alex Mitchell[11], Pablo Sanchez[4], Kimmo Siren[12], Martin Steinegger[13,14], Frank Oliver Glöckner[15,16,17], Antonio Fernandez-Guerra[1,18]

[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]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]University of Bremen, MARUM, Bibliothekstraße 1, 28359 Bremen
[18]Lundbeck GeoGenetics Centre, The Globe Institute, University of Copenhagen, 1350 Copenhagen, Denmark

Abstract

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

Introduction

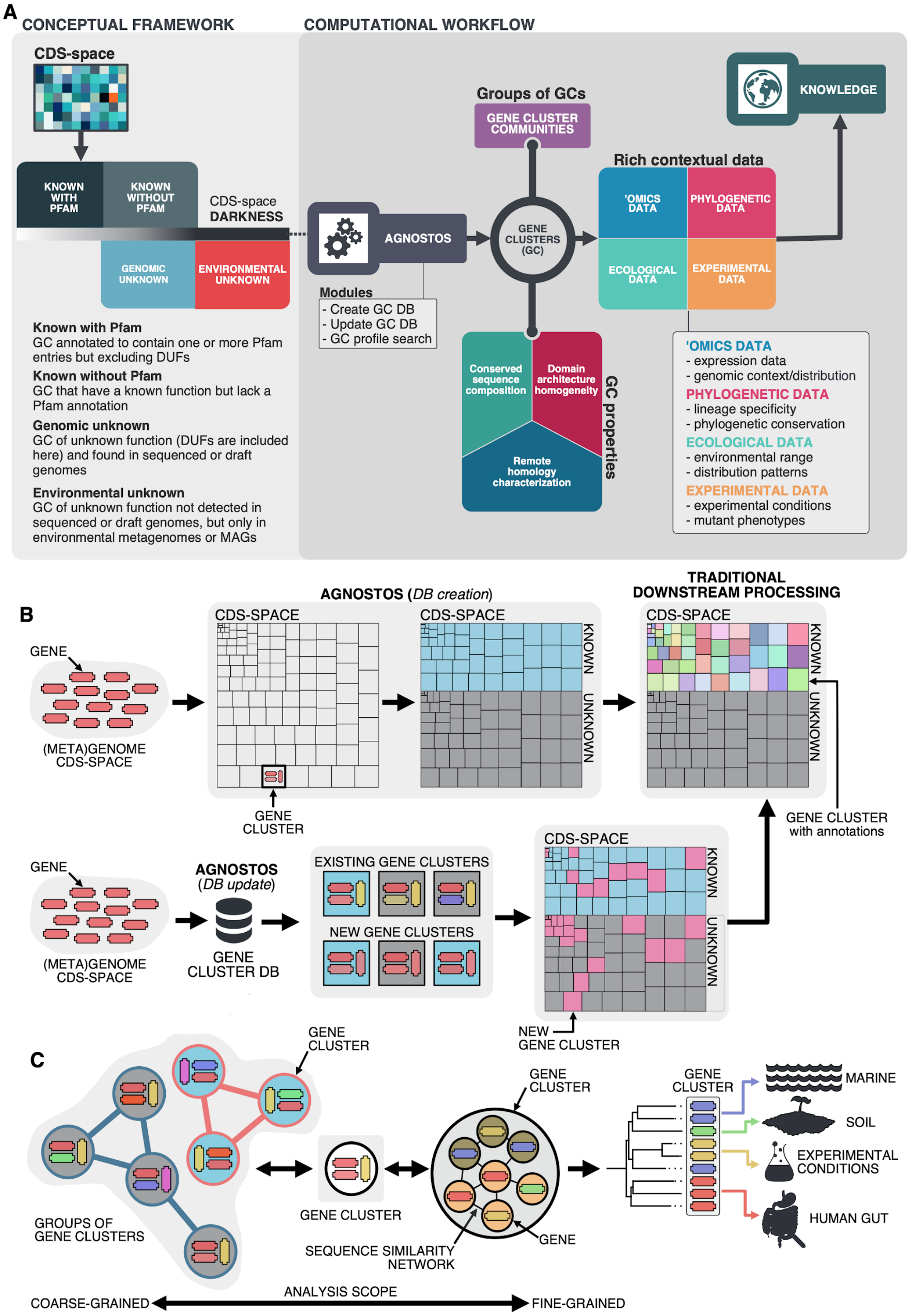
Thousands of isolate, single-cell, and metagenome-assembled genomes are guiding us towards a better understanding of how microbes shape life on Earth [1–7], thus bringing about a golden age of microbial genomics. An ever increasing number of genomes and metagenomes are unlocking uncharted regions of microbial diversity[1,8,9], providing new perspectives on the evolution of life[10,11]. However, our rapidly growing inventories of new genes have a glaring issue: between 40% to 60% cannot be assigned to a known function[12–14]. Current analytical approaches for genomic and metagenomic data[15–19] generally do not include this uncharacterized fraction in downstream analyses, constraining their results to conserved pathways and housekeeping functions[16]. This inability to handle shades of the unknown is an immense impediment to realizing the potential for discovery of microbial genomics and microbiology at large[12,20]. Predicting function from traditional sequence similarity appears to have yielded all it can[21–23], thus several groups have attempted to resolve gene function by other means. Such efforts include combining biochemistry and crystallography[24]; using environmental co-occurrence[25]; by grouping those genes into evolutionarily related families[26–29]; and using remote homologies[30,31]. In 2018, Price et al.[13] developed a high-throughput experimental pipeline that provides mutant phenotypes for thousands of bacterial genes of unknown function being one of the most promising methods to tackle the unknown. Despite their promise, experimental methods are labor-intensive and require novel computational methods that could bridge the existing gap between the known and unknown coding sequence space (CDS-space).

Here we present a conceptual framework and a computational workflow that closes the gap between the known and unknown CDS-space by connecting genomic and metagenomic gene clusters. Our approach adds context to vast amounts of unknown biology, providing an invaluable resource to get a better understanding of the unknown functional fraction and boost the current methods for its experimental characterization. The application of our approach to 415,971,742 genes predicted from 1,749 metagenomes and 28,941 bacterial and archaeal genomes shows that (1) the extent of the unknown fraction is smaller than expected, (2) that the diversity of gene clusters in the unknown fraction is higher than in the known fraction, and that (3) the unknown fraction is phylogenetically more conserved and is predominantly lineage-specific at the species level. Finally, we show how we can connect all the outputs produced by our approach to augment the results from experimental data and add context to genes of unknown function through hypothesis-driven molecular investigations.

Results

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

We created the conceptual and technical foundations to unify the known and unknown CDS-space and provide a practical solution to one of the most significant ongoing challenges in microbiome analyses. First, we developed a conceptual framework to partition the genomic and metagenomic CDS-space based on its level of characterization, and that simultaneously combines genomic and metagenomic data (Fig. 1A). We conceptually partitioned the known and unknown fractions into (1) Knowns with Pfam annotations (K), (2) Knowns without Pfam annotations (KWP), (3) Genomic unknowns (GU), and (4) Environmental unknowns (EU) (Fig. 1A). The framework introduces a subtle change of paradigm compared to traditional approaches, our objective is to provide the best representation of the unknown space and 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[32]. With this objective in mind, we use gene clusters (GCs) instead of genes as the fundamental unit to compartmentalize the CDS-space owing to their unique characteristics (Fig. 1B). GCs produce a structured CDS-space reducing its complexity (Fig. 1B), are independent of the known and unknown fraction, are conserved across environments and organisms, and can be used to aggregate information from different sources (Fig. 1A). Moreover, the GCs provide a good compromise in terms of resolution for analytical purposes and owing to their special 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).



Known with Pfam
GC annotated to contain one or more Pfam entries but excluding DUFs

Known without Pfam
GC that have a known function but lack a Pfam annotation

Genomic unknown
GC of unknown function (DUFs are included here) and found in sequenced or draft genomes

Environmental unknown
GC of unknown function not detected in sequenced or draft genomes, but only in environmental metagenomes or MAGs

'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

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

Driven by the concepts defined in the conceptual framework, we developed AGNOSTOS, a computational workflow that infers, validates, refines, and classifies GCs in the four proposed categories (Fig. 1A; Fig. 1B; Supp. Fig 1). AGNOSTOS provides two operational modules (*DB creation* and *DB update*) to produce GCs with a highly conserved intra-homogeneous structure (Fig. 1B), both in terms of sequence similarity and domain architecture homogeneity; it exhausts any existing homology to known genes and provides a proper delimitation of the unknown CDS-space before classifying each GC in one of the four categories. In the last step, we decorate each GC with a rich collection of contextual data that we compile from different sources, or that we generate by analyzing the GC contents in different contexts (Fig. 1A). For each GC, we also offer several products that can be used for analytical purposes like improved representative sequences, consensus sequences, sequence profiles for MMseqs2[33] and HHblits[34], or the GC members as a sequence similarity network (see Online Methods). To complement the collection, we also provide a subset of what we define as *high-quality* GCs. In those GCs, the representative is a complete gene and complete genes make more than one-third of genes within a GC.

Partitioning and contextualizing the coding sequence space of genomes and metagenomes

We used our approach to explore the unknown CDS-space of 1,749 microbial metagenomes derived from human and marine environments, and 28,941 genomes from GTDB_r86 (Supp Fig 2A).
The initial gene prediction of AGNOSTOS (Supp Fig 1) produced 322,248,552 genes from the environmental dataset and assigned a Pfam annotation to 44% of them. Next, it clustered the predicted genes in 32,465,074 GCs. For the downstream processing, we kept 3,003,897 GCs (83% of the original genes) after filtering out any GC that contained less than 10 genes[35] removing 9,549,853 clusters and 19,911,324 singletons (Supp Fig 2A; Supp. Note 1). The validation process selected 2,940,257 *good-*

quality clusters (Fig. 1B; Supp. Table 1; Supp. Note 2) which resulted in 43% of them being members of the unknown CDS-space after the classification and remote homology refinement steps (Supp Fig 2A, Supp. Note 3).

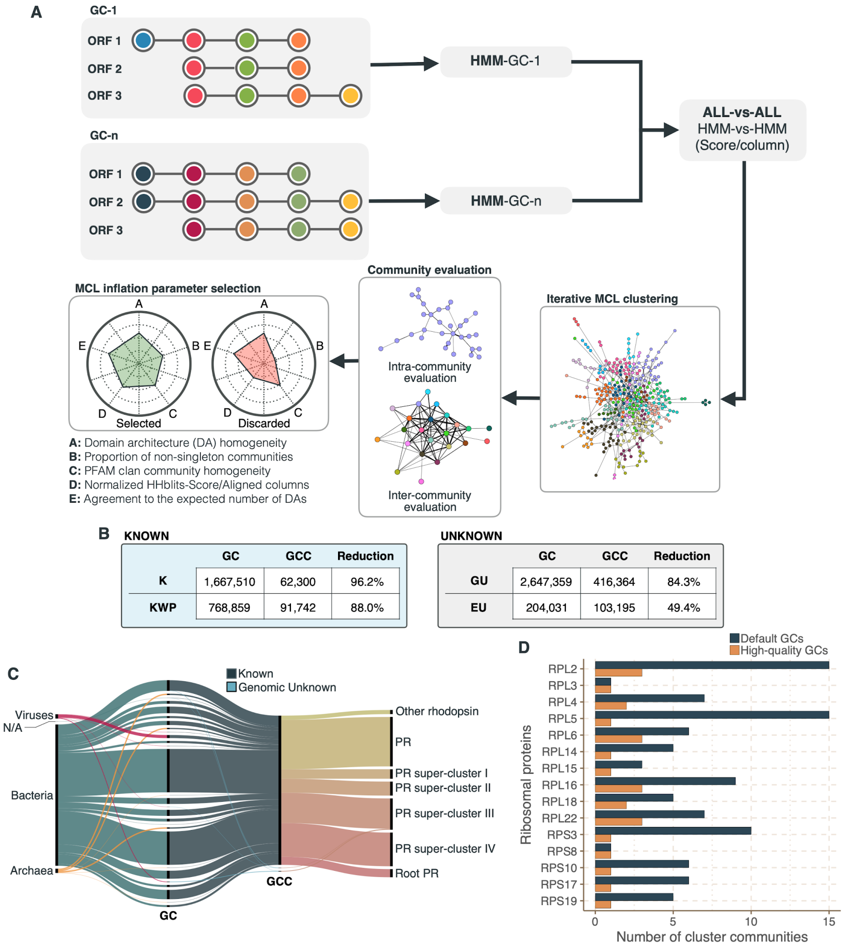
We build the link between the environmental and genomic CDS-space by expanding the final collection of GCs with the genes predicted from GTDB_r86 (Supp Fig 2A). Our environmental GCs already included 72% of the genes from GTDB_r86, 22% of them created 2,400,037 new GCs and the rest 6% resulted in singleton GCs (Supp Fig 2A; Supp. Note 4; Supp. Note 5). The final dataset includes 5,287,759 GCs (Supp Fig 2A), with both datasets sharing only 922,599 GCs (Supp Fig 2B). The addition of the GTDB_r86 genes increased the proportion of GCs in the unknown CDS-space to 54%. As the final step, the workflow generated a subset of 203,217 *high-quality* GCs (Supp Table 2; Supp Fig 3). In these *high-quality* clusters, we identified 12,313 clusters potentially encoding for small proteins (<= 50 amino acids). Most of these GCs are unknown (66% of them), which agrees with recent findings on novel small proteins from metagenomes[36].

The KWP category contains the largest proportion of incomplete ORFs (Supp. Table 3), impeding the detection and assignment of Pfam domains. But it also incorporates sequences with an unusual amino acid composition that have homologs to proteins with high levels of disorder in the DPD database[37] and that have characteristic functions of the intrinsically disordered proteins[38] (IDP) like cellular processes and signaling as predicted by eggNOG annotations (Supp. Table 4).

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

Beyond the twilight zone, communities of gene clusters

The method we developed to group GCs in gene cluster communities (GCCs) (Fig. 2A) reduced the final collection of GCs by 87%, producing 673,601 GCCs (Fig. 2B; Supp. Note 7). We validated the ability of our approach to capture remote homologies between related GCs using two well-known gene families present in our environmental datasets, proteorhodopsins[39] and bacterial ribosomal proteins[40]. In our dataset, 64 GCs (12,184 genes) and 3 GCCs (Supp Note 8) contained sequences classified as proteorhodopsin (PR). One *Known* GCC contained 99% of the PR annotated genes (Fig. 2C), with the only exception of twenty genes taxonomically annotated as viral and assigned to the *PR Supercluster I*[41] enclosed in two GU communities (five GU gene clusters). For the ribosomal proteins, the results were not so satisfactory. We identified 1,843 GCs (781,579 genes) and 98 GCCs. The number of GCCs compared to the expected number of ribosomal proteins families (16) used for the validation. When we use *high-quality* GCs (Supp. Note 8), we get closer to the expected number of GCCs (Fig. 2D). With this subset, we identified 26 GCCs and 145 GCs (1,687 genes). The cross-validation of our method against the approach used in Méheust et al.[40] (Supp. Note 9) confirmed the intrinsic complexity of analyzing metagenomic data. Both approaches showed a high agreement in the GCCs identified (Supplementary Table 9-1) but our method inferred less GCCs for each of the ribosomal protein families (Supplementary Figure 9-3), coping better with the nuisances of a metagenomic setup, like 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%

Figure 2: Overview and validation of the workflow to aggregate GCs in communities. (A) We inferred a gene cluster homology network using the results of an all-vs-all HMM gene cluster comparison with HHBLITS. The edges of the network are based on the HHblits-score/Aligned-columns. Communities are identified by an iterative screening of different MCL inflation parameters and evaluated using five different metrics that take into account the inter- and intra-community properties. (B) Comparison of the number of GCs and GCCs for each of the functional categories. (C) Validation of the GCCs inference based on the environmental genes annotated as proteorhodopsins. Ribbons in the alluvial plot are genes, and each stacked bar corresponds (from left to right) to the (1) gene taxonomic classification at domain level, (2) GC membership, (3) GCC membership and (4) MicRhoDE operational

classification. (D) Validation of the GCCs inference based on ribosomal proteins based on standard and high-quality GCs.

A smaller but highly diverse unknown coding sequence space

Combining clustering and remote homology searches reduces the extent of the unknown CDS-space compared to the traditional genomic and metagenomic analysis approaches (Fig. 3A). Our workflow recruited as much as 72% of genes in human-related metagenomic samples and 66% of the genes in marine metagenomes into the known CDS-space. In both human and marine microbiomes, the genomic unknown fraction shows a similar proportion of genes (21%, Fig. 3A). The number of genes corresponding to EU gene clusters is higher in marine metagenomes; in total, 12% of the genes are part of this GC category. We observed a similar outcome evaluating genes from the GTDB_r86, where 75% of bacterial and 64% of archaeal genes were in the Known group. Archaeal genomes contain more unknowns than those from the Bacteria, where 30% of the genes are classified as genomic unknowns in Archaea, and only 20% in Bacteria (Fig. 3A; Supp. Table 7). To evaluate the coverage of our dataset, we calculated the accumulation rates of GCs and GCCs. For the metagenomic dataset we used 1,264 metagenomes (18,566,675 GCs and 282,580 GCCs) and for the genomic dataset 28,941 genomes (9,586,109 GCs and 496,930 GCCs). The rate of accumulation of unknown GCs was three times higher than the known (2 times for the genomic), and both cases were far from reaching a plateau (Fig. 3B). This is not the case for the GCC accumulation curves (Supp Fig 4B), where they reached a plateau. The rate of accumulation is largely determined by the large number of singletons, and especially singletons from EUs (Supp note 11 and Supp Fig 5). While the accumulation rate of known GCs between marine and human metagenomes is almost identical, there are striking differences for the unknown GCs (Fig. 3C). These differences are maintained even when we remove the virus-enriched samples from the marine metagenomes (Supp Fig 4A). Although the marine metagenomes include a large variety of environments, from coastal to the deep sea, the known space remains quite constrained.

Despite only including marine and human metagenomes in our database, our coverage to other databases and environments is quite comprehensive, with an overall coverage of the 76% (Supp. Note 12). The lowest covered biomes are freshwater, soil and human non-digestive as revealed by the screening of MGnify[15] (release 2018_09; 11 biomes; 843,535,6116 proteins) where we assigned 74% of the MGnify proteins into one of our categories (Supplementary Fig. 6).

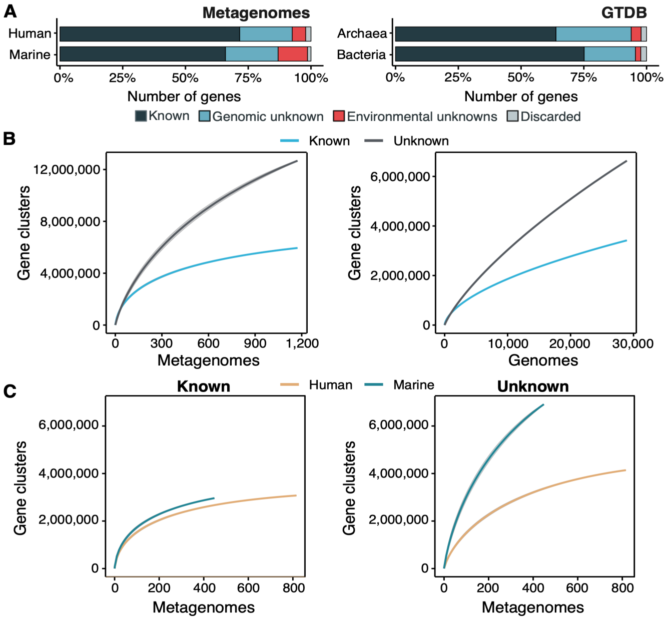


Figure 3: The extent of the known and unknown coding sequence space (A) Proportion of genes in each cluster category. (B) Accumulation curves for the known and unknown CDS-space at the GC- level for the metagenomic and genomic data. from TARA, MALASPINA, OSD2014 and HMP-I/II projects. (C) Collector curves comparing the human and marine biomes. Non-abundant singleton clusters were excluded from the calculations.

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

Although the role of the unknown fraction in the environment is still a mystery, the large number of gene counts and abundance observed underlines its inherent ecological relevance (Fig. 4A). In some samples the genomic unknown fraction can account for more than 40% of the total gene abundance observed (Fig. 4A). The environmental unknown fraction is also relevant in several samples, where singleton GCs are the majority (Fig. 4A). We identified two metagenomes with an unusual composition in terms of environmental unknown singletons. The marine metagenome corresponds to a sample from Lake Faro (OSD42), a meromictic saline with a unique extreme environment where Archaea plays an important role[42]. The HMP metagenome (SRS143565) corresponds to a human sample from the right cubital fossa from a healthy female subject. To understand the unusual composition of this metagenome, we should perform further analyses to discard potential technical artifacts like sample contamination.

The ratio between the unknown and known GCs revealed that the metagenomes located at the upper left quadrant in Fig. 4B-C are enriched in GCs of unknown function. In human metagenomes, we can distinguish between body sites, with the gastrointestinal tract, where microbial communities are expected to be more diverse and complex, especially enriched with genomic unknowns. The HMP metagenomes with the largest ratio of unknowns are those samples identified to contain crAssphages[43,44] and HPV viruses[45] (Supp. Table 8; Supp. Fig. 7). Consistently, in marine metagenomes (Fig. 4D) we can separate between size fractions, where the highest ratio in genomic and environmental unknowns correspond to the ones enriched with viruses and giant viruses.

To complement the previous findings, we performed a large-scale analysis to investigate the GC occurrence patterns in the environment. The narrow distribution of the unknown fraction (Fig. 4D) suggests that these GCs might provide a selective advantage and be important for the adaptation to specific environmental conditions. But the pool of broadly distributed environmental unknowns is the most interesting result. We identified traces of potential ubiquitous organisms left uncharacterized by traditional approaches, as more than 80% of these GCs cannot be associated with a MAG (Supp Table 9, Supp. Note 10).

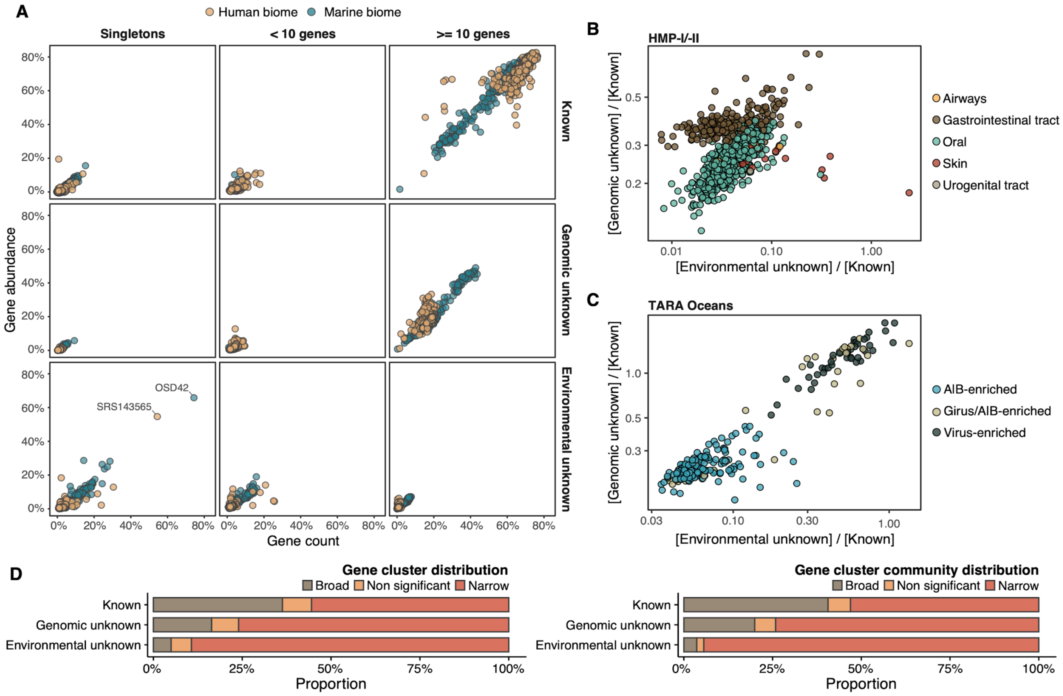


Figure 4: Distribution of the unknown coding sequence space in the human and marine metagenomes (A) Ratio between the proportion of the number of genes and their estimated abundances per cluster category and biome. Columns represented in the facet depicts three cluster categories based on the size of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

The genomic unknown coding sequence space is lineage-specific

We already showed that the unknown CDS-space is habitat-specific and might be relevant for organism adaptation. With the inclusion of the genomes from GTDB_r86, we have accessed a phylogenomic framework to assess the phylogenetic conservation level and lineage-specificity of the GCs[46,47]. We identified 782,142 lineage-specific GCs and 465,148 phylogenetically conserved GCs in Bacteria (Supplementary Table 10; Supp. Note 13 for Archaea). The number of lineage-specific GCs increases with the Relative Evolutionary Distance[11] (Fig. 5A) and differences between the known and the unknown fraction start to be evident at the Family level. The unknown GCs are more phylogenetically conserved than the known (Fig. 5B, $p < 0.0001$), revealing the importance of the genome's uncharacterized fraction. This is not the case for the lineage-specific and phylogenetically conserved GCs, where the unknown GCs are less phylogenetically conserved (Fig. 5B), agreeing with the large number of lineage-specific GCs at Genus and Species level. To discard the possibility that the lineage-specific GCs of unknown function have a viral origin, we screened all GTDB_r86 genomes for prophages. We only found 37,163 lineage-specific GCs in prophage genomic regions, being 86% of them GCs of unknown function. After unveiling the potential relevance of the GCs of unknown function in bacterial genomes, we identified phyla in GTDB_r86 enriched with these types of clusters. A clear pattern emerged when we partitioned the phyla based on the ratio of known to unknown GCs and vice versa (Fig. 5D), the phyla with a larger number of MAGs are enriched in GCs of unknown function Fig. 5D. Phyla with a high proportion of non-classified GCs (those discarded during the validation steps) contain a small number of genomes and are primarily composed by MAGs. These groups of phyla highly enriched in unknowns and represented mainly by MAGs include newly described phyla such as *Cand.* Riflebacteria and *Cand.* Patescibacteria[9,48,49], both with the largest unknown to known ratio.

We demonstrate the possibility to bridge genomic and metagenomic data and simultaneously unify the known and unknown CDS-space by integrating the new Ocean Microbial Reference Gene Catalog[50] (OM-RGC v2) in our database. We assigned 26,170,875 genes to known GCs, 11,422,975 to genomic unknowns, 8,661,221 to environmental unknown and 520,083 were discarded. From the 11,422,975 genes classified as genomic unknowns, we could associate 3,261,741 to a GTDB_r86 genome and we identified 56,402 as lineage-specific. The alluvial plot in Fig. 5E depicts the new organization of the OM-RGC v2 after being integrated into our and how we can provide context to the two original types of unknowns in the OM-RGC (those annotated as category S in eggNOG[51] and those without known homologs in the eggNOG database[50]) that can lead to potential experimental targets at the organism level to complement the metatranscriptomic approach proposed by Salazar et al[50].

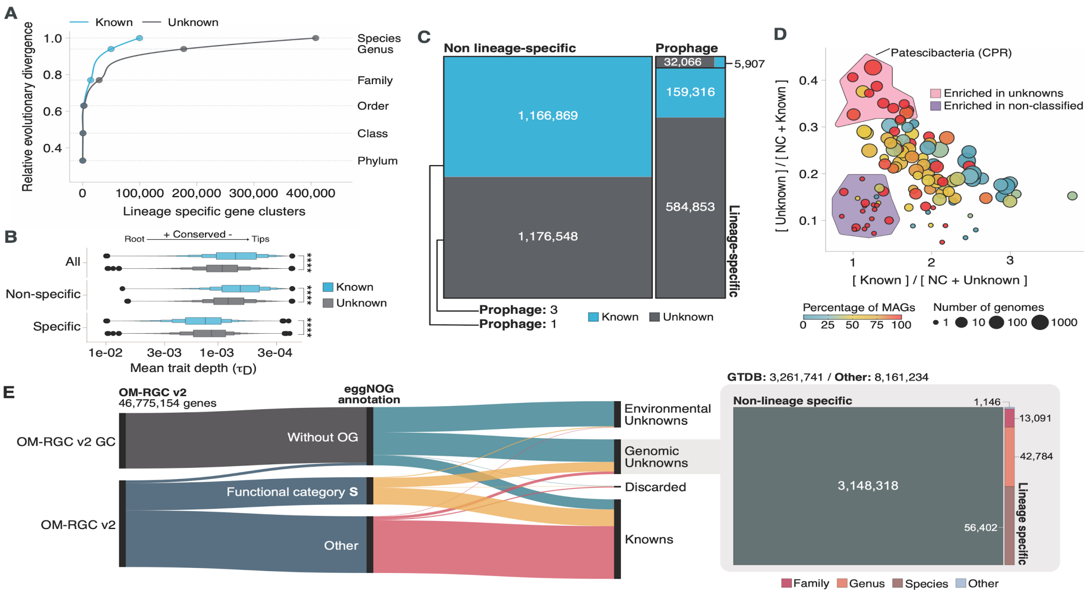


Figure 5: Phylogenomic exploration of the unknown coding sequence space. (A) Distribution of the lineage-specific GCs by taxonomic level. Lineage-specific unknown GCs are more abundant in the lower taxonomic levels (genus, species). (B) Phylogenetic conservation of the known and unknown coding sequence space in 27,372 bacterial genomes from GTDB_r86. There are differences in the conservation between the known and the unknown coding sequence space for lineage- and non-lineage specific GCs (paired Wilcoxon rank-sum test; all p-values < 0.0001). (C) The majority of the lineage-specific clusters are part of the unknown coding sequence space, being a small proportion found in prophages present in the GTDB_r86 genomes. (D) Known and unknown coding sequence space of the 27,732 GTDB_r86 bacterial genomes grouped by bacterial phyla. Phyla are partitioned based on the ratio of known to unknown GCs and vice versa. Phyla enriched in MAGs have higher proportions in GCs of unknown function. Phyla with a high proportion of non-classified clusters (NC; discarded during validation) tend to contain a small number of genomes. (E) The left side of the alluvial plot shows the uncharacterized (OM-RGC v2 GC) and characterized (OM-RGC v2) fraction of the gene catalog. The functional annotation is based on eggNOG annotations[50]. The right side of the alluvial plot shows the new organization of the OM-RGC v2 coding sequence space based on the approach described in this study. The treemap in the right links the metagenomic and genomic space adding context to the unknown fraction of the OM-RGC v2.

Augmenting experimental data through a structured coding sequence space

We selected one of the experimental conditions tested in Price et al.[13] to demonstrate the potential of our approach to augment experimental data. We compared the fitness values in plain rich medium with added Spectinomycin dihydrochloride pentahydrate to the fitness in plain rich medium (LB) in *Pseudomonas fluorescens FW300-N2C3* (Fig. 6A). This antibiotic inhibits protein synthesis and elongation by binding to the bacterial 30S ribosomal subunit and interferes with the peptidyl tRNA translocation. We identified the gene with locus id AO356_08590 that presents a strong phenotype (fitness = -3.1; t = -9.1) and has no known function. This gene belongs to the genomic unknown GC GU_19737823. We can track this GC into the environment and explore the occurrence in the different samples we have in our database. As expected, the GC is mostly found in non-human metagenomes (Fig. 6B) as *Pseudomonas* are common inhabitants of soil and water environments[52]. However, finding this GC also in human related samples is very interesting, due to the potential association of *P. fluorescens* and human disease where Crohn's disease patients develop serum antibodies to this microbe[53].

We can add more information to the selected GC by exploiting the remote homologies found in the GCC GU_c_21103 (Fig. 6C). We identified all the genes in the GTDB_r86 genomes that belong to the GCC GU_c_21103 (Supplementary table 11) and explored their genomic neighborhoods. All members from GU_c_21103 are constrained to the class *Gammaproteobacteria*, and interestingly 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 30S ribosomal proteins, the prime target of spectinomycin. The combination of the experimental evidence and the associated data inferred by our approach provides strong support to generate the hypothesis that the gene AO356_08590 might be involved in the resistance to spectinomycin.

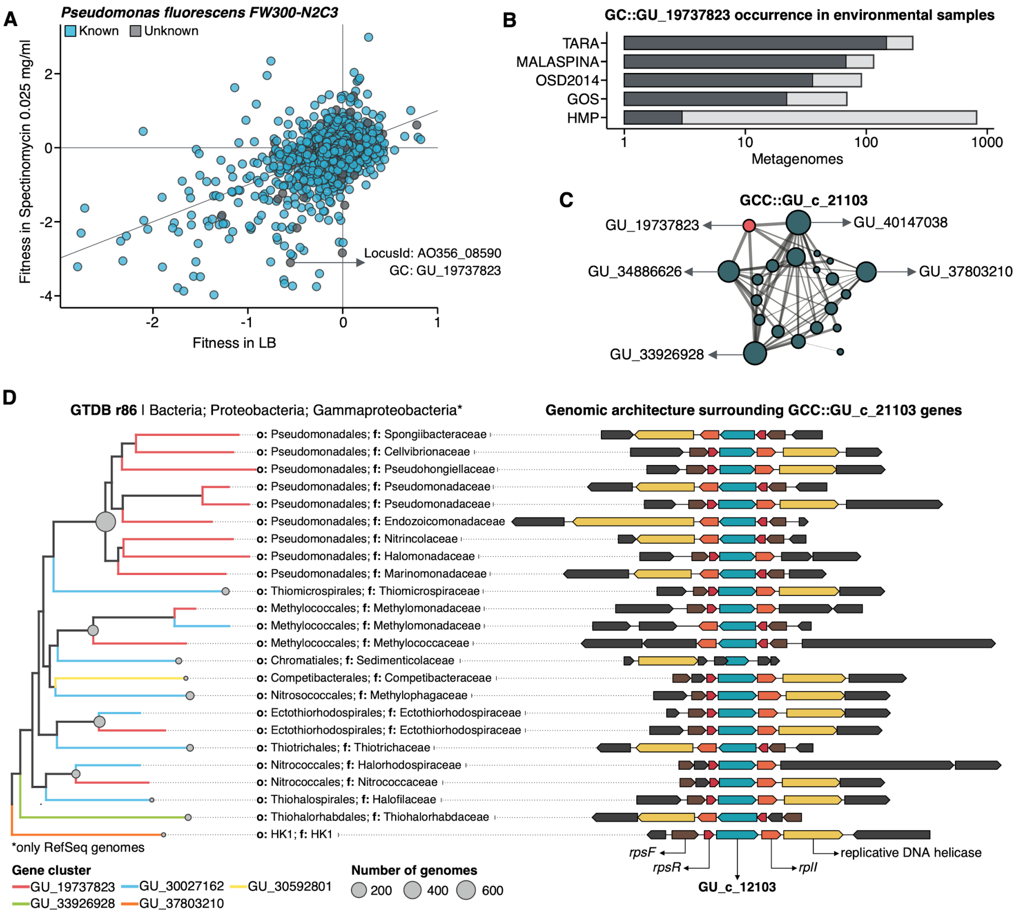


Figure 6: Augmenting experimental data with GCs of unknown function. (A) We used the fitness values from the experiments from Price et al.[13] to identify genes of unknown function that are important for fitness under certain experimental conditions. The selected gene belongs to the genomic unknown GC GU_19737823 and presents a strong phenotype (fitness = -3.1; t = -9.1) (B) Occurrence of GU_19737823 in the metagenomes used in this study. Darker bars depict the number of metagenomes where the GC is found. (C) GU_19737823 is a member of the GCC GU_c_21103. The network shows the relationships between the different GCs members of the gene cluster community GU_c_21103. The size of the node corresponds to the node degree of each GC. Edge thickness corresponds to the bitscore/column metric. Highlighted in red is GU_19737823. (D) We identified all the genes in the GTDB_r86 genomes that belong to the GCC GU_c_21103 and explored their genomic neighborhoods. GU_c_21103 members were constrained to the class Gammaproteobacteria, and GU_19737823 is mostly exclusive to the order Pseudomonadales. The gene order in the different genomes analyzed is highly conserved, finding GU_19737823 after the *rpsF::rpsR* operon and before *rplI. rpsF* and *rpsR* encode for the *30S ribosomal protein S6* and *30S ribosomal protein S18* respectively. The GTDB_r86 subtree only shows RefSeq genomes. Branch colors correspond to the different GCs found in GU_c_21103. Bubble plot depicts the number of genomes with a gene that belongs to GU_c_21103.

Discussion

We present a new conceptual framework and computational workflow to unify the known and unknown CDS-space in microbial analyses. Using this framework, we performed an in-depth exploration of the microbial unknown CDS-space and demonstrated that we can link the unknown fraction of metagenomic studies to specific genomes and provide a powerful tool for hypothesis generation. During the last years the microbiome community has established a standard operating procedure[16] for analyzing metagenomes that can briefly be summarized into (1) assembly, (2) gene prediction, (3) gene catalog inference, (4) binning, and (5) characterization. Thanks to recent computational developments[54] we envisioned an alternative to this workflow where we can maximize the information used when analyzing genomic and metagenomic data. With a well-structured CDS-space as the one proposed by our framework, we can also provide a mechanism to reconcile top-down and bottom-up approaches. With the large amount of data available, AGNOSTOS can create environmental- and organism- specific variations of a seed GC database. Then integrate the predicted genes of new genomes and metagenomes while dynamically creating and classifying new GCs with the genes that couldn't be integrated in the initial step (Fig. 1B). Afterwards, the potential functions of the known GCs can be carefully characterized by integrating them into the traditional workflows.

One of the most appealing characteristics of our approach, is that the GCs provide unified groups of homologous genes across environments and organisms indifferently if they belong to the known or unknown CDS-space, and the contextualization of the unknown fraction allows its integration in our analyses. Our combination of partitioning and contextualization features a smaller unknown CDS-space than we expected. On average, for our genomic and metagenomic data, only 30% of the genes fall in the unknown fraction. One hypothesis to reconcile this surprising finding is that until recently, the methodologies to identify remotely homologous sequences in large datasets were computationally prohibitive. New methods[55] like the ones integrated in AGNOSTOS, are enabling large scale distant homology searches but one has to apply conservative measures to control the trade-off between specificity and sensitivity to avoid overclassification.

However, despite the fact that we found a reduced unknown CDS-space, it presents a high diversity as shown in the GC accumulation curves, highlighting the vast remaining untapped microbial fraction and its potential importance for niche adaptation owing to its narrow distribution. In a genomic context, the unknown fraction is predominantly species' lineage-specific and phylogenetically more conserved than the known fraction, supporting the signal observed in the environmental data and emphasizing that the unknown fraction should not be ignored. We also ruled out the effect of prophages, strengthening the hypothesis that the lineage-specific GCs of unknown function might be associated with the mechanisms of microbial diversification and niche adaptation as a result of the constant diversification of gene families and the survival of new gene lineages[55,56]. Metagenome-assembled genomes are not only unveiling new regions of the microbial universe (42% of the genomes in GTDB_r86), but they are also enriching genes of unknown function in the tree of life. We investigated the unknown CDS-space of *Cand*. Patescibacteria, more commonly known as Candidate Phyla Radiation (CPR), a phylum that has raised considerable interest due to their unusual biology[9]. We provide a collection of 54,350 lineage-specific GCs of unknown function at different taxonomic level resolutions (Supp. Table 12; Supplementary Note 14) which will be a valuable resource for the advancement of knowledge in the CPR research efforts.

Our effort to tackle the unknown provides a pathway to unlock a large pool of likely relevant data that remains untapped to analysis and discovery. With the identification of a potential target gene of unknown function for antibiotic resistance we demonstrate the value of our approach and how it can boost insights from model organism experiments. But severe challenges remain, such as the dependence on the quality of the assemblies and their gene predictions as shown by the analysis of the ribosomal protein GCCs where many of the recovered genes are incomplete. While sequence assembly has been an active area of research[57], this has not been the case for gene prediction methods[57], which are becoming outdated[58] and cannot cope with the current amount of data. Alternatives like protein-level assembly[59] combined with the exploration of the assembly graphs' neighborhoods[60] become very attractive for our purposes. In any case, we still face the challenge of discriminating between real and artifactual singletons[61]. At the moment, there are no methods available to provide a plausible solution and, at the same time, being scalable. We devise a potential solution in the recent developments in unsupervised deep learning methods where they use large corpora of proteins to define a language model *embedding* for protein sequences[62]. These models could be applied to predict *embeddings* in singletons, which could be clustered or used to determine their coding potential. Furthermore, the fragmented nature of the short-read based metagenomic assemblies can inflate the number of GCs recovered, and especially after our conservative approach to avoid the inclusion of fragmented genes that could be unrelated to the GCs in multidomain proteins. Not only splitting can be a problem, but also lumping unrelated genes or GCs owing to the use of remote homologies. Although the inference of GCCs is using very sensitive methods to compare HMM profiles, low sequence diversity in GCs can limit its effectiveness. Our method is affected by the presence and propagation of contamination in reference databases, a major problem in 'omics [63,64]. In our case, we only use Pfam as a source for annotation owing to its high-quality and manual curation process. The categorization process of our GCs depends on the information from other databases, and to minimize the potential impact of contamination, we apply methods that weight the annotations of the identified homologs to discriminate if a GC belongs to the known or unknown CDS-space. We foresee the integration of our approach to assist in the manual curation process and increase the quality of the recovered MAGs[65].

The work presented here should incentivize the scientific community to build a common effort to define the different levels of unknown[66] where clear guidelines and protocols should be established. Our work proves that the integration of the unknown fraction is possible and aims to provide a new brighter future for microbiome analyses.

Material and methods

Genomic and metagenomic dataset

We used a set of 583 marine metagenomes from four of the major metagenomic surveys of the ocean microbiome: Tara Oceans expedition (TARA)[66], Malaspina expedition[67], Ocean Sampling Day (OSD)[67], and Global Ocean Sampling Expedition (GOS)[68]. We complemented this set with 1,246 metagenomes obtained from the Human Microbiome Project (HMP) phase I and II[69]. We used the assemblies provided by TARA, Malaspina, OSD and HMP projects and the long Sanger reads from GOS[70]. A total of 156M (156,422,969) contigs and 12.8M long-reads were collected (Supplementary Table 6).

For the genomic dataset, we used the 28,941 prokaryotic genomes (27,372 bacterial and 1,569 archaeal) from the Genome Taxonomy Database[11] (GTDB) Release 03-RS86 (19th August 2018).

Computational workflow development

We implemented a computation workflow based on Snakemake[71] for the easy processing of large datasets in a reproducible manner. The workflow provides three different strategies to analyze the data. The module *DB-creation* creates the gene cluster database, validates and partitions the gene clusters (GCs) in the main functional categories. The module *DB-update* allows the integration of new sequences (either at the contig or predicted gene level) in the existing gene cluster database. In addition, the workflow has a *profile-search* function to quickly screen the gene cluster PSSM profiles in the database

Metagenomic and genomic gene prediction

We used Prodigal (v2.6.3)[72] in metagenomic mode to predict the genes from the metagenomic dataset. For the genomic dataset, we used the gene predictions provided by Annotree[46], since they were obtained, consistently, with Prodigal v2.6.3. We identified potential spurious genes using the *AntiFam* database[74]. Furthermore, we screened for '*shadow' genes* using the procedure described in Yooseph et al.[75].

PFAM annotation

We annotated the predicted genes using the *hmmsearch* program from the *HMMER* package (version: 3.1b2)[76] in combination with the Pfam database v31[76] We kept the matches exceeding the internal gathering threshold and presenting an independent e-value < 1e-5 and coverage > 0.4. In addition, we took in account multi-domain annotations and we removed overlapping annotations when the overlap is larger than 50%, keeping the ones with the smaller e-value.

Determination of the gene clusters

We clustered the metagenomic predicted genes using the cascaded-clustering workflow of the MMseqs2 software[77] ("*--cov-mode 2 -c 0.8 --min-seq-id 0.3"*). We discarded from downstream analyses the singletons and clusters with a size below a threshold identified after applying a broken-stick model[77]. We integrated the genomic data into the metagenomic cluster database using the "DB-update" module of the workflow. This module uses the *clusterupdate* module of MMseqs2[78], with the same parameters used for the metagenomic clustering.

Quality-screening of gene clusters

We examined the GCs to ensure their high intra-cluster homogeneity. We applied two methodologies to validate their cluster sequence composition and functional annotation homogeneity. We identified non-homologous sequences inside each cluster combining the identification of a new cluster representative sequence via a sequence similarity network (SSN) analysis, and the investigation of intra-cluster multiple sequence alignments (MSAs), given the new representative. Initially, we generated a SSN for each cluster, using the semi-global alignment methods implemented in *PARASAIL*[78] (version 2.1.5). The SSN was then trimmed, using a custom algorithm[79,80] that removes edges while maintaining the network structural integrity and obtaining the smallest connected graph formed by a single component. Finally, the new cluster representative was identified as the most central node of the trimmed SSN by the eigenvector centrality algorithm as implemented in igraph[81]. After this step, we built a multiple sequence alignment for each cluster using *FAMSA*[82] (version 1.1). Then, we screened each cluster-MSA for non-homologous sequences to the new cluster representative. Owing to computational limitations we used two different approaches to screen the cluster-MSAs. We used *LEON-BIS*[83] for the clusters with a size ranging from 10 to 1,000 genes and OD-SEQ[84] for the clusters with more than 1,000 genes. In the end, we applied a broken-stick model[77] to determine the cut-off number above which a cluster is identified as discarded.

The predicted genes can have multi-domain annotations in different orders, therefore to validate the consistency of intra-cluster Pfam annotations, we applied a combination of w-shingling[85] and Jaccard similarity. We used w-shingling (k-shingle = 2) to group consecutive domain annotations as a single object. We measured the homogeneity of the *shingle sets* (sets of domains) between genes using the Jaccard similarity and reported the median similarity value for each cluster. Moreover, we took into consideration the Clan membership of the Pfam domains and that a gene might contain N-, C- and M-terminal domains for the functional homogeneity validation. Clusters with a median similarity < 1 were discarded.

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

Remote homology classification of gene clusters

To partition the validated GCs into the four main categories we processed the set of GCs containing Pfam annotated genes and the set of not annotated GCs separately. For the annotated GCs, we inferred a consensus protein domain architecture (DA) (an ordered combination of protein domains) for each annotated gene cluster. To identify each gene cluster consensus DA, we created directed acyclic graphs connecting the Pfam domains based on their topological order on the genes using *igraph*[83]. We collapsed the repetitions of the same domain. Then we used the gene completeness as a positive-weighting value for the selection of the cluster consensus DA. Within this step we divided the GCs into "Knowns" (Known) if annotated to at least one Pfam domains of known function (DKFs), and "Genomic unknowns" (GU) if annotated entirely to Pfam domains of unknown function (DUFs).

We aligned the sequences of the non-annotated GCs with FAMSA[85] and obtained cluster consensus sequences with the *hhconsensus* program from *HH-SUITE*[86]. We used the cluster consensus sequences to perform a nested search against the UniRef90 database (release 2017_11)[86] and NCBI *nr* database (release 2017_12)[87] to retrieve non-Pfam annotations with *MMSeqs2*[88] ("*-e 1e-05 --cov-mode 2 -c 0.6*"). We kept the hits within 60% of the Log(best-e-value) and searched the annotations for any of the terms commonly used to define proteins of unknown function (Supp table 12). We used a quorum majority voting approach to decide if a gene cluster would be classified as *Genomic Unknown* or *Known without Pfams* based on the annotations retrieved. We searched the consensus sequences without any homologs in the UniRef90 database against NCBI *nr*. We applied the same approach and criteria described for the first search. Ultimately, we classified as *Environmental Unknown* those GCs whose consensus sequences did not align to any of the NCBI *nr* entries.

In addition, we developed some conservative measures to control the trade-off between specificity and sensitivity for the remote homology searches such as (1) a modification of the algorithm described in Hingamp et al.[88] to get a confident group of homologs to determine if a query protein is known or unknown by a quorum majority voting approach (Supp Note 3); (2) strict parameters in terms of iterations, bidirectional coverage and probability thresholds for the HHblits alignments to minimize the inclusion of non-homologous sequences; and (3) avoid providing annotations for our gene clusters, as we believe that annotation should be a careful process done on a smaller scale and with experimental context.

Gene cluster remote homology refinement

We refined the *Environmental Unknown* GCs to ensure the lack of any characterization by searching for remote homologies in the Uniclust database (release 30_2017_10) using the HMM/HMM alignment method *HHblits*[89]. We created the HMM profiles with the *hhmake* program from the *HH-SUITE*[89]. We only accepted those hits with a *HHblits-probability* ≥ 90% and we re-classified them following the same majority vote approach as previously described. The clusters with no hits remained as the refined set of EUs. We applied a similar refinement approach to the KWP clusters to identify GCs with remote homologies to Pfam protein domains. The KWP HMM profiles were searched against the Pfam *HH-SUITE* database (version 31), using *HHblits*. We accepted hits with a probability ≥ 90% and a target coverage > 60% and removed overlapping domains as described earlier. We moved the KWP with remote homologies to known Pfams to the Known set, and those showing remote homologies to Pfam DUFs to the GUs. The clusters with no hits remained as the refined set of KWP.

Gene cluster characterization

To retrieve the taxonomic composition of our clusters we applied the *MMseqs2 taxonomy* program (version: b43de8b7559a3b45c8e5e9e02cb3023dd339231a), which allows computing the lowest common ancestor through the implementation of the 2bLCA protocol [88]. We searched all cluster genes against UniProtKB (release of January 2018) [90]. The taxonomic search was then performed using "*-e 1e-05 --cov-mode 0 -c 0.6"*. We parsed the results to keep only the hits within the 60% of the log10(best-e-value). To retrieve the taxonomic lineages we used the R package *CHNOSZ*[91]. We measured the intra-cluster taxonomic admixture by applying the *entropy.empirical()* function from the *entropy* R package [92]. This function estimates the Shannon entropy based on the different taxonomic annotation frequencies. For each cluster we also retrieved the cluster consensus taxonomic annotation, which we defined as the taxonomic annotation of the majority of the genes in the cluster.

In addition to the taxonomy, we evaluated the clusters level of darkness and disorder using the Dark Proteome Database (DPD)[92] as reference. We searched the cluster genes against the DPD, applying the MMseqs2 search program[88] with "*-e 1e-20 --cov-mode 0 -c 0.6*". For each cluster we then retrieved the mean and the median level of darkness, based on the gene DPD annotations.

High-quality clusters

We defined a subset of high-quality clusters based on the completeness of the cluster genes and their representatives. We identified the minimum required percentage of complete genes per cluster by a broken-stick model[77] applied to the percentage distribution. Then, we selected the GCs found above the threshold and with a complete representative.

A set of non-redundant domain architectures

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

Inference of gene cluster communities

We aggregated distant homologous GCs into GCCs. The community inference approach combined an all-vs-all HMM gene cluster comparison with Markov Cluster Algorithm (MCL)[89] community identification.

We started performing the inference on the Known GCs to use the Pfam DAs as constraints. We aligned the gene cluster HMMs using HHblits[93] (-n 2 -Z 10000000 -B 10000000 -e 1) and selected the cluster pairs with probability ≥ 50% and bidirectional coverage > 60% were used to build a homology graph. We used the ratio between HHblits-bitscore and aligned-columns as the edge weights (Supp. Note 9). We used MCL[91] (v. 12-068) to identify the communities present in the graph. We developed an iterative method to identify the optimal MCL inflation parameter that tries to maximize the relationship of five intra-/inter-community properties: (1) the proportion of MCL communities with one single DA, based on the consensus DAs of the cluster members; (2) the proportion of MCL communities with more than one cluster; (3) the proportion of MCL communities with a PFAM clan entropy equal to 0; (4) the intra-community HHblits-score/Aligned-columns score (normalized by the maximum value); and (5) the number of MCL communities, which should, at the end, reflect the number of non-redundant DAs. We iterated through values ranging from 1.2 to 3.0, with incremental steps of 0.1. During the inference process, some of the GCs became orphans in the graph. We applied a three-step approach to assign a community membership to these GCs. First, we applied less stringent conditions (probability ≥ 50% and coverage >= 40%) to find homologs in the already existing GCCs. Then, we ran a second iteration to find secondary relationships between the newly assigned GCs and the missing ones. Lastly, we created new communities with the remaining GCs. We repeated the whole process with the other categories (KWP, GU and EU), applying the optimal inflation value found for the Known (2.2 for metagenomic and 2.5 for genomic data).

Gene cluster communities validation

We tested the biological significance of the GCCs using the phylogeny of proteorhodopsin[93] (PR). We used the proteorhodopsin HMM profiles[92] to screen the marine metagenomic datasets using *hmmsearch* (version 3.1b2)[94]. We kept the hits with a coverage > 0.4 and e-value <= 1e-5. We removed identical duplicates from the sequences assigned to PR with CD-HIT[95] (v4.6) and cleaned from sequences with less than 100 amino acids. To place the identified PR sequences into the MicRhode[95] PR tree first we optimized the initial tree parameters and branch lengths with RAxML (v8.2.12)[96]. We used PaPaRA (v2.5)[97] to incrementally align the query PR sequences against the MicRhode PR reference alignment and *pplacer*[96] (v1.1.alpha19-0-g807f6f3) to place the sequences into the tree. Finally, we assigned the query PR sequences to the MicRhode PR Superclusters based on the phylogenetic placement. As an additional evaluation, we investigated the distributions of standard GCCs and HQ GCCs within ribosomal protein families. We obtained the ribosomal proteins used for the analysis combining the set of 16 ribosomal proteins from Méheust et al.[98] and those contained in the collection of bacterial single-copy genes of

Anvi'o[99]. In addition, for the ribosomal proteins, we compared the outcome of our method to the one proposed by Méheust et al.[99] (Supp. Note 9).

Rate of genomic and metagenomic gene clusters accumulation

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

Gene cluster abundance profiles in genomes and metagenomes

We estimated abundance profiles for the metagenomic cluster categories using the read coverage to each predicted gene as a proxy for abundance. We calculated the coverage by mapping the reads against the assembly contigs using the *bwa-mem* algorithm from *BWA mapper*[100]. Then, we used *BEDTOOLS*[101], to find the intersection of the gene coordinates to the assemblies, and normalize the per-base coverage by the length of the gene. We calculated the cluster abundance in a sample as the sum of the cluster gene abundances in that sample, and the cluster category abundance in a sample as the sum of the cluster abundances. We obtained the proportions of the different gene cluster categories applying a total-sum-scaling normalization. For the genomic abundance profiles, we used the number of genes in the genomes and normalized by the total gene counts per genome.

Occurrence of gene clusters in the environment

We used 1,264 metagenomes from the TARA Oceans, MALASPINA Expedition, OSD2014 and HMP-I/II to explore the properties of the unknown CDS-space in the environment. We applied the Levins Niche Breadth (NB) index[102] to investigate the GCs and GCCs environmental distributions. We removed the GCs and cluster communities with a mean relative abundance < 1e-5. We followed a divide-and-conquer strategy to avoid the computational burden of generating the null-models to test the significance of the distributions owing to the large number of metagenomes and GCs. First, we grouped similar samples based on the gene cluster content using the Bray-Curtis dissimilarity[103] in combination with the *Dynamic*

Tree Cut[103] R package. We created 100 random datasets picking up one random sample from each group. For each of the 100 random datasets we created 100 random abundance matrices using the *nullmodel* function of the *quasiswap* count method[104]. Then we calculated the *observed* NB and obtained the 2.5% and 97.5% quantiles based on the randomized sets. We compared the observed and quantile values for each gene cluster, and defined it to have a *Narrow distribution* when the *observed* was smaller than the 2.5% quantile and to have a *Broad distribution* when it was larger than the 97.5% quantile. Otherwise we classified the cluster as *Non significant*. We used a majority voting approach to get a consensus distribution classification based on the 10 random datasets.

Identification of prophages in genomic sequences

We used PhageBoost (https://github.com/ku-cbd/PhageBoost/) 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[46] to identify the lineage-specific GCs and to which rank they are specific. Following a similar criteria to the ones used in Mendler et al.[46], 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*[47] function implemented in the R package *castor*[47]. 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)[50] into our cluster database using the same procedure as for the genomic data. We evaluated the uncharacterized fraction and the genes classified into the eggNOG[51] category S within the context of our database.

Augmenting experimental data

We searched the 37,684 genes of unknown function associated with mutant phenotypes from Price et al.[13] against our gene cluster profiles. We kept the hits with e-value ≤ 1e-20 and a query coverage > 60%. Then we filtered the results to keep the hits within 90% of the Log(best-e-value), and we used a majority vote function to retrieve the consensus category for each hit. Lastly, we selected the best-hits based on the smallest e-value and the largest query and target coverage values. We used the fitness values from the RB-TnSeq experiments from Price et al. to identify genes of unknown function that are important for fitness under certain experimental conditions.

Code and data availability

All the code used for the analyses is available at https://github.com/functional-dark-side/functional-dark-side.github.io/tree/master/scripts. We also provide a website https://dark.metagenomics.eu with detailed descriptions of the methods applied in this paper and a wider overview of the results.

Data files for the MG+GTDB database are available at (https://doi.org/10.6084/m9.figshare.12459056).

AGNOSTOS is available at https://github.com/functional-dark-side/agnostos-wf. The workflow can be used to create a database of categorized GCs and GCCs from genomes and metagenomic assemblies. We also compiled a seed database that can be used to integrate new genomic or metagenomic data. The database can be downloaded from https://doi.org/10.6084/m9.figshare.12459056.

Supplementary information is available at https://doi.org/10.6084/m9.figshare.12588263.

Acknowledgments

The author thankfully acknowledges 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) and the University of Oxford Advanced Research Computing (http://dx.doi.org/10.5281/zenodo.22558)

References

1. Hug, L. A. *et al.* A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).

2. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science*

348, 1261359 (2015).

3. Kopf, A. *et al.* The ocean sampling day consortium. *Gigascience* **4**, 27 (2015).
4. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).
5. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e20 (2019).
6. Pachiadaki, M. G. *et al.* Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell* **179**, 1623-1635.e11 (2019).
7. Cross, K. L. *et al.* Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. *Nat. Biotechnol.* **37**, 1314–1321 (2019).
8. Eloe-Fadrosh, E. A. *et al.* Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. *Nat. Commun.* **7**, 10476 (2016).
9. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
10. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
11. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
12. Bernard, G., Pathmanathan, J. S., Lannes, R., Lopez, P. & Bapteste, E. Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. *Genome Biol. Evol.* **10**, 707–715 (2018).
13. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557**, 503–509 (2018).
14. Carradec, Q. *et al.* A global ocean atlas of eukaryotic genes. *Nat. Commun.* **9**, 373 (2018).
15. Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* **48**,

D570–D578 (2020).

16. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

17. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).

18. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).

19. Chen, I.-M. A. *et al.* IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–D677 (2019).

20. Hanson, A. D., Pribat, A., Waller, J. C. & Crécy-Lagard, V. de. 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list--and how to find it. *Biochem. J* **425**, 1–11 (2010).

21. Arnold, F. H. Design by Directed Evolution. *Acc. Chem. Res.* **31**, 125–131 (1998).

22. Brandenberg, O. F., Fasan, R. & Arnold, F. H. Exploiting and engineering hemoproteins for abiological carbene and nitrene transfer reactions. *Curr. Opin. Biotechnol.* **47**, 102–111 (2017).

23. Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem. Int. Ed Engl.* **57**, 4143–4148 (2018).

24. Jaroszewski, L. *et al.* Exploration of uncharted regions of the protein universe. *PLoS Biol.* **7**, (2009).

25. Buttigieg, L. P. *et al.* Ecogenomic Perspectives on Domains of Unknown Function: Correlation-Based Exploration of Marine Metagenomes. *PLoS One* **8**, (2013).

26. Yooseph, S. *et al.* The Sorcerer II global ocean sampling expedition: Expanding the universe of protein families. *PLoS Biol.* **5**, 0432–0466 (2007).

27. Wyman, S. K., Avila-Herrera, A., Nayfach, S. & Pollard, K. S. A most wanted list of conserved microbial protein families with no known domains. *PLoS One* **13**, e0205749 (2018).

28. Brum, J. R. *et al.* Illuminating structural proteins in viral "dark matter" with metaproteomics. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 2436–2441 (2016).

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

30. Lobb, B., Kurtz, D. A., Moreno-Hagelsieb, G. & Doxey, A. C. Remote homology and the functions of metagenomic dark matter. *Front. Genet.* **6**, 1–12 (2015).

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

32. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng. Des. Sel.* **12**, 85–94 (1999).

33. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **advance on**, (2017).

34. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019).

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

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

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

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

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

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

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

42. La Cono, V. *et al.* Partaking of Archaea to biogeochemical cycling in oxygen-deficient zones of

meromictic saline Lake Faro (Messina, Italy). *Environ. Microbiol.* **15**, 1717–1733 (2013).

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

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

45. Ma, Y. *et al.* Human papillomavirus community in healthy persons, defined by metagenomics analysis of human microbiome project shotgun sequencing data sets. *J. Virol.* **88**, 4786–4797 (2014).

46. Mendler, K. *et al.* AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Res.* **47**, 4442–4448 (2019).

47. Martiny, A. C., Treseder, K. & Pusch, G. Phylogenetic conservatism of functional traits in microorganisms. *ISME J.* **7**, 830–838 (2013).

48. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).

49. Anantharaman, K. *et al.* Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. *ISME J.* **12**, 1715–1728 (2018).

50. Salazar, G. *et al.* Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* **179**, 1068-1083.e21 (2019).

51. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

52. Heffernan, B., Murphy, C. D. & Casey, E. Comparison of planktonic and biofilm cultures of Pseudomonas fluorescens DSM 8341 cells grown on fluoroacetate. *Appl. Environ. Microbiol.* **75**, 2899–2907 (2009).

53. Scales, B. S., Dickson, R. P., LiPuma, J. J. & Huffnagle, G. B. Microbiology, genomics, and clinical significance of the Pseudomonas fluorescens species complex, an unappreciated colonizer of

humans. *Clin. Microbiol. Rev.* **27**, 927–948 (2014).

54. Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 2542 (2018).

55. Francino, M. P. The ecology of bacterial genes and the survival of the new. *Int. J. Evol. Biol.* **2012**, 394026 (2012).

56. Muller, E. E. L. Determining Microbial Niche Breadth in the Environment for Better Ecosystem Fate Predictions. *mSystems* **4**, (2019).

57. Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I. & Watson, M. A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. *Front. Genet.* **8**, 23 (2017).

58. Ivanova, N. N. *et al.* Stop codon reassignments in the wild. *Science* **344**, 909–913 (2014).

59. Steinegger, M., Mirdita, M. & Söding, J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods* **16**, 603–606 (2019).

60. Titus Brown, C. *et al.* Exploring neighborhoods in large metagenome assembly graphs reveals hidden sequence diversity. *bioRxiv* 462788 (2018) doi:10.1101/462788.

61. Höps, W., Jeffryes, M. & Bateman, A. Gene Unprediction with Spurio: A tool to identify spurious protein sequences. *F1000Res.* **7**, 261 (2018).

62. Heinzinger, M. *et al.* Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* **20**, 723 (2019).

63. Breitwieser, F. P., Pertea, M., Zimin, A. & Salzberg, S. L. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.* (2019) doi:10.1101/gr.245373.118.

64. Steinegger, M. & Salzberg, S. L. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol.* **21**, 115 (2020).

65. Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).

66. Thomas, A. M. & Segata, N. Multiple levels of the unknown in microbiome research. *BMC Biol.* **17**, 48 (2019).

67. Duarte, C. M. Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation Expedition. *Limnol. Oceanog. Bull.* **24**, 11–14 (2015).

68. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol.* **5**, 1–34 (2007).

69. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).

70. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).

71. Köster, J. Reproducible data analysis with Snakemake. *F1000Res.* **7**, (2018).

72. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119–119 (2010).

73. Eberhardt, R. Y. *et al.* AntiFam: a tool to help identify spurious ORFs in protein annotation. *Database* **2012**, bas003–bas003 (2012).

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

75. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).

76. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).

77. Bennett, K. D. Determination of the number of zones in a biostratigraphical sequence. *New Phytol.* **132**, 155–170 (1996).

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

79. Žure, M., Fernandez-Guerra, A., Munn, C. B. & Harder, J. Geographic distribution at subspecies

resolution level: closely related Rhodopirellula species in European coastal sediments. *ISME J.* **11**, 478–489 (2017).

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

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

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

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

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

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

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

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

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

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

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

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

92. Hausser, J. & Strimmer, K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *arXiv [stat.ML]* (2008).

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

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

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

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

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

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

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

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

101. Levins, R. The strategy of model building in population biology. *Am. Sci.* **54**, 421–431 (1966).

102. Bray, J. R., Roger Bray, J. & Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs* vol. 27 325–349 (1957).

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

104. Miklós, I. & Podani, J. Randomization of presence-absence matrices: comments and new algorithms. *Ecology* vol. 85 86–92 (2004).