Network studies unveil new groups of highly divergent proteins in families as old as cellular life with important biological functions in the ocean

Duncan Sussfeld[1,2,*], Romain Lannes[1], Eduardo Corel[1], Guillaume Bernard[1], Pierre Martin[1], Eric Bapteste[1], Eric Pelletier[2,3], Philippe Lopez[1]

[1]Institut de Systématique, Evolution, Biodiversité (ISYEB), Sorbonne Université, CNRS, Museum National d'Histoire Naturelle, EPHE, Université des Antilles, Paris, France

[2]Génomique Métabolique, Genoscope, Institut François-Jacob, CEA, CNRS, Université d'Evry, Université Paris-Saclay, 91000 Evry, France.

[3]Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE, 75016 Paris, France.

***Corresponding author**: Duncan Sussfeld, duncan.sussfeld@univ-evry.fr

Keywords: Microbial dark matter, Sequence similarity networks, Distant homology, Microbiome

Abstract

Metagenomics has considerably broadened our knowledge of microbial diversity, unravelling fascinating microbial adaptations and characterising multiple novel major taxonomic groups, e.g. CPR bacteria, DPANN and Asgard archaea, and novel viruses. Such findings profoundly reshaped the structure of the tree of life and emphasised the central role of investigating uncultured organisms. However, despite significant progresses, a large portion of proteins predicted from metagenomes remain today unannotated, both taxonomically and functionally, across many biomes and in particular in oceanic waters, including at relatively lenient clustering thresholds. Here, we used an iterative network-based approach to probe a dataset of 40 million ORFs, predicted in marine environments. We assessed the environmental diversity of a selection of 53 gene families as old as cellular life, broadly distributed across the Tree of Life. About half of them harboured clusters of environmental homologs that diverged significantly from the known diversity of published complete genomes, with representatives distributed across all the ocean. Network analyses also enabled the detection of environmental clades with new structural variants of essential genes (SMC), divergent polymerase subunits forming deep branching clades in the polymerase tree, and potentially deep taxonomic groups at the basis of Archaea or Bacteria, from ultra-small microorganisms distinct from CPR and DPANN. These diverse results provide a strong proof of concept for the capabilities of sequence network approaches to make better sense of microbial dark matter across taxonomical scales. Still, further methodological developments appear warranted in order to better comprehend the most divergent members of the oceanic microbial dark matter.

Introduction

Over the last decades, novel sequencing methods have allowed microbiologists to appreciate the ubiquity and abundance of uncultured organisms [1–5] and access microorganisms' genomes beyond the restrictions imposed by the isolation and cultivation dogma issued from the Koch principles [6] that underpinned microbiological studies for decades. Metagenomic studies [7] have led to an unprecedented broadening of our knowledge of microbial diversity [8], from the unravelling of microbial adaptations and interactions in numerous environments [9–12] to the characterisation of multiple novel major taxonomic groups [13–16] – most notably CPR bacteria [13, 17, 18], DPANN archaea [17, 19, 20] and Asgard archaea [21–23], profoundly reshaping the structure of the tree of life. Large groups of novel viruses [24–26] and mobile elements [27] have also been unearthed. Together, these major discoveries emphasise the central role of investigating yet uncultured organisms, believed to constitute the majority of overall microbial lineages [3, 28], in addressing many fundamental questions of biology and evolutionary biology.

Over time, as cultivation-independent sequencing efforts are carried out in an increasing range of ecosystems, discovery events of novel branches near the base of the tree of life are predicted to become less frequent [8]. In accordance with this perspective, an extensive study of over 50,000 MAGs, assembled from a vast ensemble of metagenomes and including 12,556 novel candidate species-level OTUs, found no reliable evidence of novel prokaryote phylum content [29].

Contrasting with these observations, it still persists that a large portion of proteins predicted from metagenomes remain unannotated, both taxonomically and functionally, across most biomes, including at relatively lenient clustering thresholds [30]. This vast pool of uncharacterised sequences constitutes a significant blind spot in our grasp of the diversity of microbial life forms and biological processes, and interrogates the need for novel approaches complementing the current techniques to mine metagenomes for highly divergent groups.

In particular, interrogations remain on the ability of current multiple alignment-based methods to comprehensively classify the vast diversity of uncharacterised environmental sequences [31, 32]. In addition to potential tractability issues owed to the vast amounts of data to process, a more fundamental complication stems from the sequences themselves, which lack taxonomical and/or functional annotation precisely due to a higher divergence from well-characterised sequences. It may then prove challenging to produce significant alignments between primary sequences of cultured organisms and their divergent environmental variants, just as it may prove challenging to reconstruct a tree of life with both well-known and uncharacterised extant taxa, based on single genes.

Accordingly, some alternative strategies have emerged to explore the unknown microbial matter. Various network-based approaches [33], in particular, have been developed to address these methodological concerns. Co-occurrence networks can help assessing ecological roles of unknown taxa [34]. Sequence similarity networks (SSNs), wherein pairs of

primary sequences are connected depending on given similarity criteria, can also be employed to compare sequences from cultured and uncultured organisms [35, 36]. In 2012, Lynch et al. used SSNs to identify several candidate new lineages from environmental 16S rRNA [37]. In 2015, Lopez et al. designed an SSN-based exploratory analysis to probe metagenomes for distant homologs of well-distributed gene families [38]. In this study, 86 clusters of genes broadly distributed across domains of life were used as seeds for a two-step BLAST search inside a metagenome collection. Seed sequences were then gathered in SSNs together with their direct and indirect environmental homologs. Subsequent analysis of these networks confirmed that environmental sequences gathered in the second alignment step were more divergent from their cultured relatives than those gathered in the first round, and found several hundred groups of highly divergent environmental variants, some of them potentially compatible with novel major divisions of life. Consequently, (i) iterative explorations of environmental datasets may allow the retrieval of increasingly divergent variants (Fig. 1), and (ii) network-based methods may be well-suited to handle this type of data, by integrating sequences with various levels of divergence within homologous gene families.

Accordingly, we further developed the iterative explorative strategy of environmental sequence datasets by Lopez et al. [38], by allowing distant homolog search iterations to continue indefinitely until convergence. Convergence was reached after 7 rounds on average for 53 old, conserved, functionally important gene families. Thus, we identified highly divergent variants of protein families as old as cellular life in a large oceanic metagenome dataset, uncovering new structural and sequence variants of biologically important genes and new taxonomical groups bearing divergent genes across taxonomical scales.

Results and Discussion

Oceanic metagenomes harbour distant homologs of highly conserved protein families

We developed an iterative mining procedure to accumulate highly divergent environmental variants for families of genes or proteins of interest. From an initial set of 9,737,821 protein sequences gathered from a broad diversity of prokaryotic, eukaryotic, viral and plasmidic complete genomes (Table SI-1), we derived a collection of 891,459 protein clusters. We selected those displaying both a broad taxonomic representation, a strong evolutionary conservation and minimal evidence of horizontal transfer between Domains of life, i.e., gene families at least as old as cellular life. These families represent the best candidate seeds for identifying novel important lineages, potentially deep branching, in distant environmental homologs [38]. This resulted in a final set of 125,774 sequences distributed in 53 protein families, including 12 ribosomal proteins (Table SI-2). On average, bacterial and archaeal sequences in these families had 34.9% identity to their closest homolog in the other prokaryotic Domain, reflecting the clear distinction between sequences of archaeal and bacterial origin.

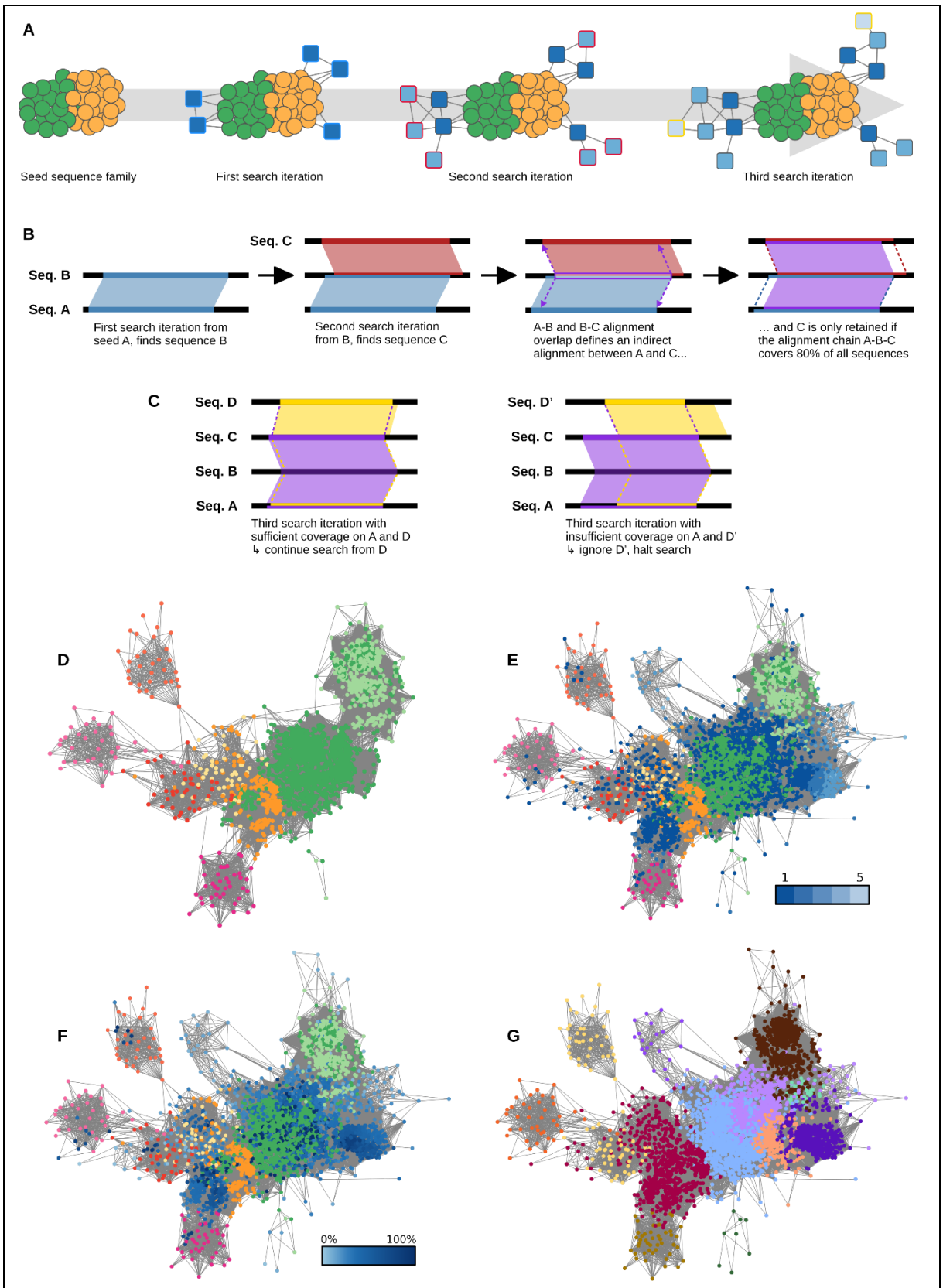


Figure 1: Iterative homolog search procedure. (**A**) Iterative aggregation of environmental homologs around seed sequences in a similarity network. From a set of seed sequences belonging to a given protein family (green and orange nodes), a first search iteration finds environmental homologs (dark blue nodes) for some of the seeds. A second search iteration then uses these environmental sequences as queries to find more homologs (medium blue nodes, red frame), which are themselves used as queries for a third search iteration finding further environmental homologs (light blue nodes, yellow frame). (**B**) At each iteration of the search, newly found homologs are only retained if their aligned region can be mapped back onto a seed sequence in a way that ensures >80% coverage on all sequences along the chain of aligned

sequences. (**C**) **Left**: sequence D is found after three search iterations from seed A, and its alignment with sequence C can be mapped back to A in a way that preserves 80% coverage on all sequences along the "alignment chain". Sequence D is therefore retained and will be used as query for the next iteration of the search. **Right**: sequence D' is found after three search iterations from seed A, but its aligned region cannot be mapped back to A without breaking the 80% coverage requirement. D' is thus not retained as a distant homolog of A in this round of search. (**D-G**) Sequence similarity networks for SMC proteins. (D) shows seed sequences only, (E-G) show seed and environmental sequences. In (D-F), nodes representing seed sequences are coloured according to their taxonomic origin (yellow: non-DPANN archaea; orange: DPANN archaea; light green: CPR bacteria; dark green: non-CPR bacteria; shades of red: four eukaryotic SMC paralogs). In (E), environmental nodes are coloured in blue, with darker shades for sequences retrieved in earlier iterations of the search, and lighter shades for sequences retrieved later. In (F), environmental nodes are coloured in blue, with darker shades for sequences with higher similarity to the known cultured diversity, and lighter shades for sequences with less similarity. In (G), all nodes are coloured according to Louvain clusters inferred in the SSN (one arbitrary colour per cluster).

Each of these selected protein families was used as the seed for a deep mining procedure (described below and in Methods) against the Ocean Microbial Reference Gene Catalog (OM-RGC) dataset [39]. This extensive, non-redundant oceanic record contains sequences for 40,154,822 proteins, predicted from metagenomic sequencing of a large variety of marine environments across the world. At the time of initial publication, around 45% of these sequences lacked taxonomical annotation at or below the Domain level and 43% lacked functional annotation to an eggNOG orthologous group (OG). This highlights the existence of a vast, undescribed diversity in the global oceanic microbiome, as well as the necessity of additional efforts to improve its characterisation.

This iterative search aims at aggregating around each seed family the diversity of its environmental homologs, including variants too divergent to produce a significant direct alignment to any seed sequence. For each family, a first round of BLAST alignment (using Diamond [40]) was performed between primary seed sequences and the OM-RGC dataset, thus retrieving direct environmental homologs to these seed sequences. Subsequent rounds of alignments were then conducted against the OM-RGC dataset, each time using as queries the environmental sequences newly retained in the previous round, to mine the oceanic dataset for increasingly divergent variants of the seed family, all aligning over a common region of the seed sequence covering >80% of its length, until the procedure converged to find no additional environmental homologs (Fig. 1). This search expanded the selected 53 seed families by a total of 826,717 environmental sequences from OM-RGC (Fig. SI-1). All seed families had their own set of environmental homologs, retrieved after an average of 7 rounds of iterative search.

Environmental homologs of these 53 families were then compared against proteins from the NCBI non-redundant (*nr*) database to find their closest relative amongst all published sequences with taxonomically resolved annotations (Fig. SI-2; Supplementary Text SI-1). Only 6.7% of all retrieved environmental sequences were >90% similar to their closest characterised relative, implying that a large majority of environmental proteins cannot be accurately represented by organisms captured by current cultivation or isolation techniques. Additionally, 20.5% of environmental variants had less than 34.9% similarity with their closest *nr* relative, i.e., they diverged more from any well-characterised organism than bacteria and archaea diverged from one another on average in the retained dataset. These remarkably divergent sequences might be indicative of substantially distant oceanic lineages, and/or of

significantly increased gene evolution rates in these universally conserved families. Interestingly, environmental homologs of ribosomal protein families had generally higher similarity to their closest characterised relative than non-ribosomal environmental sequences (one-sided Kolmogorov-Smirnov test, p-value $<1.6\ 10^{-22}$; Fig. SI-3), possibly owing to their reputedly high evolutionary conservation. Still, even ribosomal protein families included very divergent oceanic variants (Fig. SI-3). Moreover, all sampled oceanic sites revealed similar proportions (but not absolute numbers) of divergent and highly divergent sequences from prokaryotes, Bacteria and Archaea (Fig. SI-1). Thus, any location in the global ocean could be a prolific reserve of new microbial variants, including microorganisms that may belong to a different Domain that their closest known relative, as they present sequences that are as distant from this relative as bacterial sequences are on average distant from archaeal ones. Contrasting with predictions that suggest that we have now exhausted the diversity of prokaryotic major groups to be discovered on Earth, our observations indicate that the discovery of deep branching groups of both oceanic bacteria and archaea remains possible in the future.

Highly divergent clusters of environmental variants expand the diversity of multiple universal protein families

Sequences within each SSN were then partitioned into network communities using Louvain clustering [41]. This higher-level view of network topologies allows an easier assessment of the environmental diversity, including to distinguish potential novelty in these protein families from homologs closely related to known sequences. In particular, protein clusters consisting exclusively or predominantly of environmental sequences (<10% of cultured sequences), with little similarity to published sequence records (<40% sequence identity to any non-environmental sequence found in the nr database), and containing enough proteins to be unlikely to result from sequencing inaccuracies, intuitively constitute the best candidates for *bona fide* discoveries of novel groups of divergent homologs. Out of the 691 Louvain clusters of sequences inferred in total in the 53 SSNs, we retained 80 such clusters of proteins with a degree of divergence to their closest published relatives comparable to that observed between bacterial and archaeal sequences in the seed family networks. These 80 clusters of highly divergent sequences were distributed across 25 ancient, conserved protein families, suggesting that numerous key biological processes are carried out by a currently underestimated diversity of protein primary structures. In other words, the “functional dark matter” of proteins likely consists of both unknown functions and unknown actors of known functions [30, 42]. Remarkably, none of these divergent clusters were found in networks of ribosomal proteins, possibly due to a superior level of conservation or a higher coverage of their diversity in public sequence databases.

To assess where these groups of divergent sequences may branch in the tree of life, we reconstructed phylogenetic trees regrouping seed and environmental sequences from each of the 80 selected clusters (see Materials & Methods). This selection exposed an additional phylogenetic diversity in conserved protein families when environmental contributions are considered. In particular, in some families, the phylogenetic placement of some divergent

clusters, branching between or beside the main groups of archaeal and bacterial sequences, conformed to typical expectations for novel major lineages.

Detection of variant oceanic DNA polymerases

From one seed family, consisting of several AAA+ ATPases (bacterial DNA polymerase III subunits δ' (HolB) and γ/τ (DnaX); archaeal and eukaryotic replication factor C (RfcS); bacterial replication-associated recombination protein A (RarA)) mainly involved in the DNA replication clamp loading systems of their respective Domains of life [43], we found a total of 10 new network clusters of divergent environmental homologs (Fig. 2; Fig. SI-4). These included one cluster branching within archaeal and eukaryotic RfcS sequences, as well as two clusters branching within HolB/DnaX bacterial sequences.

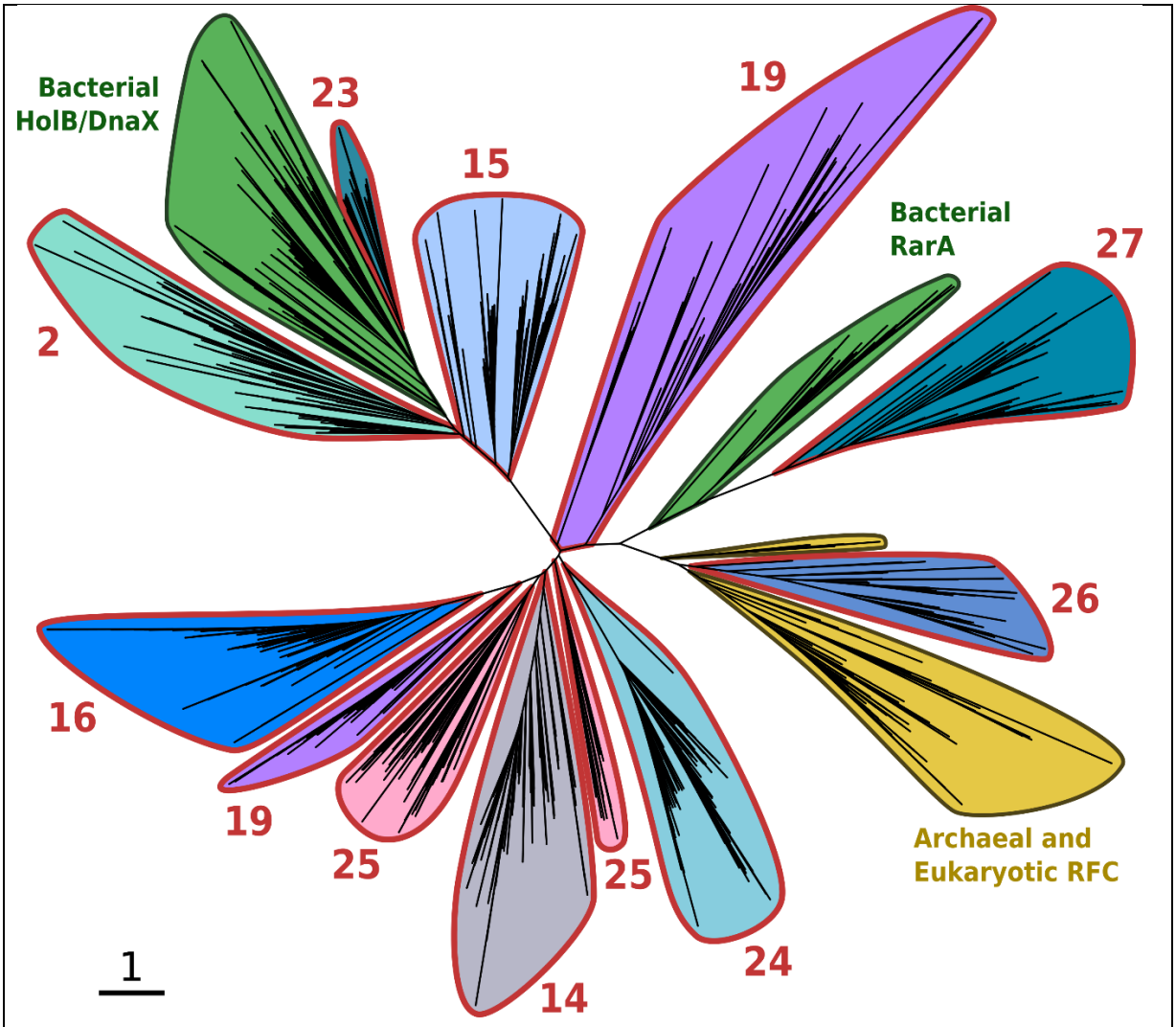


Figure 2: Alignment-free phylogeny of the DNA Pol III subunits: HolB/DnaX/RarA/RfcS sequences and environmental homologs from significantly divergent clusters. Seed sequences are coloured according to the Domain of life of their host organism (green: Bacteria, yellow: Archaea and Eukaryotes). Environmental sequences are coloured according to the network cluster they belong to in the family SSN, and outlined in red. Numerical cluster labels are shared with Fig. 3 and Fig. SI-4. Note: environmental network clusters 19 and 25 are both split into two groups in this phylogenetic tree.

Additionally, one environmental cluster branched next to bacterial RarA sequences, and its sequences were annotated as belonging to the B subunit of the Holliday junction resolving

complex RuvABC, which has already been demonstrated to cluster near clamp loading proteins in sequence networks [44]. Finally, six divergent clusters were situated outside the bacterial and archaeal/eukaryotic seed sequence clans [45] in the phylogeny. Eggnog annotations for sequences in these clusters mapped them predominantly to HolB (COG0470), though it should be noted that one particular cluster contained 96% of functionally unassigned sequences. Interestingly, many of the tertiary structures predicted for representatives of these clusters were similar to reference HolB structures from cultured organisms. However, while a few of these divergent oceanic HolB branched in a clade comprising the references in our dendrogram analyses, most variant HolB oceanic sequences branched outside this clade of references, indicating that oceanic HolB did not only diverge in primary sequence, but also in tertiary structure (Fig. 3). Such structural differences and sequence-based phylogenetic placements for those groups of highly divergent environmental HolB homologs are compatible with either highly divergent paralogs or major basal divisions of life, and are in any case indicative of a diversity of uncultivated marine organisms replicating DNA using several unusual proteic machineries, suggesting that unknown polymerase may operate in the ocean.

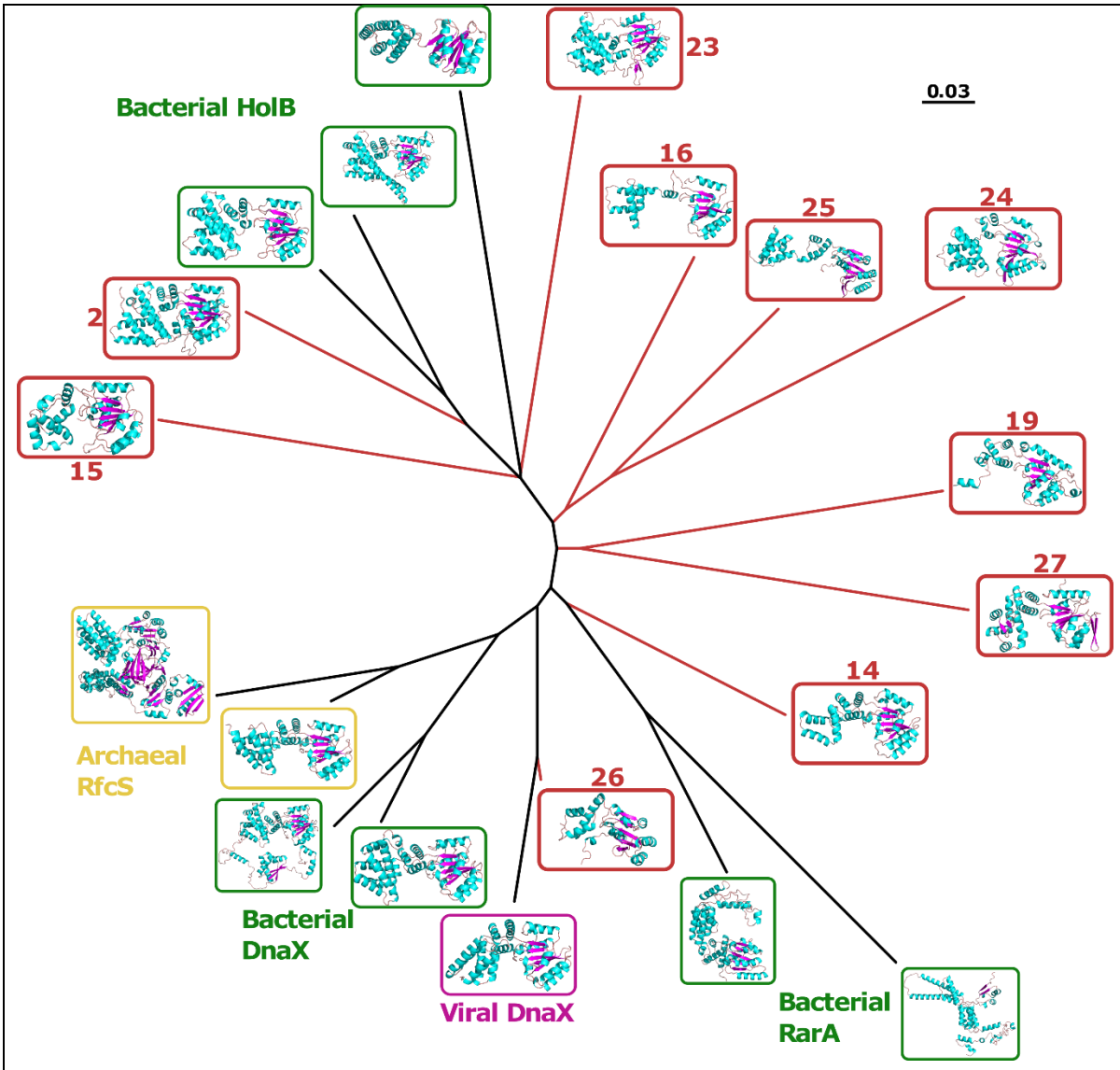


Figure 3: Dendrogram of tertiary structures of DNA Pol III subunits: HolB/DnaX/RarA/RfcS sequences and environmental homologs from significantly divergent clusters. Leaves and structures are boxed according to the Domain of life of their

host organism (green: Bacteria, yellow: Archaea, magenta: Viruses). Environmental leaves and structures are boxed in red, with numerical labels corresponding to the SSN cluster they belong to, in accordance with Fig. 2 and Fig. SI-4.

Detection of a new, abundant clade within Actinobacteria with alternative SMC proteins

A second remarkable seed family consisted of SMC (structural maintenance of chromosomes) proteins, found in all Domains of life with one copy in prokaryotes and six paralogs in eukaryotes [46]. From seed sequences in this family, we retrieved a rather limited amount of environmental homologs (0.97 environmental homolog per seed sequence in this family, compared to a median value of 2.6 across all families, see Table SI-2), though one small cluster of distant environmental homologs could still be identified. In the phylogenetic tree we produced to make sense of this cluster (Fig. 4; Fig. SI-5), environmental sequences gathered in a homogeneous clan (i.e., a subtree containing all environmental sequences and no other sequences, equivalent to a monophyletic clade in rooted phylogenies [45]) branching close to the root of seed sequences from Actinobacteria.

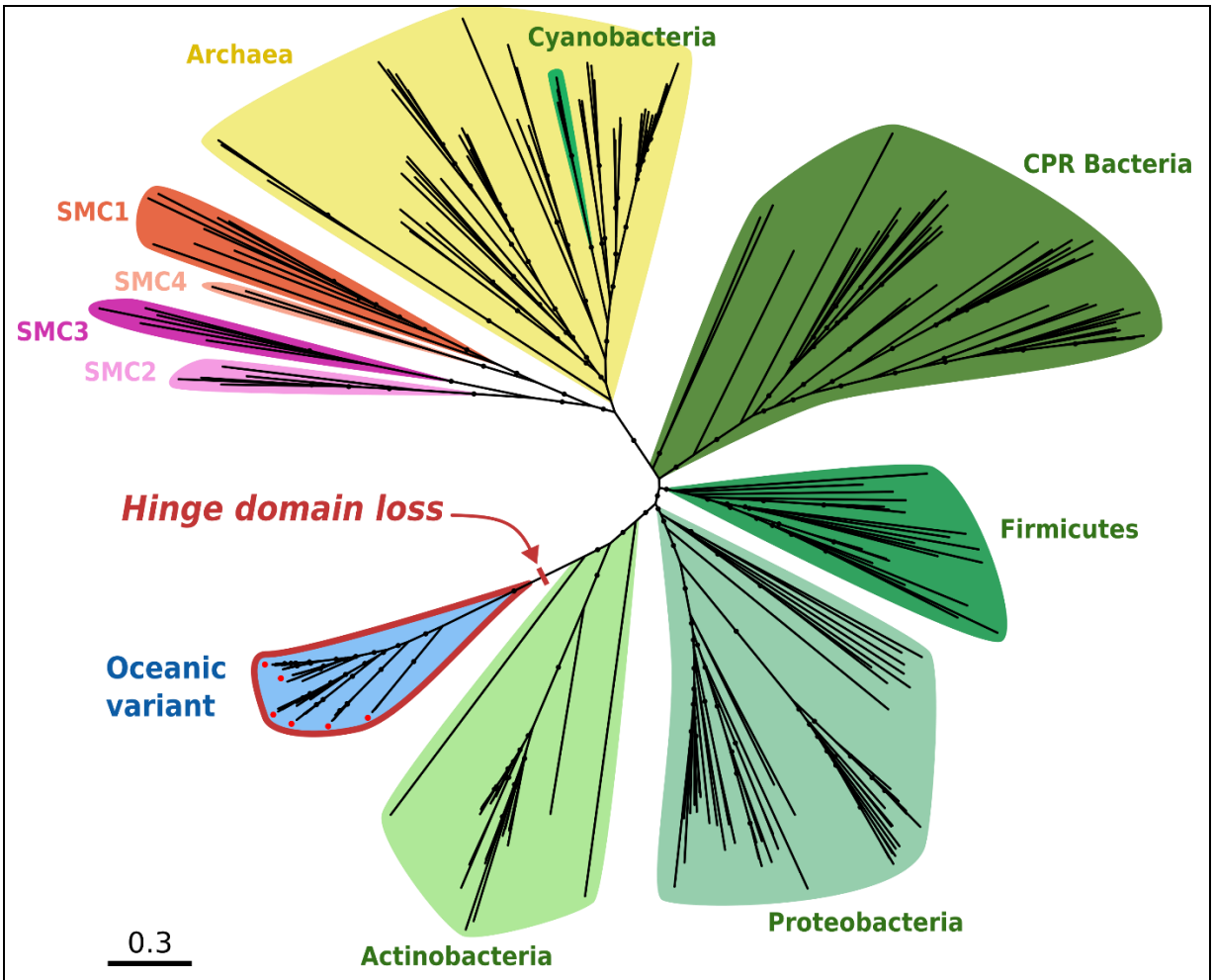


Figure 4: Maximum likelihood phylogenetic tree of SMC sequences and environmental homologs from significantly divergent clusters. Seed sequences are coloured according to the Domain of life of their host organism (green tones: Bacteria, yellow: Archaea, orange and purple tones: Eukaryotes). Environmental sequences are coloured in blue and outlined in red. Red dots indicate environmental sequences for which 3D structures were inferred. Black dots indicate branches with >85% bootstrap support.

These divergent environmental sequences were functionally annotated as SMC proteins (COG1196). Additionally, they appeared to be strikingly abundant in the sequencing data, nearly seven times more so than the rest of OM-RGC homologs of SMC proteins. Several evolutionary scenarios could be offered as to the origin of this new bacterial cluster. Firstly, it could be indicative of some paralog of SMC existing in Actinobacteria. This would then be the first description of SMC duplication in prokaryotes. Indeed, some Bacteria encode for different analogous complexes than SMC to preserve their chromosomal organisation, namely MukBEF and MksBEF [47], and at least one actinobacterium has been proven to encode both SMC and MksBEF complexes [48]; however, despite the convergent similarity between MukBEF/MksBEF and SMC proteins observed at the 3D structure level, homology between these protein families has been deemed unlikely based on primary sequence comparisons alone [46]. Alternatively, oceanic divergent SMC sequences could also stem from a different SMC duplication or lateral gene transfer from a distant, currently unknown relative. Finally, this divergent cluster could indicate the existence of an unknown lineage, supposedly branching within Actinobacteria. In any case, the substantial divergence of these environmental sequences to any gene published from a well-characterised organism, together with their remarkably high abundance in the sampling data, suggests that we identified a new kind of biology within the SMC family. Moreover, this novel oceanic clade harbours SMC-related proteins that are structurally critically different from canonical SMC proteins (Fig. 5).

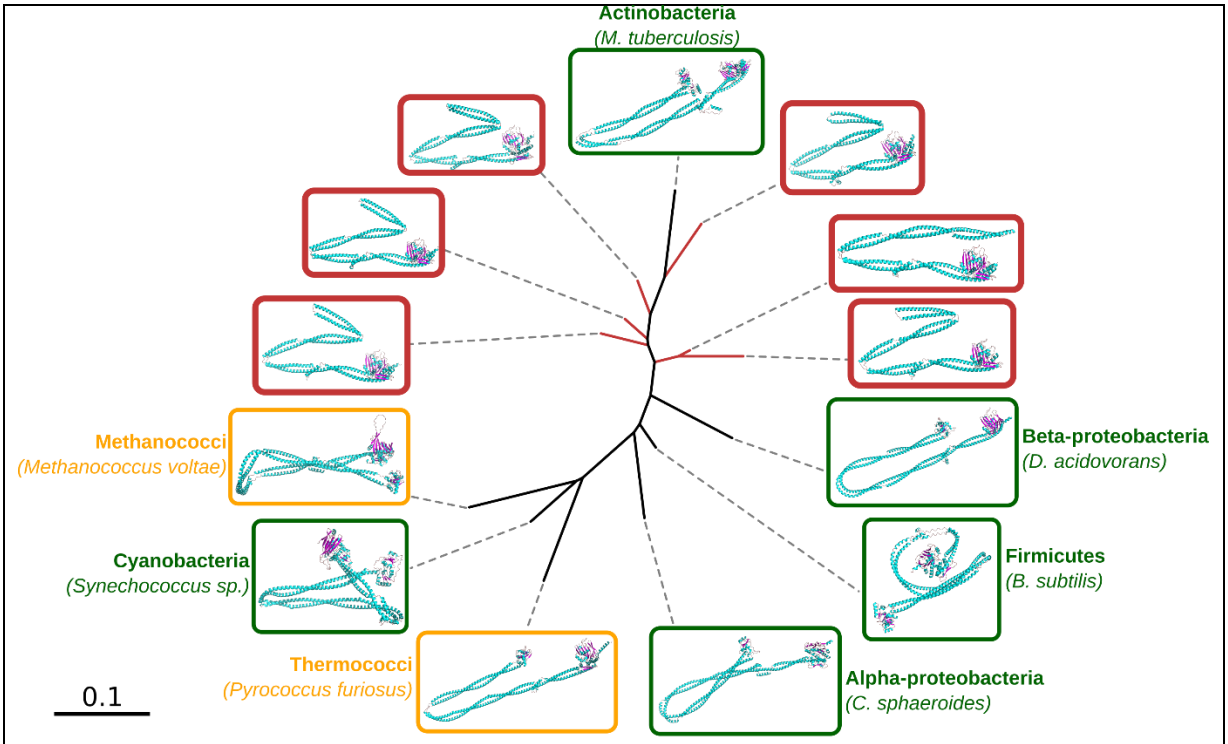


Figure 5: Dendrogram of tertiary structures of SMC sequences and selected environmental homologs from significantly divergent clusters. Leaves and structures are boxed according to the Domain of life of their host organism (green: Bacteria, yellow: Archaea). Environmental leaves and structures are highlighted in red.

Namely, these oceanic variants lack a *hinge* domain, which was experimentally demonstrated to be necessary for the interaction between DNA and SMC proteins *in vitro* in *Bacillus subtilis* [49] and in yeast [50]. On that ground, one would speculate that these

"hingeless" SMC-related oceanic proteins should either perform a different function than known SMC or bind DNA through different mechanisms. The broad distribution of hingeless SMC variants across the oceans, their monophyly and their relative abundance in the ocean microbiome suggest that hingeless SMC variants play an important, underappreciated function in this oceanic clade.

Detection of very deep branching major groups with RecA/RadA variants

In a third family, consisting of DNA recombinases (bacterial and eukaryotic (organellar) recombinase RecA; archaeal and eukaryotic (nuclear) recombinase RadA) [51], we identified other possible sources of novel diversity, including within ultrasmall cell size fractions (Fig. 6; Fig. SI-6).

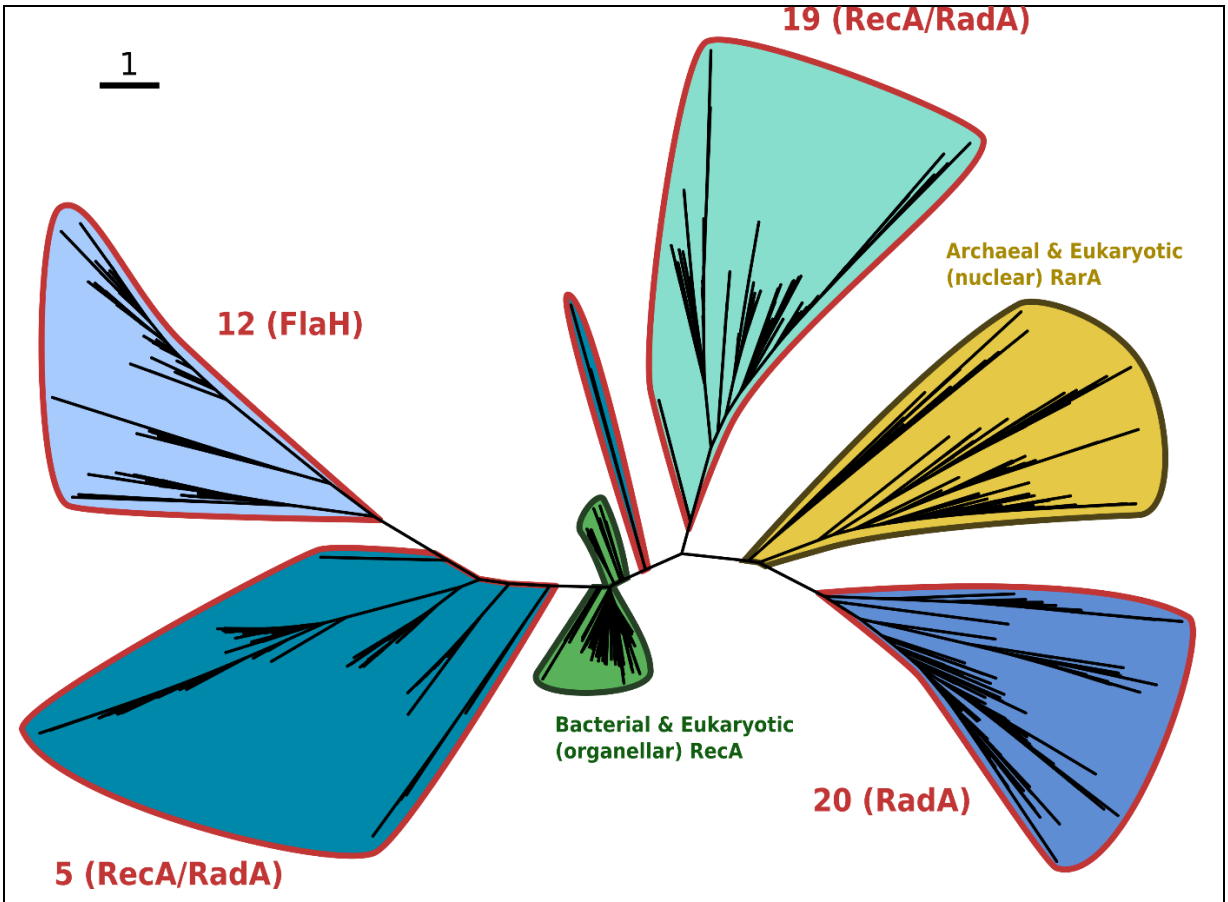


Figure 6: Alignment-free phylogeny of RecA/RadA sequences and environmental homologs from significantly divergent clusters. Seed sequences are coloured according to the Domain of life of their host organism (green: Bacteria and eukaryotic organelles, yellow: Archaea and eukaryotic nuclei). Environmental sequences are coloured according to the network cluster they belong to in the family SSN, and outlined in red. Numerical cluster labels are shared with Fig. SI-6.

Four clusters of environmental sequences in this family were retained as significantly divergent, totalling 1700 sequences. One cluster branched near the root of the archaeal seed sequence clan and was functionally categorised as RadA (COG1066) in accordance with that placement. One cluster of distant environmental RecA homologs (COG0468) branched within bacterial sequences, and a second cluster annotated as RecA branched between the archaeal and bacterial clans. Interestingly, both of these clusters were composed of >50% of sequences from the ultrasmall size fraction of cells with diameters <0.2 µm. Such cellular sizes are akin

to those of CPR bacteria and DPANN archaea [52]; however, seed sequences from these superphyla branch clearly within the clans of their respective Domains of life. Additionally, environmental sequences from these "ultrasmall clusters" bore no remarkable similarity to viral sequences recorded in the NCBI Virus sequence database (accessed in February 2023) and only 11 of them (out of 1700) matched to a single oceanic virus from the GVMAG database [53], suggesting that environmental sequences from that ultrasmall cluster may belong to new *bona fide* uncharacterised, deep-branching cellular lineages of sub-micrometre cell size. Finally, the last cluster of divergent environmental sequences branched within the environmental "ultrasmall cluster" in Bacteria. Interestingly, this cluster was predominantly annotated as ArlH (COG2874), an archaeal protein involved in the biogenesis of the archaellum, a cellular motility structure analogous to bacterial flagella [54]. Structural and sequence similarities between ArlH and bacterial RecA have previously been described [55] but, to the best of our knowledge, no evolutionary hypothesis has yet been put forth to explain this surprising homology.

Conclusion

The prevalence of biological unknowns in environmental metagenomes remains, to this day, vast; vast indeed to the extent that "known unknowns" and "unknown unknowns" constitute a relevant distinction to address the genes, organisms, processes and interactions at play in the uncultured microbial world. With our network-based, multi-marker iterative approach, we sought to understand the structure of environmental genetic variation for a range of ancient, conserved gene families. We found that environmental variants for those gene families could exist in marine microbiomes with considerable divergence to the known diversity. Moreover, these highly divergent gene sequences often organised in (sometimes vast) cohesive groups of homology, indicating that they are harboured by (sometimes vast) groups of related genomes, as illustrated by the oceanic variants of DNA polymerase, the hingeless SMC clade and deep-branching divergent RecA/RadA variants from the ultra-small size-fraction. Scenarios of evolution and diversification that could have given rise to such diversity are varied, and doubtlessly not one may underlie all of the divergent groups mentioned in this study. Some of these environmental variants may belong to uncharacterised lineages, including deep branching major groups when they diverge from any known organism to a similar or higher degree than archaea and bacteria diverge from each other. However, such levels of divergence might also point to radical genetic change and unknown duplications undergone by organisms of known lineages in adapting to different environments for protein familis as old as cellular life.

Altogether, our protocol demonstrates the capability of iterative network searches to mine metagenomes and unravel the underlying organisation of the genetic space, in particular to aggregate increasingly divergent variants around a set of primary sequences of interest. Although the analysis presented here focused on the implications of such methods for the ongoing exploration of microbial dark matter, owing to the high genetic diversity of the ocean

microbiomes, future methodological extensions to make more sense of the oceanic microbial diversity that rely less on primary sequence comparisons still appear warranted. As 3D structures tend to be more conserved than primary sequences during evolution, the development of 3D similarity networks connecting protein structures from cultured organisms to structures predicted from metagenomes, especially for protein families as old as cellular life, could greatly benefit microbial dark matter analyses, offering a new mining strategy to unveil novel potential major groups, and maybe even Domains of Life, from environmental microbiomes.

Materials & Methods

Constitution of a conserved protein families dataset

We constituted an initial dataset of 9,737,821 protein sequences annotated in published complete genomes from 4403 bacteria (including genomes from the CPR superphylum), 567 archaea (including genomes from the DPANN and Asgard superphyla) and 120 eukaryotes, as well as 18,020 viral and 1586 plasmidic genomes, that were acquired from public NCBI databases [56] (Table SI-1). The sequence similarity network of this protein sequence collection was reconstructed by an all-against-all DIAMOND blastp alignment [40] (using thresholds: E-value ≤ 10^{-5}, sequence identity ≥ 30%, mutual coverage ≥ 80%). This defined 891,459 protein clusters (connected components in the similarity network), including 1170 clusters containing ten or more sequences from each of the two prokaryotic Domains. The assortative mixing between Domains of life within each of these clusters, quantifying the degree of separation in the corresponding SSN between archaeal and bacterial sequences, was then computed using the Python package networkx [57] (version 2.8.8). We retained a final set of 53 families meeting thresholds of (i) Domain assortativity of 0.65 or above, and (ii) 150 or more sequences from each of the bacterial and archaeal Domains. These 53 protein families comprised a total of 125,774 sequences.

Iterative retrieval of environmental homologs

Sequences for 40,154,822 environmental genes from the Ocean Microbial Reference Gene Catalog (OM-RGC) dataset [39] were collected, alongside their corresponding sampling metadata and eggNOG [58] functional annotation, and translated into amino-acid sequences. An iterative search for environmental homologs in the OM-RGC dataset was conducted for the selected 53 protein families independently (building upon [38]). For each family, seed sequences were aligned against the environmental protein sequences with DIAMOND (using thresholds: E-value ≤ 10^{-5}, sequence identity ≥ 30%, mutual coverage ≥ 80%). The environmental sequences retrieved were then used as a base for a new round of DIAMOND alignment (with identical parameters) against the OM-RGC dataset. The procedure was then iterated, each round using as queries the environmental sequences newly retrieved during the previous round until one eventually found no additional sequence. At each step, the aligned

regions of the matched environmental sequences were projected onto those of previous rounds and checked to map back to a region covering at least 80% of a seed protein sequence, in order to maintain the plausibility of distant homology between indirectly linked sequences. A total of 826,717 sequences in the OM-RGC dataset were assigned to the selected protein families, with convergence being reached after seven rounds on average.

Comparison of retrieved environmental sequences to cultured diversity

The environmental sequences retrieved for each of the 53 selected seed families were compared to published sequences from taxonomically-resolved organisms in the NCBI *nr* database (downloaded in March 2020) via a DIAMOND alignment search (E-value ≤ 10^{-5}). Similarity values between environmental sequences and their closest published relative were calculated as the product of the amino-acid identity within the aligned region times the alignment coverage on the shortest sequence.

Sequence similarity network reconstruction and analysis

Sequence similarity networks were computed for each environmentally expanded protein family by conducting an all-against-all DIAMOND blastp alignment of seed and environmental sequences (E-value ≤ 10^{-5}, sequence identity ≥ 30%, mutual coverage ≥ 80%). The product of sequence identity times alignment coverage was used as edge lengths in the resulting networks. We then inferred, using Louvain clustering [41], node communities in those networks, i.e. groups of sequences that were tightly connected by homology links. This clustering step retained 691 communities in total across the 53 families in our dataset. Out of these 691 clusters, we further selected those containing at least 30 sequences, of which at most a tenth were initially present in the seed dataset, and with environmental sequences having on average 40% identity with their closest published counterpart or less. This selection yielded a subset of 80 sequence clusters, from 25 different families, that we judged to constitute the best candidates for harbouring major genetic novelty.

Similarity networks were rendered on Cytoscape [59]. However larger networks, usually comprising tens or hundreds of thousands of nodes and millions of edges, made the visualisation on Cytoscape intractable on standard computers. To represent and inspect visually the overall topology of these networks, we built “meta-networks” of SSNs instead. Rather than showing interconnections between all sequences, these meta-networks represented the connections between sequence clusters (as defined above). Specifically, each Louvain cluster inferred in the SSN of a protein family was represented by a single “meta-node” in the meta-network of this family. Two of these meta-nodes were then linked by a “meta-edge” if there existed at least one edge connecting the corresponding clusters in the SSN. To provide more insight into the strength of a connection between any two clusters, meta-edges were given a numeric weight representing the proportion of edges existing between sequences of the two connected clusters, relative to the total possible number of edges if the clusters had been fully connected together. These meta-networks were much lighter data structures, and could be displayed using Cytoscape.

Phylogenetic analysis of divergent clusters

Sequences from divergent clusters were then gathered in phylogenetic trees along with their seed sequences to determine their relative proximity. We started by dereplicating sequences from each of the 80 selected clusters, as well as seed sequences from each of the 25 families with at least one such cluster, using CD-HIT [60] with a 90% identity threshold. Up to 100 sequences per environmental cluster and up to 200 sequences per seed family were selected as representatives, and two types of phylogenies were subsequently inferred. First, we computed cluster-specific maximum likelihood phylogenies. A multiple sequence alignment was therefore produced, for each divergent cluster, between its own representative sequences and those of the corresponding seed family, using Mafft [61] with 1000 cycles of iterative refinement. These alignments were then trimmed using trimAl [62], and the result was used to produce a phylogeny using IQ-TREE [63–65] with 1000 bootstrap replicates. The second kind of phylogenetic trees we inferred were family-wide trees, grouping together (representatives of) seed sequences and all divergent clusters from any given family. To cope with tractability issues, we opted for an alignment-free approach [31, 32]. Kmer-based distance matrices were computed between all representative sequences of a family using jD2Stat [66], and then used to infer Neighbour-Joining phylogenetic trees with RapidNJ [67]. All trees were rendered and annotated in iTOL [68].

Inferrence and comparison of protein tertiary structures

3D structures were inferred for a selection of representative sequences in the SSNs of SMC proteins and DNA polymerase subunits.

For polymerases, one sequence was selected as representative for each Louvain cluster in the SSN. Clusters of highly divergent environmental variants were represented by the environmental sequence with the highest degree (number of edges in the SSN) to other environmental sequences within the cluster; other clusters were represented by the reference sequence with the highest degree to other references in the cluster. For SMC proteins, which have a significantly longer primary sequence (around 1200 amino-acids), we sought to reduce the number of structures to infer *de novo.* Six sequences from the divergent cluster of environmental SMC variants were chosen arbitrarily (all with maximal degree, since the cluster was fully connected), and public AlphaFold structures [69, 70] were acquired from UniProt [71] to represent reference SMC sequences (namely, and indicated by their UniProtKB accession: P9WGF2, Q5N0D2, A3PMS2, A9BZW2, P51834, Q69GZ5, Q8TZY2).

Structures were inferred for selected polymerase and environmental SMC sequences using ColabFold (v1.5.2, default parameters) [72]. Then, reference and environmental polymerase structures were compared together in an all-against-all manner using FoldSeek (easy-search mode, no pre-filter, alignment by TM-Align) [73]. Similarly, the inferred environmental SMC structures were compared with reference structures acquired from UniProt following the same protocol. For both protein families, these comparisons were used to construct dendrograms with RapidNJ [67], taking as distance metrics between two structures the average local distance difference test (lDDT) score of the corresponding bidirectional

structural alignment. These dendrograms thus represented the divergence in tertiary structures amongst the set of selected sequences. Dendrograms were plotted in iTOL [68] and annotated with 3D models of the protein structures rendered by PyMOL.

Data Availability Statement: The OM-RGC dataset analysed in this study is available from this webpage: http://ocean-microbiome.embl.de/companion.html. The dataset of conserved protein families constructed for this analysis is available from this Figshare project: https://doi.org/10.6084/m9.figshare.24893910.v1.

Competing Interests statement: The authors declare no competing financial interests.

Acknowledgements: RL, GB, PM and EB were supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement # 615274, category LS8).

References

1. Staley JT, Konopka A. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol* 1985; **39**: 321–346.
2. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 1995; **59**: 143–169.
3. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: The unseen majority. *Proc Natl Acad Sci* 1998; **95**: 6578–6583.
4. Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG, et al. Dissecting biological 'dark matter' with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci U S A* 2007; **104**: 11889–11894.
5. Alain K, Querellou J. Cultivating the uncultured: limits, advances and future challenges. *Extremophiles* 2009; **13**: 583–594.
6. Koch R. Untersuchungen uber bakterien V. Die aetiologie der milzbrand-krankheit, begrunder auf die entwicklungegeschichte Bacillus anthracis. *Beitrage Zur Biol Pflanz* 1877; **2**: 277–310.
7. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* 1998; **5**: R245–R249.
8. Castelle CJ, Banfield JF. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. *Cell* 2018; **172**: 1181–1197.
9. Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, et al. Accessing the Soil Metagenome for Studies of Microbial Diversity. *Appl Environ Microbiol* 2011; **77**: 1315–1324.
10. Ventosa A, de la Haba RR, Sánchez-Porro C, Papke RT. Microbial diversity of hypersaline environments: a metagenomic approach. *Curr Opin Microbiol* 2015; **25**: 80–87.
11. Behzad H, Gojobori T, Mineta K. Challenges and Opportunities of Airborne Metagenomics. *Genome Biol Evol* 2015; **7**: 1216–1226.
12. Sunagawa S, Acinas SG, Bork P, Bowler C, Eveillard D, Gorsky G, et al. Tara Oceans: towards global ocean ecosystems biology. *Nat Rev Microbiol* 2020; **18**: 428–445.
13. Hugenholtz P, Pitulle C, Hershberger KL, Pace NR. Novel division level bacterial diversity in a Yellowstone hot spring. *J Bacteriol* 1998; **180**: 366–376.

14. Chouari R, Le Paslier D, Dauga C, Daegelen P, Weissenbach J, Sghir A. Novel major bacterial candidate division within a municipal anaerobic sludge digester. *Appl Environ Microbiol* 2005; **71**: 2145–2153.
15. Pelletier E, Kreimeyer A, Bocs S, Rouy Z, Gyapay G, Chouari R, et al. "Candidatus Cloacamonas Acidaminovorans": Genome Sequence Reconstruction Provides a First Glimpse of a New Bacterial Division. *J Bacteriol* 2008; **190**: 2572–2579.
16. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. *Nat Microbiol* 2016; **1**: 1–6.
17. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 2013; **499**: 431–437.
18. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 2015; **523**: 208–211.
19. Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* 2002; **417**: 63–67.
20. Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD, et al. Enigmatic, ultrasmall, uncultivated Archaea. *Proc Natl Acad Sci* 2010; **107**: 8806–8811.
21. Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 2015; **521**: 173–179.
22. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 2017; **541**: 353–358.
23. Imachi H, Nobu MK, Nakahara N, Morono Y, Ogawara M, Takaki Y, et al. Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature* 2020; **577**: 519–525.
24. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. *Nature* 2016; **536**: 425–430.
25. Zhou Y, Zhou L, Yan S, Chen L, Krupovic M, Wang Y. Diverse viruses of marine archaea discovered using metagenomics. *Environ Microbiol* 2023; **25**: 367–382.
26. Gaïa M, Meng L, Pelletier E, Forterre P, Vanni C, Fernandez-Guerra A, et al. Mirusviruses link herpesviruses to giant viruses. *Nature* 2023; **616**: 783–789.
27. Al-Shayeb B, Schoelmerich MC, West-Roberts J, Valentin-Alvarado LE, Sachdeva R, Mullen S, et al. Borgs are giant genetic elements with potential to expand metabolic capacity. *Nature* 2022; **610**: 731–736.
28. Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems* 2018; **3**.
29. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. *Nat Biotechnol* 2020; **39**: 499–509.
30. Bernard G, Pathmanathan JS, 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* 2018; **10**: 707–715.
31. Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol* 2017; **18**: 1–17.
32. Ren J, Bai X, Lu YY, Tang K, Wang Y, Reinert G, et al. Alignment-Free Sequence Analysis and Applications. *Annu Rev Biomed Data Sci* 2018; **1**: 93–114.
33. Liu Z, Ma A, Mathé E, Merling M, Ma Q, Liu B. Network analyses in microbiome based on high-throughput multi-omics data. *Brief Bioinform* 2021; **22**: 1639–1655.
34. Zamkovaya T, Foster JS, de Crécy-Lagard V, Conesa A. A network approach to elucidate and prioritize microbial dark matter in microbial communities. *ISME J* 2021; **15**: 228–244.

35. Forster D, Bittner L, Karkar S, Dunthorn M, Romac S, Audic S, et al. Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *BMC Biol* 2015; **13**: 16.
36. Arroyo AS, Iannes R, Bapteste E, Ruiz-Trillo I. Gene Similarity Networks Unveil a Potential Novel Unicellular Group Closely Related to Animals from the Tara Oceans Expedition. *Genome Biol Evol* 2020; **12**: 1664–1678.
37. Lynch MDJ, Bartram AK, Neufeld JD. Targeted recovery of novel phylogenetic diversity from next-generation sequence data. *ISME J* 2012; **6**: 2067–2077.
38. Lopez P, Halary S, Bapteste E. Highly divergent ancient gene families in metagenomic samples are compatible with additional divisions of life. *Biol Direct* 2015; **10**: 64.
39. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. *Science* 2015; **348**: 1261359.
40. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 2021; **18**: 366–368.
41. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech Theory Exp* 2008; **2008**: P10008.
42. Vanni C, Schechter MS, Acinas SG, Barberán A, Buttigieg PL, Casamayor EO, et al. Unifying the known and unknown microbial coding sequence space. *eLife* 2022; **11**.
43. Iyer LM, Leipe DD, Koonin EV, Aravind L. Evolutionary history and higher order classification of AAA+ ATPases. *J Struct Biol* 2004; **146**: 11–31.
44. Frickey T, Lupas AN. Phylogenetic analysis of AAA proteins. *J Struct Biol* 2004; **146**: 2–10.
45. Lapointe F-J, Lopez P, Boucher Y, Koenig J, Bapteste E. Clanistics: a multi-level perspective for harvesting unrooted gene trees. *Trends Microbiol* 2010; **18**: 341–347.
46. Cobbe N, Heck MMS. The Evolution of SMC Proteins: Phylogenetic Analysis and Structural Implications. *Mol Biol Evol* 2004; **21**: 332–347.
47. Rybenkov VV, Herrera V, Petrushenko ZM, Zhao H. MukBEF, a Chromosomal Organizer. *Microb Physiol* 2014; **24**: 371–383.
48. Böhm K, Giacomelli G, Schmidt A, Imhof A, Koszul R, Marbouty M, et al. Chromosome organization by a conserved condensin-ParB system in the actinobacterium Corynebacterium glutamicum. *Nat Commun* 2020; **11**: 1485.
49. Hirano T. The ABCs of SMC proteins: two-armed ATPases for chromosome condensation, cohesion, and repair. *Genes Dev* 2002; **16**: 399–414.
50. Gruber S, Arumugam P, Katou Y, Kuglitsch D, Helmhart W, Shirahige K, et al. Evidence that Loading of Cohesin Onto Chromosomes Involves Opening of Its SMC Hinge. *Cell* 2006; **127**: 523–537.
51. Lin Z, Kong H, Nei M, Ma H. Origins and evolution of the recA/RAD51 gene family: Evidence for ancient gene duplication and endosymbiotic gene transfer. *Proc Natl Acad Sci* 2006; **103**: 10328–10333.
52. Castelle CJ, Brown CT, Anantharaman K, Probst AJ, Huang RH, Banfield JF. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat Rev Microbiol* 2018; **16**: 629–645.
53. Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, et al. Giant virus diversity and host interactions through global metagenomics. *Nature* 2020; **578**: 432–436.
54. Banerjee A, Neiner T, Tripp P, Albers S-V. Insights into subunit interactions in the Sulfolobus acidocaldarius archaellum cytoplasmic complex. *FEBS J* 2013; **280**: 6141–6149.
55. Chaudhury P, Does C van der, Albers S-V. Characterization of the ATPase FlaI of the motor complex of the Pyrococcus furiosus archaellum and its interactions between the ATP-binding protein FlaH. *PeerJ* 2018; **6**: e4984.

56. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2022; **50**: D20–D26.
57. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function using NetworkX. In: Varoquaux G, Vaught T, Millman J (eds). *Proc. 7th Python Sci. Conf.* 2008. Pasadena, CA USA, pp 11–15.
58. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019; **47**: D309–D314.
59. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003; **13**: 2498–2504.
60. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinforma Oxf Engl* 2006; **22**: 1658–1659.
61. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 2013; **30**: 772–780.
62. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009; **25**: 1972–1973.
63. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 2020; **37**: 1530–1534.
64. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* 2018; **35**: 518–522.
65. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017; **14**: 587–589.
66. Chan CX, Bernard G, Poirion O, Hogan JM, Ragan MA. Inferring phylogenies of evolving sequences without multiple sequence alignment. *Sci Rep* 2014; **4**: 6504.
67. Simonsen M, Mailund T, Pedersen CNS. Rapid Neighbour-Joining. In: Crandall KA, Lagergren J (eds). *Algorithms Bioinforma.* 2008. Springer, Berlin, Heidelberg, pp 113–122.
68. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* 2021; **49**: W293–W296.
69. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021; **596**: 583–589.
70. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022; **50**: D439–D444.
71. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* 2023; **51**: D523–D531.
72. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods* 2022; **19**: 679–682.
73. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* 2023; 1–4.

Supplementary Information and Figures

Supplementary Text SI-1: Exploring the protein universe using iterative network searches unveiled distant homologs, including distant paralogs

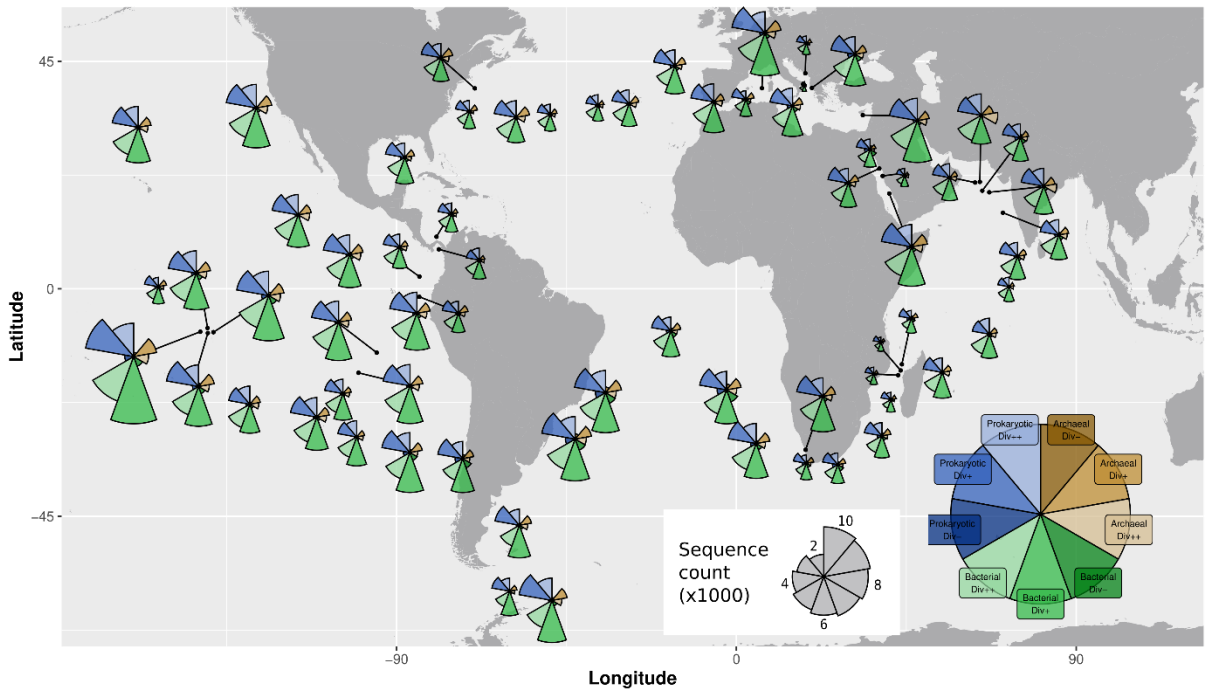
Sequences matched in earlier rounds of metagenome mining generally tended to have higher similarity to their closest *nr* relatives (one-sided Spearman rank test, correlation coefficient -0.2, p-value $<1.6x10^{-22}$) (Figure SI-2), highlighting the capacity of iterative alignment search protocols to identify distant homology relationships beyond the scope of direct BLAST searches. These earlier sequences also lay closer to seed sequences in sequence similarity networks, using alignment similarity as a network distance (see Methods). In accordance with this, the network distance from environmental sequences to their seed counterparts was generally correlated with their divergence from their closest well-characterised published relative (one-sided Spearman test, correlation 0.22, p-value $<1.6x10^{-22}$). This observation stood both for the globality of environmental sequences and for a majority (47/53) of protein families when evaluated individually. Families that did not support this correlation exhibited a common pattern of environmental homolog retrieval, wherein an initial decline in retrieved sequences after each search step was followed by a substantial retrieval uptick, as opposed to an expected monotonous decline. Sequences responsible for this later retrieval uptick were generally more similar to their published counterparts than previously retrieved homologs, but connected only distantly to the seed sequences they were retrieved from.

Moreover, sequences from these 6 families were predominantly assigned to different OGs than previously retrieved homologs, suggesting that they may belong to other protein families than that initially used to query the environmental dataset, and showing that extant protein families do not always perfectly segregate in sequence homology within the protein space. Although these indirect connections may arise between bona fide unrelated families, spuriously showing some convergence in primary sequence, some may also be indicative of a shared evolutionary history.

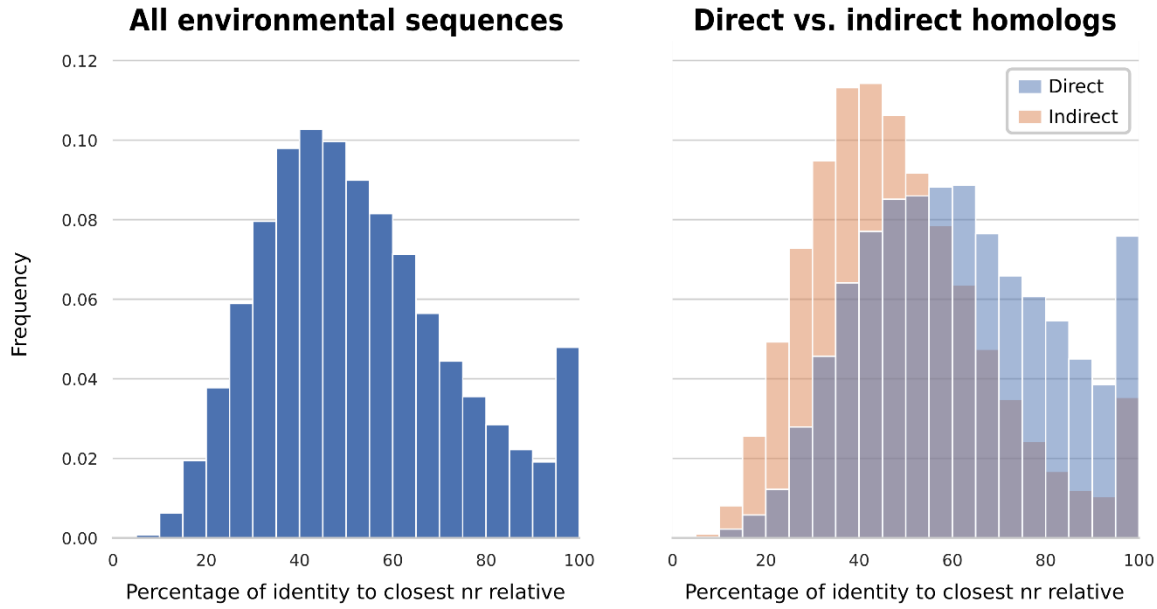
One such case is for the SecD and SecF translocation proteins families. Both families are involved in a common complex facilitating protein secretion across the cytoplasmic membrane of archaea and bacteria [1]. While these two families had been separately retained by our clustering analysis of reference sequences as highly conserved and broadly distributed in accordance with the literature [2], our detection of distant homologs recovered their paralogy relationship, unveiling 11,864 sequences retrieved in common for both families out of 15,027 and 13,048 homologs in OM-RGC (in SecD and SecF respectively, after respectively 14 and 13 search iterations). Upon further analysis, these sequences clustered in two main categories. First-type sequences were closer to seed sequences from SecD than from SecF in the similarity networks, more central in the SecD network than in SecF, and retrieved by SecD sequences in fewer rounds of search; conversely, second-type sequences were closer to SecF by those same criteria. In line with these observations, eggNOG functional prediction [3] predominantly

assigned first- and second-type sequences to OGs corresponding respectively to SecD (COG0342) and SecF (COG0341). Overall, nearly 80% of these sequences could be unambiguously mapped back to SecD or SecF by network distance, centrality, search iterations and eggNOG functional annotation. The intersecting retrieval of environmental homologs in the case of SecD and SecF proteins is thus most likely the result of their ancestral paralogous history [4], which was not reflected in the initial dataset but became apparent with the addition of environmental sequences, and assesses the efficiency of our iterative strategy to identify deep homology.

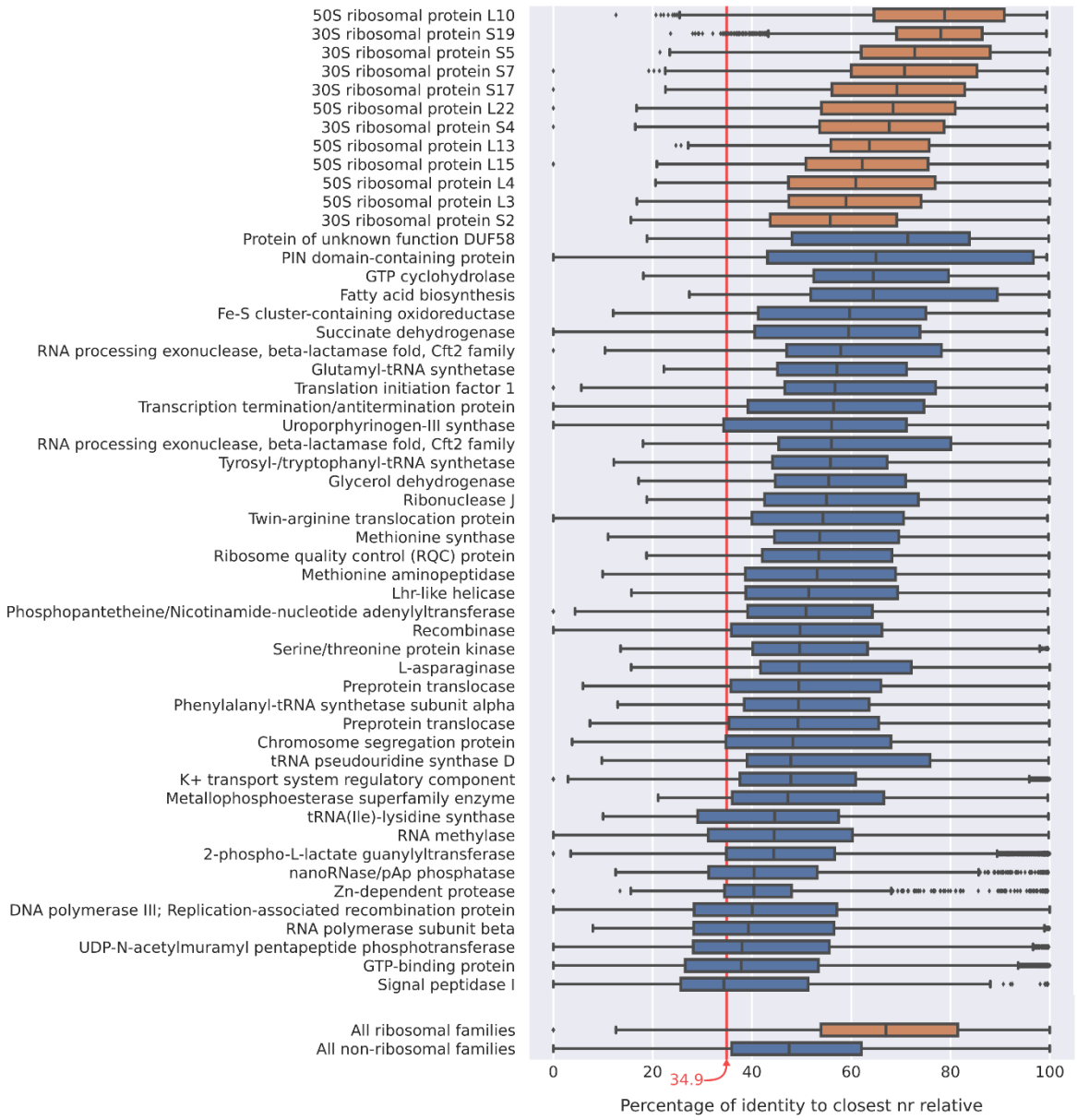
1. Pogliano JA, Beckwith J. SecD and SecF facilitate protein export in Escherichia coli. *EMBO J* 1994; **13**: 554–561.
2. Hand NJ, Klein R, Laskewitz A, Pohlschröder M. Archaeal and bacterial SecD and SecF homologs exhibit striking structural and functional conservation. *J Bacteriol* 2006; **188**: 1251–1259.
3. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019; **47**: D309–D314.
4. Tseng T-T, Gratwick KS, Kollman J, Park D, Nies DH, Goffeau A, et al. The RND Permease Superfamily: An Ancient, Ubiquitous and Diverse Family that Includes Human Disease and Development Proteins. *J Mol Microbiol Biotechnol* 1999; **1**: 107–125.



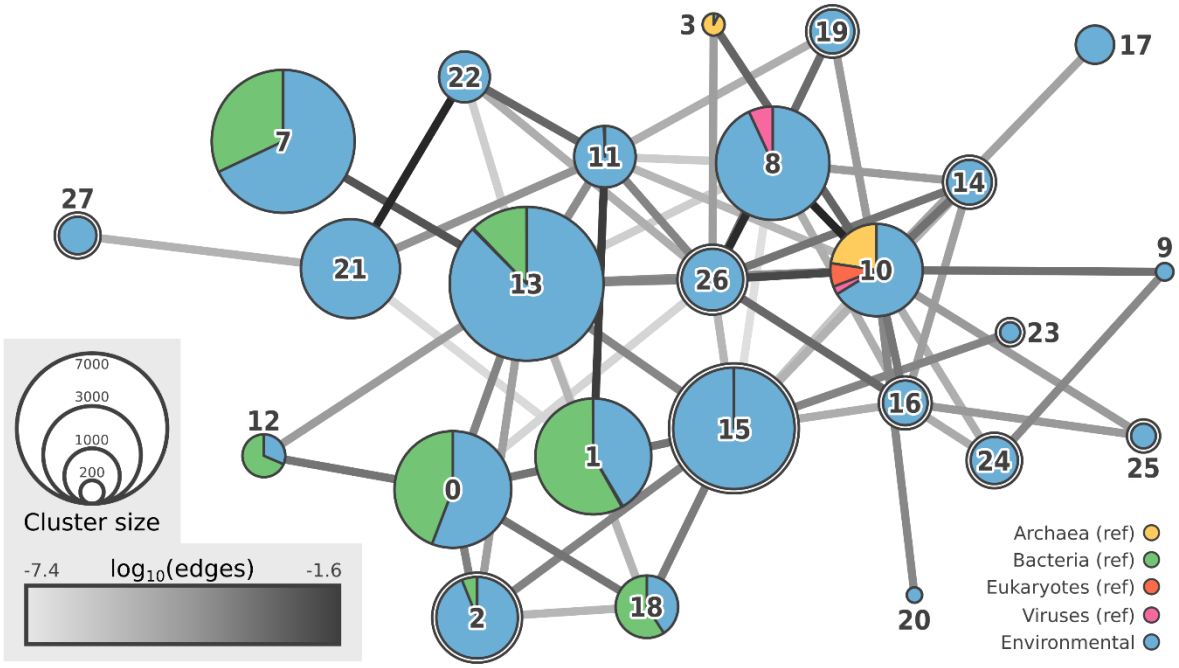
Supplementary Figure SI-1: Geographical distribution of retrieved environmental homologs. Environmental homologs were categorised as bacterial (green), archaeal (brown), or prokaryotic (blue), from highly divergent (Div++, <40% similarity with the best hit from cultured organism in the NCBI nr database) to weakly divergent (Div-, >90% similarity with the best hit from cultured organism in the NCBI nr database). The sizes of sector charts indicate the numbers of sequences falling within each category at a given sampling site.



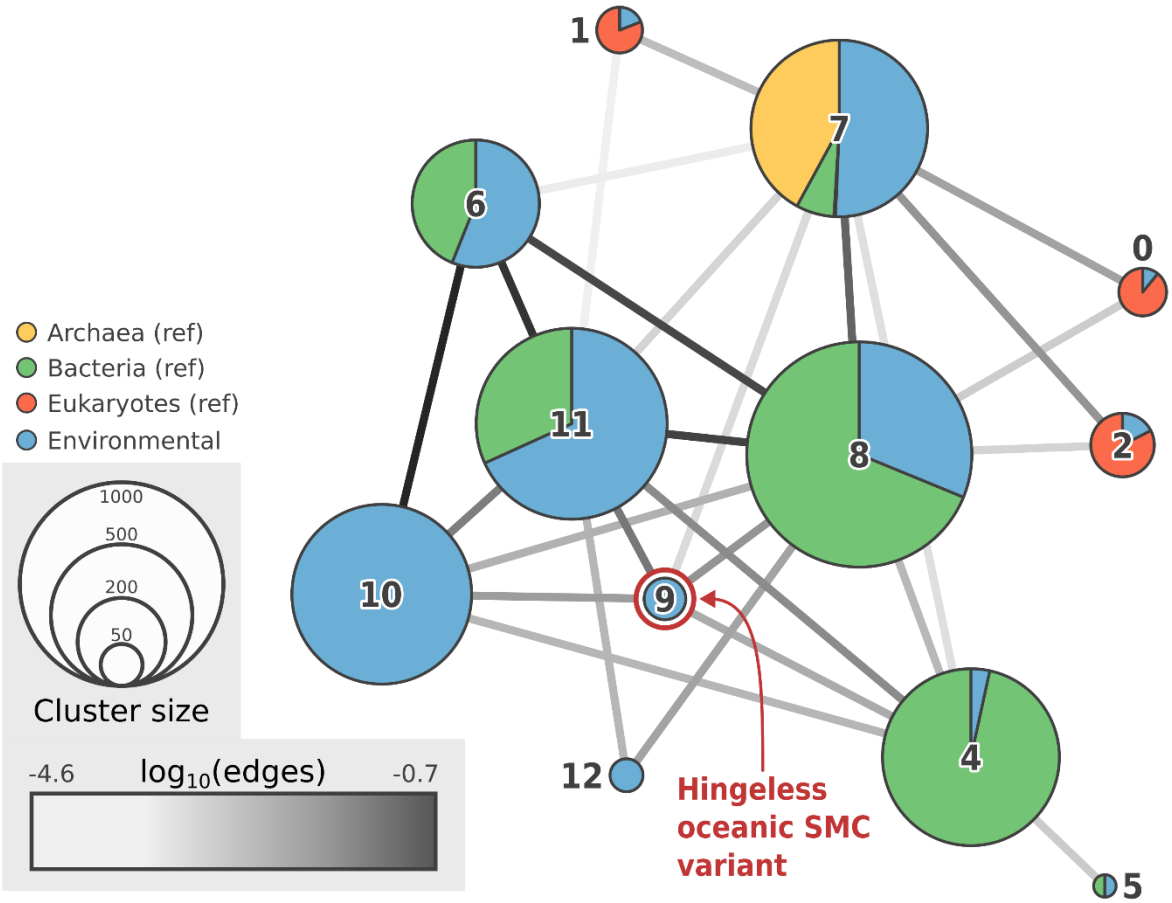
Supplementary Figure SI-2: Distribution of sequence similarity to closest published homolog for environmental sequences. Similarity is computed as amino-acid identity within the aligned region times coverage of the alignment on the shortest sequence, ranging from 0 to 100, and is indicated in abscissa. The proportion of environmental sequences with a given identity to their closest published relative is indicated in ordinate. **A**: all environmental sequences. **B**: sequences retrieved in the first round of search (blue) and in subsequent rounds (orange).



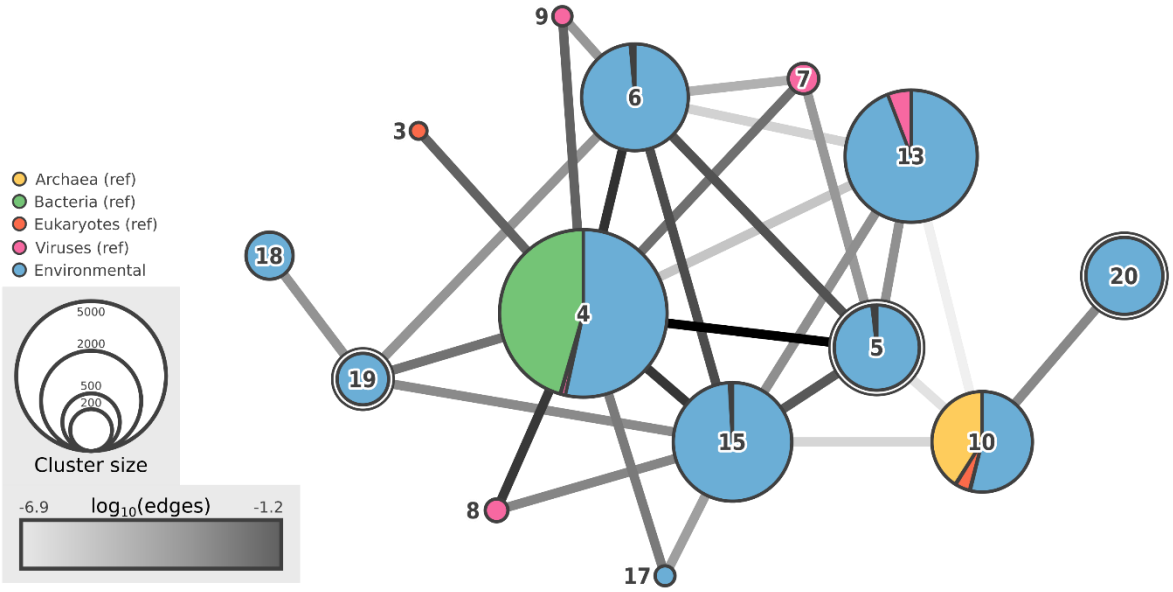
Supplementary Figure SI-3: Distribution of the divergence of environmental homologs from the known diversity. Boxplots show, for each of the 53 protein families, the distribution of sequence similarity (computed as the product of the amino-acid identity within the aligned region times the coverage of the alignment on the shortest sequence; ranging from 0% to 100%) between environmental homologs and their closest relative in the known cultured diversity. Ribosomal families are represented by brown boxes, and blue boxes indicate non-ribosomal families. The two boxes on the bottom show the distribution of sequence similarity when all ribosomal families are pooled together against all non-ribosomal families. The red vertical line indicates the 34.9% identity threshold, corresponding to the average divergence observed between archaeal and bacterial sequences within seed protein families.



Supplementary Figure SI-4: Meta-network of the DNA Polymerase III subunits sequences and their environmental homologs. Nodes correspond to Louvain clusters within the sequence similarity network of DNA Pol III subunit proteins. Nodes are arbitrarily designated with numeric labels. Node areas indicate the number of represented sequences. The interior of each node consists of a pie chart indicating the composition of the cluster in archaeal, bacterial, eukaryotic, viral and environmental sequences. Nodes representing clusters of highly divergent environmental variants are highlighted with a double outline. Edges are coloured according to the proportion of edges between two clusters in the family SSN, relative to the maximum number of possible edges between these clusters (darker edges indicate stronger connections between clusters).



Supplementary Figure SI-5: Meta-network of the SMC sequences and their environmental homologs. Nodes correspond to Louvain clusters within the sequence similarity network of SMC proteins. Nodes are arbitrarily designated with numeric labels. Node areas indicate the number of represented sequences. The interior of each node consists of a pie chart indicating the composition of the cluster in archaeal, bacterial, eukaryotic and environmental sequences. Node 9, representing the only cluster of highly divergent environmental variants in this family, is highlighted in red. Edges are coloured according to the proportion of edges between two clusters in the family SSN, relative to the maximum number of possible edges between these clusters (darker edges indicate stronger connections between clusters).



Supplementary Figure SI-6: Meta-network of the RecA/RadA sequences and their environmental homologs. Nodes correspond to Louvain clusters within the sequence similarity network of RecA/RadA proteins. Nodes are arbitrarily designated with numeric labels. Node areas indicate the number of represented sequences. The interior of each node consists of a pie chart indicating the composition of the cluster in archaeal, bacterial, eukaryotic, viral and environmental sequences. Nodes representing clusters of highly divergent environmental variants are highlighted with a double outline. Edges are coloured according to the proportion of edges between two clusters in the family SSN, relative to the maximum number of possible edges between these clusters (darker edges indicate stronger connections between clusters).

Supplementary Table SI-1: Public genomic data used to constitute the initial set of protein families. NCBI Assembly or Nucleotide IDs are indicated along with taxonomic information on the host lineage. This dataset was assembled from two prior separate sets of genome records, one with sequences linked to a Nucleotide identifier (sheet *Dataset_part1* in the Table) and one linking sequences to Assembly objects (sheet *Dataset_part2*), hence the two sheets in this Table.

Supplementary Table SI-2: Main functional annotations for protein families used in the environmental homolog search.

Supplementary Table SI-3: Sequence identifiers, iteration step of retrieval, and SSN cluster label for seed sequences and environmental homologs of DNA Polymerase III subunits, SMC proteins, and recombinases. Seed sequences are identified in accordance with the seed dataset shared in Data Availability Statement and are designated as retrieved at step 0. Environmental sequences (retrieved at steps 1 and above) are specified by their OM-RGC identifiers.