

Water mass specific genes dominate the Southern Ocean microbiome

Received: 11 November 2025

Accepted: 28 January 2026

Published online: 09 March 2026



Emile Faure^{1,2} , Jolann Pommellec¹, Cyril Noel³ , Alexandre Cormier³, Lisa-Marie Delpech^{1,4,5} , A. Murat Eren^{6,7,8,9,10}, Antonio Fernandez-Guerra^{11,12} , Chiara Vanni¹³ , Marion Fourquez¹⁴ , Marie-Noëlle Houssais¹⁵, Ulysse Guyet¹⁶ , Corinne Da Silva¹⁶ , Frederick Gavory¹⁶, Aude Perdereau¹⁷, Karine Labadie¹⁷, Patrick Wincker^{16,18} , Julie Poulain^{16,18}, Christel Hassler^{19,20,21} , Yajuan Lin^{22,23,24} , Nicolas Cassar^{23,24}  & Loïs Maignien^{1,10} 

The Southern Ocean (SO) plays a key role in regulating global biogeochemical cycles and climate, yet microbial genes sustaining its biological activity remain poorly characterized. We introduce a microbial genes collection from 218 metagenomes sampled during the Antarctic Circumnavigation Expedition, the majority of which are missing from functional databases. 38% even lack homologs in current reference marine gene catalogs, defining a singular genetic seascape. We show that SO gene assemblages exhibit a common polar signature with the Arctic Ocean while being structured by water masses at the SO-scale. We analyze genomic markers of diverse SO biomes, focusing on dimethylsulphoniopropionate (DMSP) cleavage by polar-adapted bacteria, organic matter consumption in the blooming Mertz polynya and adaptation to polar conditions in the ubiquitous bacteria *Pelagibacter*. Our work takes a step towards a comprehensive understanding of SO's plankton ecology and evolution, capturing the current state of the unique microbial diversity in this rapidly changing Ocean.

The Southern Ocean (SO) dominates other oceans in heat and carbon uptakes while being particularly exposed to climate change impacts¹. It is mainly composed of high nutrient low chlorophyll waters (HNLC) where phytoplankton growth is limited by trace elements such as iron or manganese^{2,3}. In presence of these elements, as in coastal polynyas, phytoplankton blooms can reach concentrations of 10⁸ cells per liter⁴, playing a significant role in global carbon sequestration through the biological pump⁵. Beyond phytoplankton, the extent of carbon export is impacted by the consumption and remineralization of organic matter by communities of bacteria and archaea^{6,7}. The SO bacterial communities are also key actors of the global sulfur cycle through the cleavage of dimethylsulphoniopropionate (DMSP) into dimethyl sulfide (DMS), the dominant natural source of sulfur in the

atmosphere^{8,9}. Yet, in situ abundance and diversity of microbial communities in the SO remain poorly described.

Recent large-scale environmental metagenomics projects highlighted the rich functional and taxonomic diversity of marine plankton and the structuring effect of environmental conditions on planktonic communities^{10–13}. However, only two sampling locations in the SO were included in recent efforts to compute global genes and genomes catalogs^{14,15}, underscoring the substantial undersampling of this critical ocean. A study focusing on polar oceans and including 21 metagenomics samples from the SO allowed the construction of a polar gene catalog in 2020, showing the high prevalence of polar-specific genes in the SO¹⁶. Other recent studies investigated microbial metabolism around the Kerguelen Islands and coastal western Antarctic Peninsula

A full list of affiliations appears at the end of the paper. ✉ e-mail: emile.faure@sb-roscoff.fr; Nicolas.Cassar@duke.edu; lois.maignien@univ-brest.fr

through metagenomics and metatranscriptomics^{17–19}. Yet, we still lack a realistic census of SO's microbial diversity and of the environmental factors structuring its planktonic communities. We address this important knowledge gap, identifying drivers of planktonic functional and taxonomic diversity in this area subject to major environmental changes¹.

The Antarctic Circumnavigation Expedition (ACE) circumnavigated the Southern Ocean during the 2016–2017 austral summer, producing an unprecedented amount of physical, chemical and biological observations²⁰. Analyzing 218 metagenomes, we increase by an order of magnitude the number of SO samples ever considered in a metagenomics study to present an SO-specific gene catalog (Fig. 1). Building on the seminal work of previous global metagenomics efforts^{21–23}, we first demonstrate the broad-scale uniqueness of the SO compared to other oceans, before exploring its regional variability. We then chose to illustrate this original genetic seascape using three cases with particular relevance for this polar ocean: we identify functional gene classes with strong biogeographical structure at SO scale, such as lyases cleaving DMSP into DMS and acrylate, we then characterized the genomic signature of specialist species occurring in the Mertz polynya, and finally use *Pelagibacter* to illustrate genomic adaptations to polar conditions in a ubiquitous taxon.

Results

Considerable genetic originality of the Southern Ocean

Individual assemblies of the 218 metagenomes (Figs. 1 and S1) produced 68,074,004 contigs in which we identified 175,336,776 open-reading frames (ORFs). We dereplicated these ORFs using 95%

similarity and 90% coverage thresholds, producing 89,739,060 dereplicated-genes hereafter called unigenes. 51.3% of ACE unigenes did not cluster with any unigene from the most recent Tara Ocean and Polar Circle gene catalog²¹ (OM-RGC-v2) at thresholds of 30% similarity and 80% coverage in amino acid sequence (Fig. 2A). When accounting only for ACE samples from 0.2 to 3 μm size fraction taken in the sunlit layer, therefore matching the Tara projects sampling strategy, this number remained at 37.9% (Fig. 2A). Conversely, 28.9% of OM-RGC v2 unigenes did not cluster with any ACE unigene using the same thresholds. This strong mutual exclusion between the two catalogs highlights the originality of the SO as compared to other oceans, especially considering that the OM-RGC-v2 includes unigenes assembled from Arctic metagenomes. Furthermore, we used the same approach to cluster ACE unigenes with consensus sequences from the Novel Metagenome Protein Families database (NMPFamsDB)²⁴, i.e., representative sequences from 106,198 protein families absent from both Pfam and reference genomes. Only 14,436 (13.6%) representative sequences from the NMPFamsDB clustered with 25,927 (0.03%) ACE unigenes.

We further explored distant gene homology using AGNOSTOS²⁵, a clustering and annotation tool based on a database built from 1823 metagenomes (from marine and human microbiome environments) and 28,941 bacterial and archaeal genomes (from the Genome Taxonomy Database²⁶). We clustered the 175,336,776 ORFs into 30,123,228 AGNOSTOS gene clusters (AGC), of which 64.8% were singletons, 32.6% were good-quality clusters of multiple ORFs as per AGNOSTOS standards, and 2.5% were discarded as low-quality clusters. 52.6% of the ORFs were tagged as unknowns (i.e., without functional annotation)

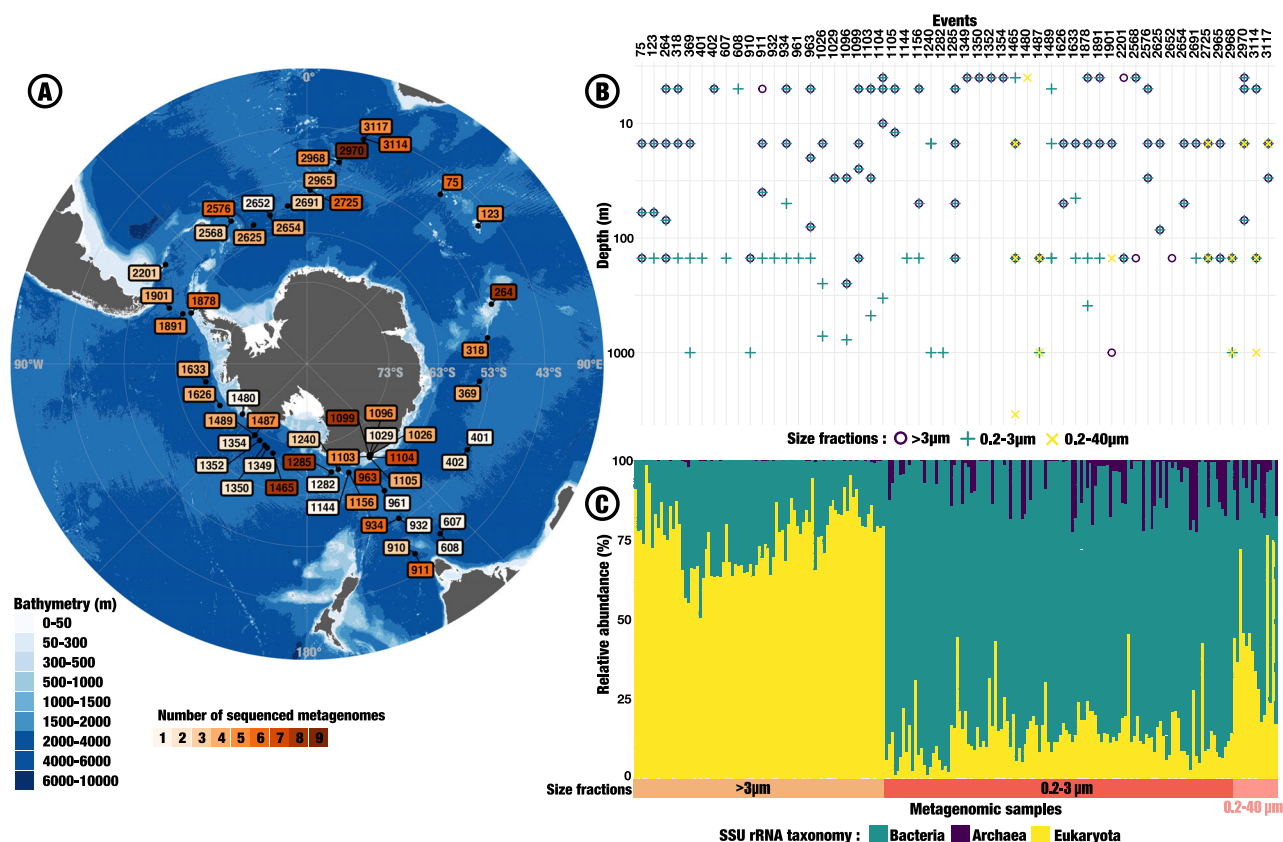


Fig. 1 | Overview of the ACE metagenomics dataset. A Map of CTD downcast events on which metagenomics samples were taken, colored according to the number of samples taken on the cast. Background colors indicate land (gray) and bathymetry (blue), data from NOAA⁹¹ obtained through the ggOceanMaps R package⁹². **B** Depth and size fraction chart of all ACE metagenomes. Each vertical line corresponds to an event as pictured on the map in (A), and each metagenome is

represented as a dot at its corresponding depth, with the shape indicating the size fraction. **C** Relative abundance of each domain of life in every sampled metagenome, as estimated through SSU rRNA assembly and annotation from metagenomics short reads. Samples were ordered according to their size fraction on the x-axis, as indicated on the bottom layer. Source data are provided as a Source data file.

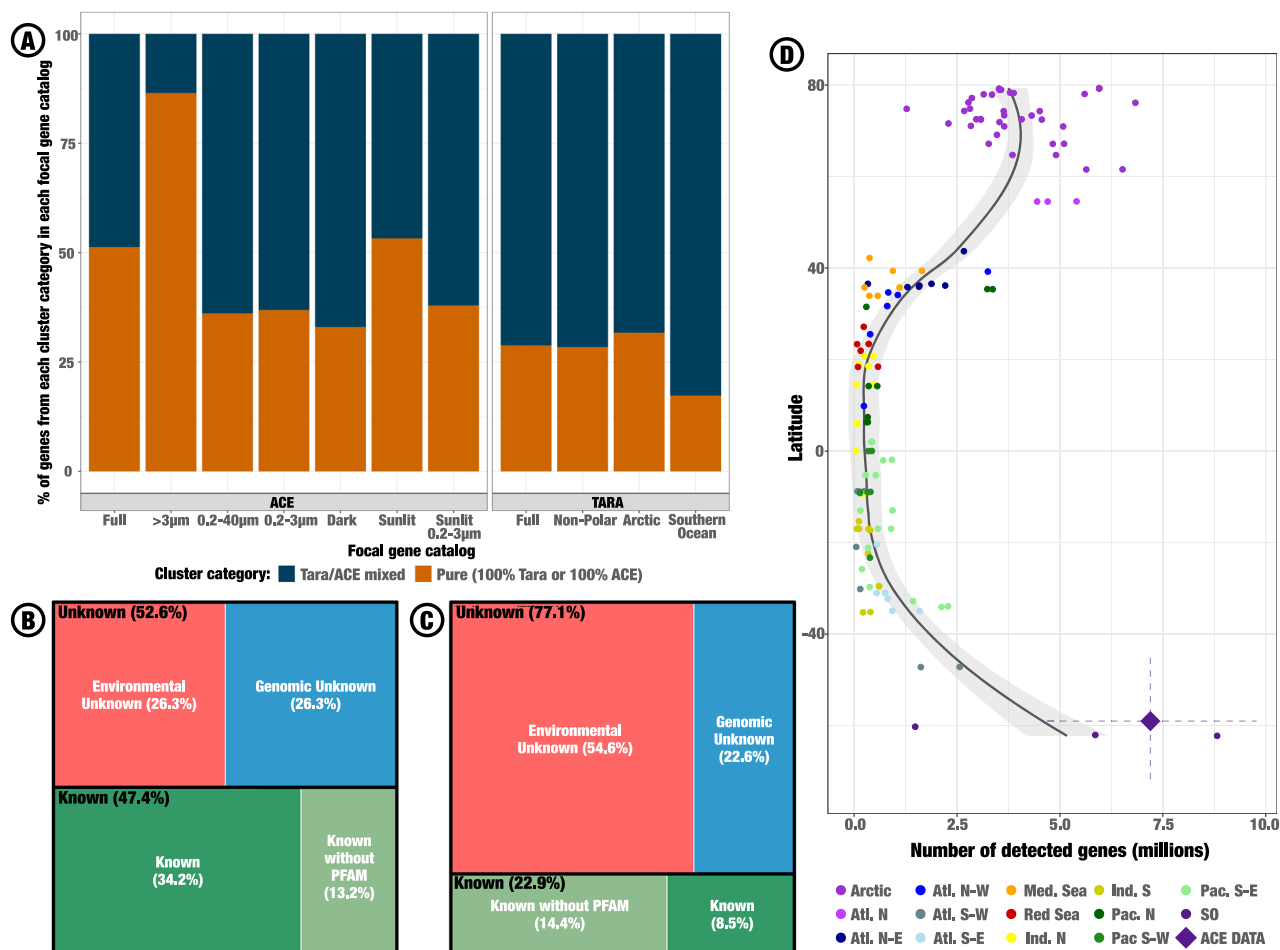


Fig. 2 | Originality of Southern Ocean microbial genes. **A** Distribution of genes from either the ACE unigenes catalog (left box) or the OM-RGC v2 (right box) into pure or mixed gene clusters. Genes from both catalogs were clustered at 30% similarity and 80% coverage thresholds in amino acid sequences, then classified as either belonging to a pure cluster, i.e., a cluster only containing genes from one catalog, or a mixed one, i.e., a cluster mixing genes from both catalogs. Results are either presented on the full catalogs, or restrained to specific gene subsets involving size fractions, depth, and geography, as described on the x axis. **B** Chart of AGNOSTOS annotations at gene level, i.e., accounting for the number of genes in each cluster annotation category: Environmental Unknowns (EU) lack functional annotations and are absent from any genome recorded in the AGNOSTOS database, Genomic Unknowns (GU) also lack functional annotations yet are recorded in

a genomic context in the AGNOSTOS database, Knowns are functionally annotated, either with (K) or without PFAM (KWP) annotations. **C** Chart of AGNOSTOS annotations at AGNOSTOS gene cluster level. **D** Latitudinal gradient of ACE genes' detection in Tara Oceans and Polar Circle samples. Genes were considered as detected if at least 60% of their sequence was covered with a depth of 1× or more. The mean number of detected genes in ACE samples is indicated by the diamond shaped point, with the horizontal dashed line spanning from first to third quartile of detected gene number and the vertical one from minimum to maximum latitudes. A loess curve (in black) with 95% confidence interval (in gray) was fitted to the number of detected genes in Tara samples, not taking into account ACE samples. Source data are provided as a Source data file.

and contributed to 77.1% of all AGC, illustrating the high prevalence of singletons among unknown ORFs compared to known ones, which clustered better together (Fig. 2B, C). The asymptotic nature of collector curves drawn at AGC level suggests that the ACE AGC catalog covers most of SO genomic diversity (Fig. S3).

Bipolar distribution of Southern Ocean microbial genes

Adaptation to polar conditions is thought to be responsible for a high genomic similarity between Arctic and Antarctic microbiomes despite dispersal isolation¹⁶. To quantify this bipolar pattern among ACE genes, we mapped 134 Tara Oceans and Tara Polar Circle metagenomes covering the surface and subsurface of most subtropical and arctic oceanic regions onto ACE contigs. The resulting ORF detection matrix showed a bipolar distribution of SO-genes at global scale (Fig. 2D). The mean number of detected ORF per ACE sample was of 7,198,029, against 3,956,716 in Tara Polar Circle samples, illustrating the high level of endemism of the SO despite similarities in gene

content between poles. This mean dropped to 334,896 ORFs for non-polar *Tara Oceans* samples. Of the 34,344,531 ACE ORFs detected in at least one sample from the Arctic Ocean, 26,353,298 were absent from all non-polar oceans sampled during *Tara Oceans* and therefore identified as polar-specific. Polar specific ORFs were distributed in 14,426,012 unigenes and 4,105,973 AGC clusters, of which 61.8% were unknown (39% environmental unknown, 22.8% genomic unknown). We identified 4314 EggNOG functions significantly enriched in polar specific unigenes compared to the rest of ACE unigenes (over a total of 54,772 functions, unilateral Fisher tests, adjusted *p*-value < 0.01). The six functions with the highest odds ratio, ranging between 4.0 and 4.3 in favor of polar-specific unigenes, were *Formate dehydrogenase (NAD⁺) activity*, *Excinuclease ABC (UV-specific endonuclease)*, *Septum formation initiator*, *cold-shock protein*, *oxidoreductase activity acting on the aldehyde or oxo group of donors*, *iron-sulfur protein as acceptor* and *Iron-binding zinc finger CDGSH type* (See Table S1 to access complete list of enriched functions).

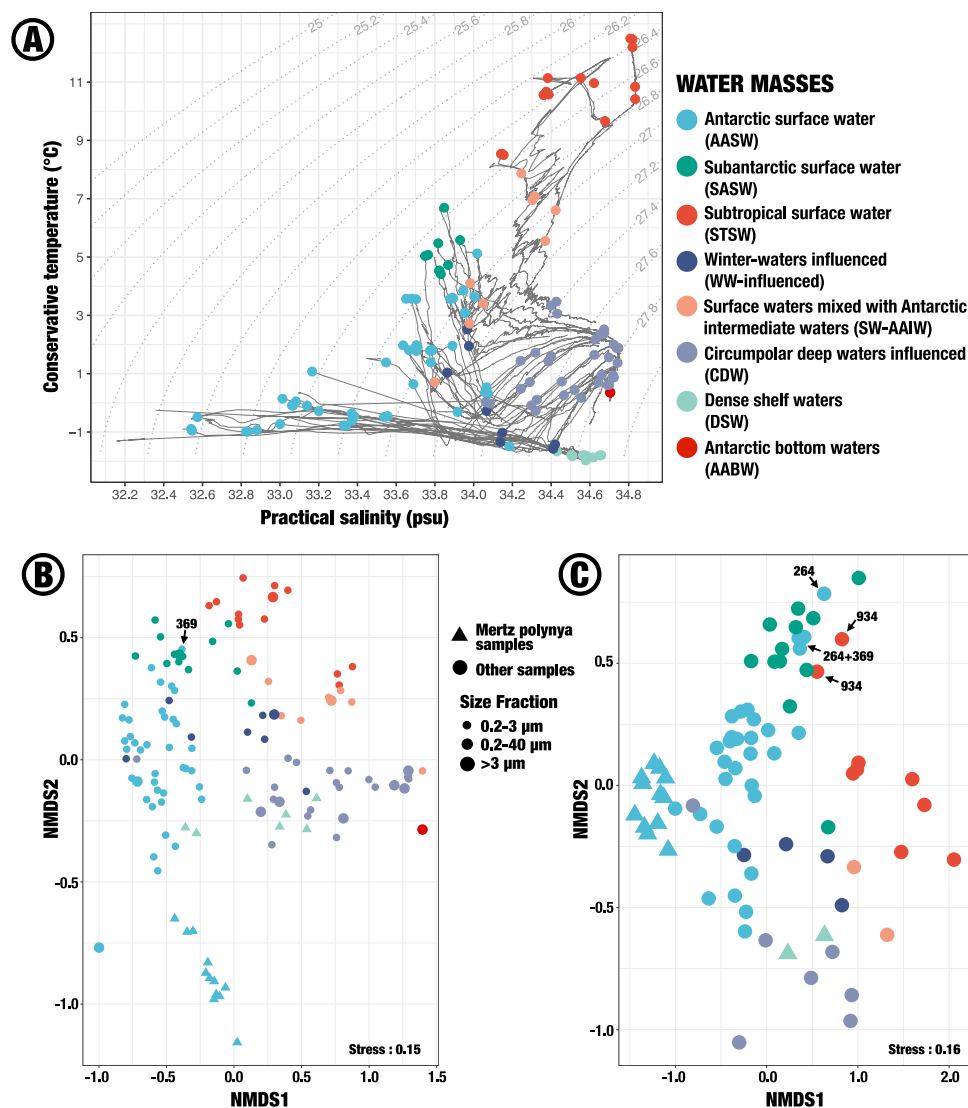


Fig. 3 | Microbial genes assemblages of the Southern Ocean are water mass specific. **A** Temperature–salinity diagram based on ACE downcast CTD data⁹³. Each gray line corresponds to a CTD cast. Dots correspond to depths at which seawater was sampled for metagenomic libraries construction, colored according to their attributed water masses. A description of each water mass is available in Table S2. Dotted lines in the background correspond to isopycnals. Non-Metric Multi-dimensional Scaling (NMDS) computed on AGC abundance matrices of small (**B**) and large (**C**) size fractions, colored according to their water masses using the same

color legend as in (**A**). Please note that the sample ACE1_264_34A_150 m was excluded from the large size fraction NMDS due to its outlying position. Positions of samples in (**B**, **C**) are only determined by their composition in AGC. The event numbers of samples taken above 150 m are indicated by black arrows when their positions do not match their water mass classification, as discussed in the results. Events 369 and 264 were taken right on the Polar Front, while event 934 was taken on the Sub-Antarctic Front.

Microbial gene assemblages are distinct across water masses at SO scale

We analyzed AGC's biogeography following three steps, (1) an unconstrained analysis of AGC's distribution across samples, (2) a univariate exploration of each AGC to detect those linked to the environment (env-AGC) and (3) a grouping of env-AGCs into co-abundant groups to allow a multivariate exploration of their response to environmental gradients. We worked independently on the small (0.2–40 + 0.2–3 μm) and large (>3 μm) size fractions considering their different taxonomic and metagenomic profiles (Figs. 1C and S2). We focused exclusively on AGCs with non-null coverage values in at least 20% of samples (1,906,624 and 2,437,988 clusters in the small and large size fractions, resp.), not taking into account rare AGCs.

The classification of our samples in water masses based on temperature-salinity-oxygen diagrams was the best grouping variable for predicting AGC abundance (Figs. 3 and S4). Water masses reflect

differences in key variables (temperature, salinity, oxygen) but also oceanic circulation and hydrographic fronts (Table S2), making it an efficient framework to investigate the interplay between microbial ecology and biogeochemistry^{27,28}. AGC assemblages were significantly distinct across water masses in both size fractions (Fig. 3B, PERMANOVA with 999 permutations, p -value < 0.001). Sub-Antarctic Surface Waters samples were separated from both Antarctic Surface Waters samples and Sub-Tropical Surface Waters samples, suggesting biogeographical barriers at both the Sub-Antarctic Front and the Polar Front. This was already observed for *Flavobacteria*²⁹ but lacked confirmation on broader taxonomic range³⁰. All surface samples showing potential mismatches between their AGC assemblage and their attributed water masses came from events located on oceanographic fronts (Fig. 3B, C), suggesting a potential mixing of microbial assemblages at these fronts.

Samples from Circumpolar Deep Waters were well separated from surface water masses in both size fractions, while surface waters influenced by colder (Winter Water-influenced and Antarctic Intermediate-influenced) layers appeared between Circumpolar Deep Waters and Surface Waters. Samples from Dense Shelf Waters were mostly similar to Circumpolar Deep Waters samples, except for the two shallowest Dense Shelf Waters samples which appeared closer to Antarctic Surface Waters in genomic composition for the small size fraction (Fig. 3B). Our results further suggest a diminution of AGC diversity in deep water masses, converging towards a “deep water” end member community (see Supplementary Results). Still, the only Antarctic Bottom Water sample (3460 m depth) had the most extreme coordinate on the NMDS first axis (Fig. 3B), suggesting a unique genomic composition. In light of this uniqueness and considering the projected decrease in Antarctic Bottom Water formation due to increasing influence of meltwater³¹ from Antarctica, a better characterization of the functional roles of microbial populations in Antarctic Bottom Waters is urgently needed.

Identifying gene classes with strong biogeographical structure at SO-scale

We built random forest regression models for each AGC, predicting coverage using 50 environmental predictors from ACE metadata. We defined R^2 thresholds based on permuted repetitions of the analysis to only consider AGCs linked to the environment (env-AGC: $R^2 > 10\%$ in the small, 15% in the large size fraction, Figure S5). 89.0% (resp. 82.1%) of the considered AGC were env-AGC in the small (resp. large) size fraction. Over both size fractions, 894,292 models (20.6%) showed R^2 values above 50%, indicating predictability of AGC abundance based only on the environmental context and opening the way for genomic-based correlative models at SO scale^{11,12}. Computing the mean R^2 across AGCs grouped by functional annotation, we identified functions showing consistently high predictability from the environment. For example, in the small size fraction 12 AGCs were annotated as *Response regulator, receiver/PhoQ Sensor / phosphorelay sensor kinase activity*, with a mean R^2 of 75.3%, and 14 AGCs were annotated as *Outer membrane receptor proteins, mostly Fe transport*, with a mean R^2 of 70.1%. Among functions associated with at least 50 AGCs, the three best predicted from the environment were *type VI secretion protein* (56 AGC, mean $R^2 = 0.59$), *large exoproteins involved in heme utilization or adhesion* (61 AGC, mean $R^2 = 0.59\%$) and *magnesium chelatase* (82 AGC, mean $R^2 = 56\%$), followed by *Dimethylsulfoniopropionate lyase* in 14th position (51 AGC, mean $R^2 = 52\%$). Most functions with homogeneously high R^2 were associated to AGCs with increased abundance in one particular set of conditions (e.g., AGCs annotated as *type VI secretion protein* and *magnesium chelatase* were best predicted by O_2 , salinity, biological silica, and sea-ice margin, all typical of Mertz polynya blooming conditions). This was not the case for the biogeochemically important *Dimethylsulfoniopropionate lyase*, associated to 51 AGCs that displayed three highly contrasted biogeographical signatures (Fig. 4). A first group was associated to cold and oxygenated waters with high biological silica, particulate organic carbon and particulate organic nitrogen, i.e., coastal Antarctic primary production conditions. A second group was linked to higher temperatures and lower latitudes, corresponding to subantarctic and subtropical surface waters. Finally, a third group was mainly correlated to depth, trace metals, and phosphate, matching deeper water masses. Highly abundant AGC were found in all three groups, and only 2 had R^2 below 30%. AGCs correlated to conditions of primary productivity in coastal Antarctic were mainly composed of ORFs from ASP10-02a contigs, an abundant lineage of Gammaproteobacteria first described in a Phaeocystis bloom of the Amundsen Sea polynya³². The group associated with warmer water masses was dominated by the globally abundant genera *Planktomarina* and *Pelagibacter*, followed by three genera of Rhodobacteraceae (GCA-002712045, GCA-2697345, and MED-G52). The most

frequent genus in the deep waters group was UBA9410, a diverse Acidimicrobiales genus commonly found in mesopelagic waters³³, followed by other deep-sea associated genera.

To identify groups of microbial genes associated with singular SO biomes without a priori selecting a specific function, we grouped env-AGC based on their abundance profiles and without reference to their sequence identity nor annotations, resulting in 156,671 and 28,756 co-abundant groups (CAGs)³⁴ in the small and large size fractions, respectively. Several CAGs had conspicuous positions in the RDA space (Fig. 5), indicating their specific association with distinct biomes: we first present CAGs specific to the Mertz polynya, before focusing on 3 CAGs illustrating a gradient of polar adaptation across latitudes. In the supplementary materials, we describe two CAGs linked with specialist species thriving in polar conditions (CAGs 131 and 34), two CAGs associated with deep water masses (CAGs 33 and 39), as well as outliers at the SO scale, including CAGs specific to sub-Antarctic islands (CAGs 136, 73614, and 177401).

The genomic signature of an active diatom bloom in the Mertz polynya

Samples taken in Mertz polynya surface waters (near the Mertz Glacier front) exhibited high biomass and chlorophyll a, indicative of an intense phytoplankton bloom. This bloom was dominated by diatoms as shown from 18S rRNA fragments read-level taxonomic annotations. Two CAGs from the large size fraction were enriched in Mertz samples (CAG 15 and 50, Fig. S6). Both mainly contained AGC associated with the diatom *Fragilariopsis cylindrus*, an indicator species of cold water adapted to the polar environment³⁵.

Four CAGs from the small size fraction were enriched in Mertz samples ($RDA2 > 0.5$, Fig. 5A): CAG 79, CAG 29, CAG 137 and CAG 85270 (Fig. 5C, D). They contained below 16% of environmental unknowns (EU) versus 54.6% in the full dataset, and more than 50% of known (K) and known without PFAM (KWP). All four CAGs were significantly enriched (Fisher test, adj. p -value < 0.01) in TonB receptors and TonB-linked outer membrane proteins specialized in the import of degradation products from proteins or carbohydrates as nutrients (SusC/RagA, SusD/RagB). They were also enriched in proteins involved in carbohydrate metabolism, e.g., glycosyl transferase, polysaccharide biosynthesis protein, or glutamine synthetase, in ABC transporters and in proteins involved in cell motility (e.g., gliding motility, morphogenesis, and elongation of the flagellar filament). A variety of metalloprotein were enriched in all four CAGs as well, including heme-binding proteins, M6 family metalloprotease or metal-dependent hydrolase. We provide a complete list of enriched functions in each CAG of interest (Table S1). ORFs from CAGs 79 and 85270 were mainly affiliated with the *Polaribacter* genus. In CAG 137, Rhodobacteraceae dominated all annotation types, mainly from the Octadecabacter and Yoonia genera. In CAG 29, HTCC2207 SAR 92, and ASP10-02a were the most represented genera. ASP10-02a is a key player in primary production co-limitation by micronutrients³⁶, and we identified it as the main carrier of DMSP lyases in the Mertz polynya. SAR 92 is a widely distributed oligotrophic clade known for its ability to consume polysaccharides in the epipelagic zone, notably through TonB-dependent receptors³⁷. It has been associated with late stages of diatom-induced bacterioplankton blooms in the North Sea, uptaking and degrading specific polysaccharides, including chrysolaminarin³⁸. *Polaribacter* and SAR 92 have both been associated with Phaeocystis-produced chrysolaminarin degradation in the SO³⁹.

In addition, the four Mertz-associated CAGs from the small size fraction were enriched in phage integrase, and three out of four were enriched in phage plasmid primase P4, suggesting a strong viral presence in the polynya. To better characterize the viral communities at Mertz, we annotated all ACE contigs using geNomad⁴⁰. Out of 68,074,004 contigs, 967,226 (1.4%) were annotated as viral. 69,804 and 20,686 viral contigs produced ORFs clustered in AGCs of the CAGs

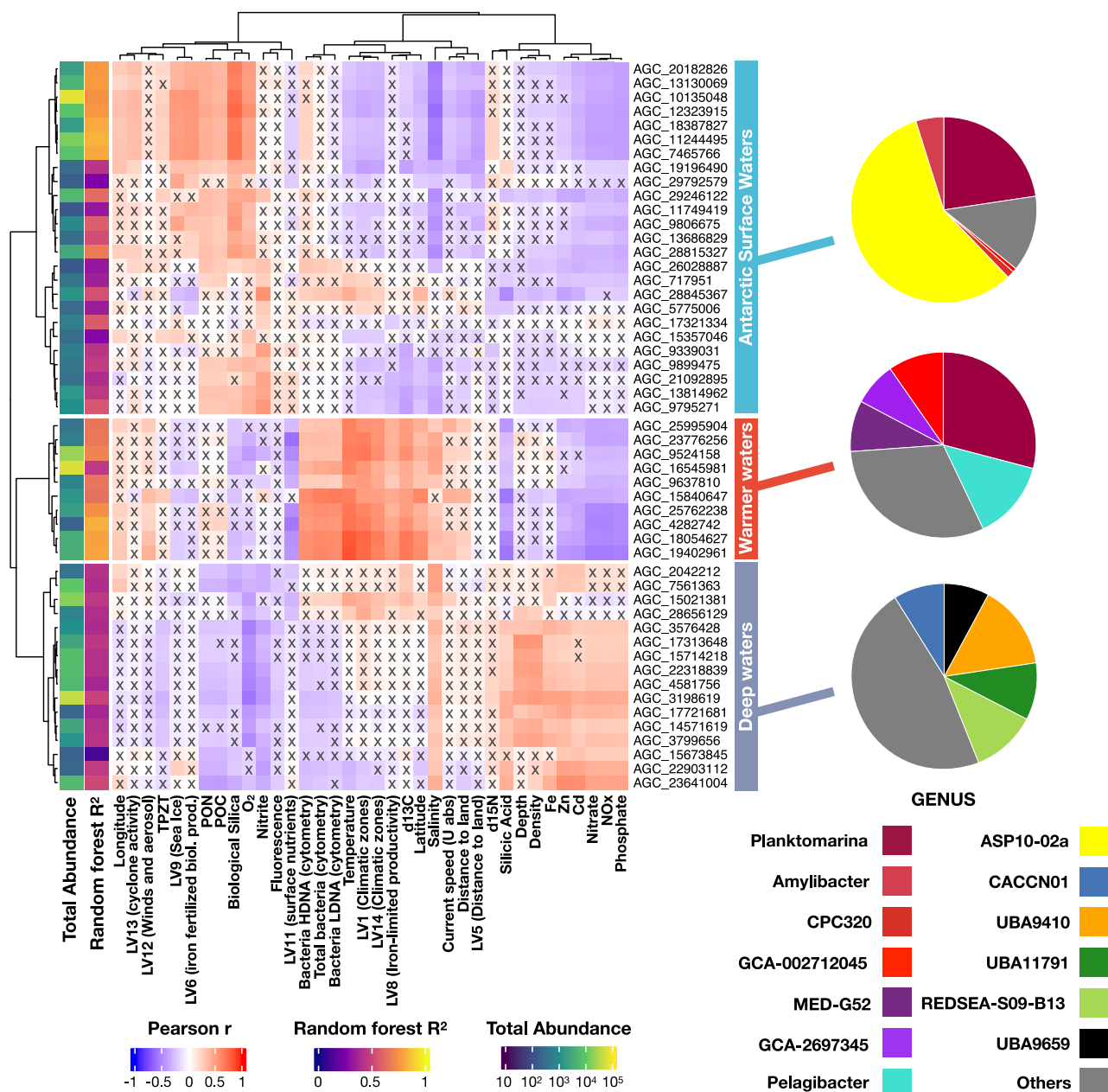


Fig. 4 | DMSP lyases form three distinct biogeographical groups at SO scale. Correlation matrix of the 51 AGC annotated as DMSP lyase versus environmental variables, in the small size fraction. AGC (lines) and environmental variables (columns) are hierarchically clustered based on their correlation patterns. Non-significant correlations (adjusted p -value > 0.05) are indicated by a gray cross. LV in environmental variables names stands for latent variables, corresponding to the ones defined in Landwehr et al.²⁰, using ACE environmental metadata. The row annotation layers on the left of the correlation matrix indicates the total normalized coverage abundance of each AGC, and the R^2 of the associated random forest

model. Three groups of AGC were defined using hierarchical clustering, as indicated by the row annotation layer on the right side of the correlation matrix. For each group of AGC, genus-level taxonomic annotations are represented as pie charts computed using contig-level taxonomic annotations based on the Genome Taxonomy DataBase (GTDB)²⁶. The taxonomy of all contigs containing an ORF clustered in an AGC of each group was considered, and only the five most occurring genera were represented for each group, the rest being regrouped in the “Others” category. Source data are provided as a Source data file.

29 and 79, respectively. 82.1% of the viral contigs associated to CAG 29 were annotated as Caudoviricetes (tailed bacteriophages), and the Autographiviridae family was particularly enriched (4.16 odds ratio when compared to all viral contigs). Caudoviricetes also dominated viral contigs associated to CAG 79, representing 65.9% of all annotations. 19.6% of the viral contigs associated to CAG 79 were annotated as Megaviricetes (giant viruses, 6.4% of annotations in CAG 29 contigs). The Mimiviridae family was particularly enriched in CAG 79 (odds ratio of 3.7 when compared to all viral contigs), indicating the presence of

giant viruses targeting eukaryotes at Mertz. To ensure a proper annotation of the contigs from this complex viral lineage, we re-annotated all contigs identified as Mimiviridae through a bioinformatic pipeline specifically designed for giant viruses annotation. Of the 12,371 contigs re-annotated, 9935 (80.3%) were confirmed as Mimiviridae or Mesomimiviridae, 1150 (9.3%) were annotated to other Megaviricetes and 928 (7.5%) could not be annotated to giant viruses. The remaining 358 contigs (2.9%) were annotated as Mirusviricota, a recently described novel Phylum⁴¹. In CAG 79, 882 contigs were

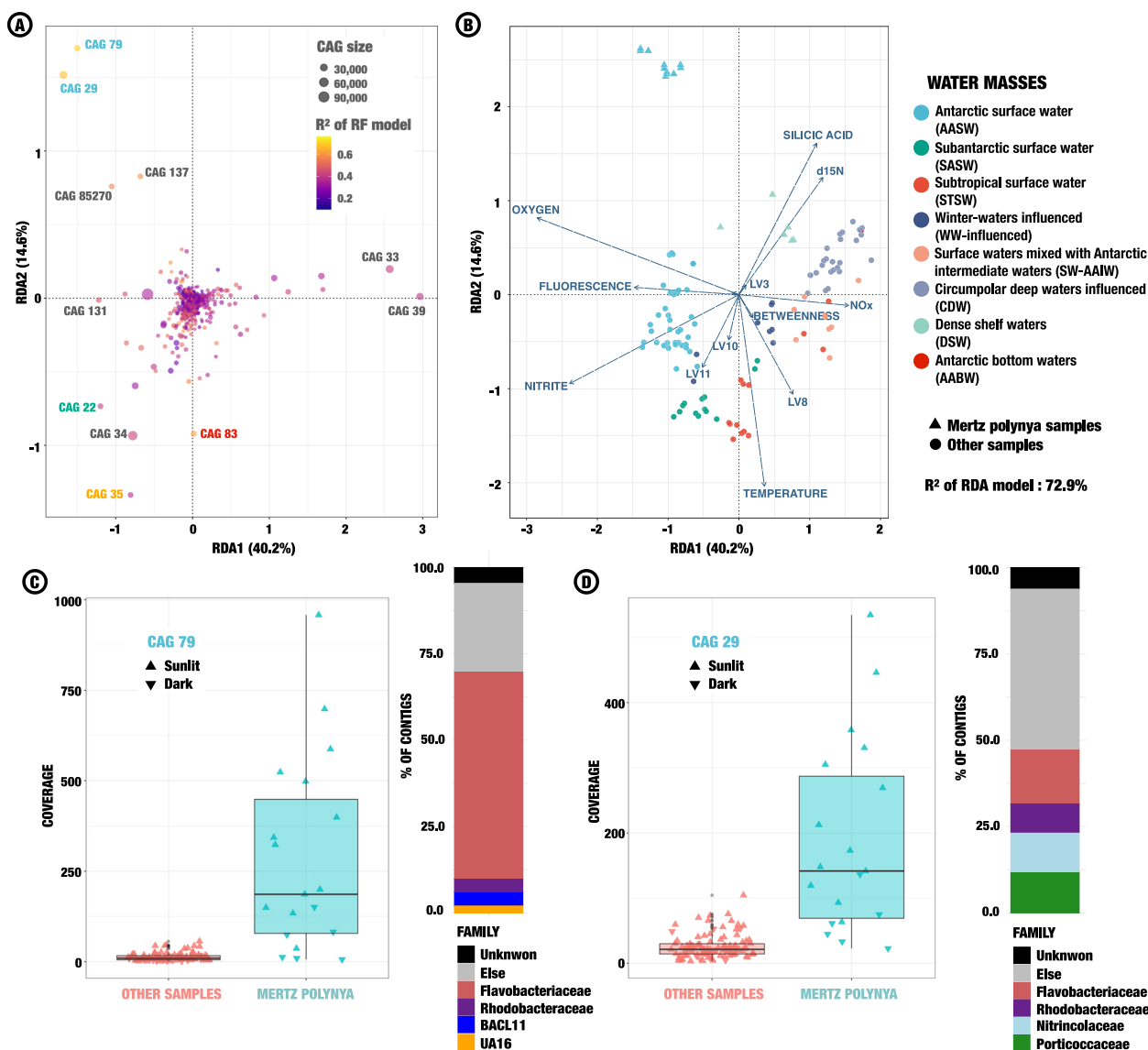


Fig. 5 | The response of co-abundant groups of env-AGCs (CAGs) to environmental gradients at SO scale highlights Mertz polynya's originality. Redundancy analysis (RDA) of CAGs abundances in response to environmental variables is shown for the small size fraction (A, B). **A** Distribution of CAGs in RDA space, colored by their mean random forest R^2 (predictability of their abundance from environmental data). Dot size is proportional to CAG's size (in number of env-AGC). CAGs of interest mentioned in this study are labeled: the two plotted in (C, D) in light blue, the three investigated in Fig. 6 in red, orange, and green (as in Fig. 6), and those covered in supplementary materials in gray. **B** Samples and environmental variables distribution in the same RDA space. Samples are colored by water mass. The first axis opposes surface samples with high fluorescence and oxygen ($RDA1 < 0$) from deeper samples showing high NO_x concentrations ($RDA1 > 0$). The second axis separates warm Subtropical Surface Waters samples ($RDA2 < 0$) from colder samples, isolating Mertz samples from other Antarctic Surface Waters samples ($RDA2 > 0$). Latent Variables (LV), defined in Landwehr et al.²⁰, using ACE

environmental metadata, include LV3 (meridional cold- and warm-air advection), LV8 (iron-limited biological productivity), LV10 (diel cycles) and LV11 (surface nutrient concentration associated with mixing events). An equivalent RDA for the large size fraction is presented in Fig. S6. Mertz polynya's originality is further illustrated in (C, D), showing the abundance and taxonomy profile of the two CAGs most linked to this biome. Boxplots compare CAG's coverage at Mertz (19 samples) versus in other samples (108 samples), displaying the median, interquartile range (1st and 3rd quartile) and whiskers extending to most extreme values within $1.5 \times$ the interquartile range. Each individual sample is plotted, shaped according to categorical depth: sunlit (150 m and above) and dark (below 150 m). Family-level taxonomic profiles are represented next to each boxplot, as estimated through contigs taxonomic annotations using GTDB as reference. For each CAG, only the four most abundant families are colored, the rest being aggregated as "Else". Source data are provided as a Source data file.

confirmed to be from mimiviruses and 90.8% of their ORFs had an eggNOG functional annotation. The most observed ones were related to host cellular machinery manipulation (*Ulp1 protease family*, *N-methyltransferase activity*, *Nuclease activity*, *Transcription factor TFIID/TATA-binding protein*). More surprisingly, we identified 25 occurrences of functions related to Zinc fingers domains, including two ORFs annotated as *Zinc finger*, *C3HC4 type (RING finger)*, found on

two similar contigs assembled in two distinct Mertz samples and both also containing a *Large eukaryotic DNA virus major capsid protein* and a *GIY-YIG catalytic domain*. Only one Mirusviricota contig was associated to CAG 79. Contrastingly, 218 (60.9%) mirusvirus contigs were associated to CAG 29. Across the 2316 ORFs of these 218 contigs, 83.4% lacked an eggNOG functional annotations.

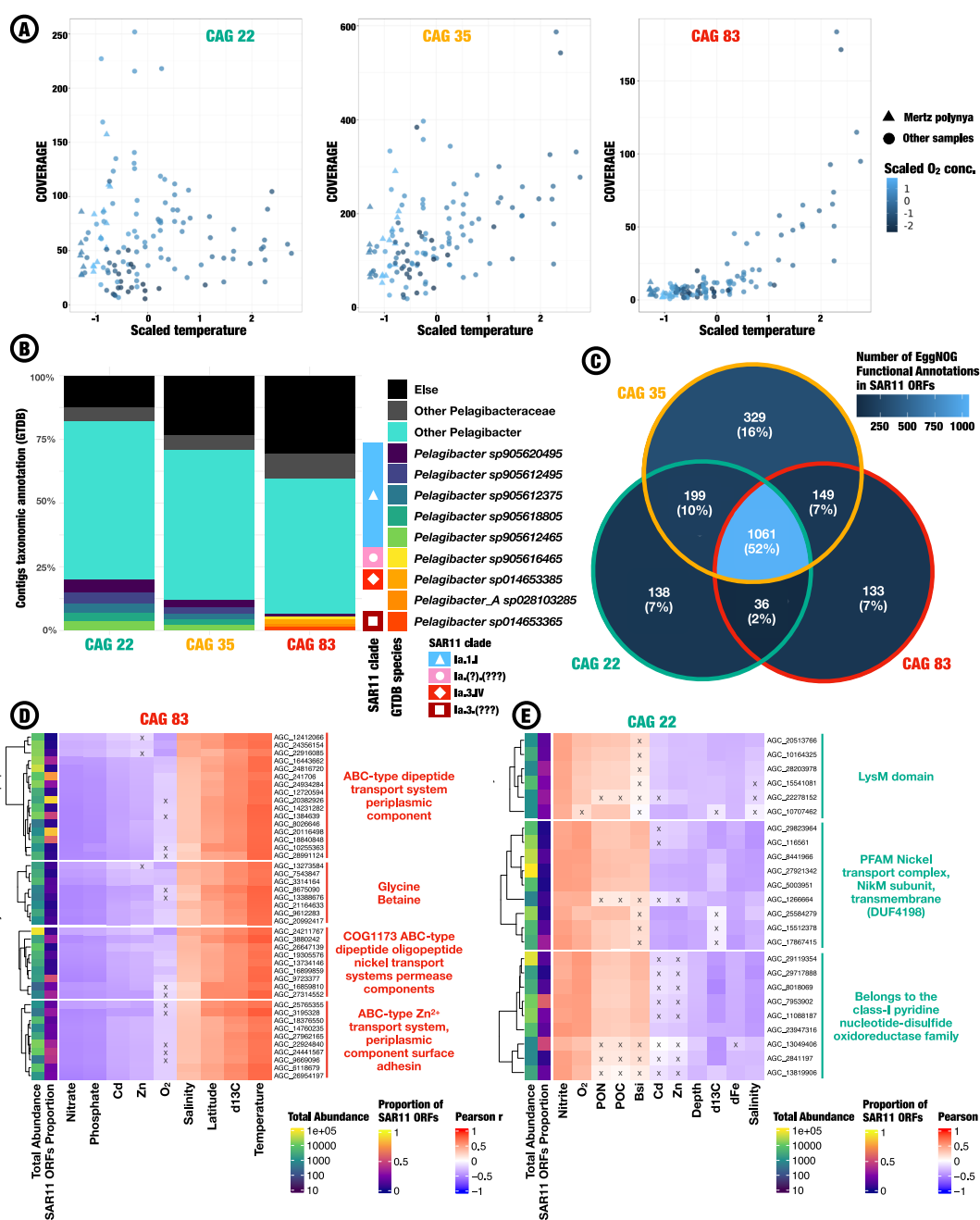


Fig. 6 | Functional changes across latitudes in Pelagibacter-related CAGs.

A Abundance of three CAGs dominated by *Candidatus Pelagibacter* (SAR11) as a function of scaled temperature and oxygen. **B** Pelagibacter species composition based on parent contigs annotations in these three CAGs (only 5 most abundant species shown). Other Pelagibacter corresponds to parent contigs annotated to the Pelagibacter genus outside of the 5 dominant species. Clades of SAR11 were assigned to each of the most abundant GTDB species based on a phylogenomic analysis of GTDB representative genomes together with the ones used in Freil et al.⁷⁹. **C** Venn diagram of Pelagibacter ORFs' functional annotations observed in the

three CAGs. Background colors and white numbers indicate the counts of unique eggNOG functional descriptions for each section of the diagram. Correlation heatmaps in **(D, E)** illustrate the relationship between environmental variables and individual AGCs annotated to the most observed functional annotations specific to CAG 83 **(D)** and 22 **(E)**. Rows of the heatmaps are decorated first with the total abundance of each CAG throughout the small-size fraction samples, and then the proportion of Pelagibacter ORFs within each AGC. Source data are provided as a Source data file.

Overall, all CAGs enriched near the Mertz glacier front showcase a distinctive polynya genetic seascape, with organisms specifically adapted to polar blooms conditions, enriched in organic matter processing genes, with indication of intense viral activity and the presence of abundant phycosphere members involved in alleviating primary production nutrient co-limitation.

Global latitudinal shifts across gene groups illustrate adaptation to polar conditions in the ubiquitous Pelagibacter

CAGs 83, 35, and 22 exhibited similar taxonomic profiles despite showing distinct responses to environmental data (Figs. 5A and 6). CAG 83 was specific to warmer waters, with low abundance in all Antarctic Surface Waters samples supposing a latitudinal boundary at

the polar front (Fig. 6A). CAG 35 was positively correlated to temperature yet ubiquitous, even in the coldest waters. Finally, CAG 22 was present in all water masses, with higher abundances in cold Antarctic Surface Waters. These three CAGs were all dominated by Pelagibacteraceae genes (Fig. 6B), and respectively 88.75%, 83.5%, and 59.48% of their AGCs contained at least one ORF from a contig annotated as *Candidatus Pelagibacter* (SAR11 and relatives) according to UniRef90. Using the species-level annotations of GTDB, we identified the three most detected species in CAG83 to be sp014653385, sp028103285, and sp014653365, all having representatives found in temperate areas (Fig. 6). Dominant species in CAGs 35 and 22 all had representatives corresponding to SAGs (single-amplified genomes) sampled beneath the Ross Ice Shelf. Through a phylogenomics analysis, we were able to associate all representative genomes from these species to clade Ia.1.I (Fig. 6B). We extracted the AGCs containing at least one ORF from a *Pelagibacter* contig in each CAG and compared their eggNOG functional annotations (Fig. 6C). A total of 600 functions were found in only one CAG, potentially illustrating functional adaptations to distinct environmental conditions. In CAG 22, the most observed specific functions were the *transmembrane NikM subunit* (in fact matching the DUF4198, see discussion) and the *class-I pyridine nucleotide-disulfide oxidoreductase*, both carried by *Pelagibacter* ORFs in 9 AGCs, followed by the *UPF0234 family* (unknown function, unshown in Fig. 6), found in 8 AGCs, and the *LysM domain*, found in 6 AGCs (Fig. 6E). In CAG 83, the three most observed specific functions were linked to ABC transporters, two of which directly linked to trace metals Nickel and Zinc (Fig. 6D). The fourth most observed was *glycine betaine* (more specifically matching ProX/OpuAC substrate-binding proteins for glycine betaine and osmoprotectant). See Table S1 to access the complete list of functions specific to each CAG.

Discussion

Analyzing 218 metagenomes from a circumpolar expedition, we were able to characterize the originality and biogeography of SO's microbial genomic diversity. We identified a set of genes distributed at both poles while being absent from most other latitudes on the planet; some of them are involved in adaptations to polar-specific constraints like UV-exposure and cold temperatures. Our gene-centric approach, involving clustering steps agnostic to gene function, illustrated how microbial gene assemblages from the SO are largely endemic and unknown. The high number of singletons in our assemblies suggests the presence of a significant proportion of rare genes in the SO, showing no deep homologies with each other. This could partly be due to ORFs from the large size fraction, which should be treated with caution due to the difficulty of detecting good quality ORFs from eukaryotic contigs⁴² (cf Methods). Yet, 53.3% of all singletons came from prokaryote-dominated samples of the small size fraction, and singletons from both size fractions were largely dominated by unknowns (83.8% in the small, 93.2% in the large), suggesting a limited impact of eukaryotic contigs on our conclusions. Collector curves' asymptotic profiles were stronger when decreasing detection thresholds (Fig. S3), suggesting that singletons do recruit reads in multiple samples independently of their size fraction of origin. Otherwise, they would remain undetected in all or most samples whatever the threshold, preventing the asymptotic form of the curve. It suggests they do share distant homologies, e.g., at domain level, with unassembled genes across multiple samples. The decrease in taxonomic diversity in the SO compared to subtropical latitudes⁴³ could then be balanced by an abundance of diverse yet individually rare genomic elements distributed at SO scale. This hypothesis will have to be confirmed by further explorations of ACE singletons, of which the majority was excluded from our biogeographical analysis to focus on widely distributed genes. This could be achieved through network-based methods allowing the characterization of distant and rare homologs⁴⁴,

by using protein structural alignment for gene cluster formation⁴⁵, or by using large language models⁴⁶ for function predictions.

Within the SO, gene assemblages were structured by water mass, supporting the observation that processes leading to water mass formation and transport exert the strongest control on microbial community composition^{28,30}. We identified gene classes with strong biogeographical structure in their distribution, highlighting the importance of functions related to cell-to-cell communication, production of chlorophyll and transport of key elements such as iron. Among these classes, genes coding for DMSP lyases were abundant in contrasting biomes, with abundances well predictable from the environmental context. Previous observations of DMS/DMSP-related enzymes in the Southern Ocean identified higher enzyme abundances at depth compared to the surface, yet did not identify significant surface biogeographical patterns⁹. Here, by exploring DMSP lyases abundance at the gene-cluster level and in a large set of conditions, we found abundant DMSP lyases carried by different organisms and structured in three biogeographical groups corresponding to distinct environmental niches. We found the ASP10-02a lineage to be the main carrier of DMSP lyase in contexts of coastal primary production in the SO. This lineage recruited up to 11% of reads in a metagenomic analysis of the Amundsen Sea polynya³², and was identified as the main meta-transcriptomic and metagenomic reads contributor to cobalamin biosynthesis genes in coastal McMurdo Sound³⁶. This lineage seems to live in close mutualistic relationship with phytoplanktonic cells³⁶, and our results suggest it could be responsible for the majority of DMS emissions during SO phytoplankton blooms. ASP10-02a therefore emerges as a key contributor to the biogeochemical processes in the SO, and quantifying its impact on global fluxes of carbon and sulfur should be considered as a priority for future studies. We also identified abundant DMSP lyases in deep samples, in lineages like Acidimicrobiales that are adapted to deep-sea environments and known to use DMSP as a carbon source, mainly through DMSP demethylation⁴⁷. The DMSP lyase DddD was identified as key for deriving carbon in coastal Oceanospirillales⁴⁸, and was already known to be abundant in deep polar samples⁹. We can thus hypothesize that lineages like Acidimicrobiales of the UBA9410 genus rely on DMSP lyase activity to retrieve carbon in the mesopelagic SO. Overall, given the distinctive prevalence, diversity and biogeography of DMSP Lyases genes, our results show that DMSP metabolism plays a central role in SO's planktonic communities.

Grouping AGC by distribution patterns instead of functions, we identified numerous gene clusters differentially abundant in the Mertz polynya. Corroborating previous findings we identified polynya bacterial communities as mostly heterotrophs exploiting residues from eukaryotic phytoplankton blooms³⁰, including taxa playing key roles in alleviating primary production limitation by iron and other micro-nutrients like cobalamin^{18,36}. Interestingly, we also found that this polynya exhibited a high viral diversity with bacteriophages and giant viruses contigs, including from the newly described Mirusviricota Phylum⁴¹. The poor level of functional annotations observed in mirusvirus contigs compared to Mimivirus contigs suggests different life styles and genomic adaptations across the two Phylum in the SO, despite their co-occurrence at Mertz polynya. Among Mimivirus contigs, we identified multiple occurrences of regulatory zinc-finger proteins, which was among the polar-specific functions we identified. Zinc-binding proteins were proposed as adaptations for life at the poles in Diatoms^{35,49}. Giant viruses carry distinct gene repertoires in polar environments, yet previous studies did not find evidence of co-adaptations between them and their hosts⁵⁰. Our results, which unlike precedent studies, include samples from a eukaryotic bloom in the SO, suggest that giant viruses and their diatom hosts could carry similar adaptations to high zinc demand.

A previous study investigating genomic composition across a transect from Tasmania to Mertz identified the polar front as the main

biogeographical boundary, but lacked samples between the front and Mertz to test the effect of the continental shelf⁵¹. Our results suggest a larger difference between populations of the polynya *versus* other Antarctic Surface Waters populations than between populations on both sides of the polar front. Coastal Antarctica thus appears to harbor a distinctive genetic repertoire, the diversity of which remains to be explored. Mertz was the only sampling location on the Antarctic shelf in our dataset: our observations could thus hardly be considered representative of shelf conditions at SO scale. We find them more likely to be polynya-specific, as they seem to be driven by the high activity of a *Fragilariopsis cylindrus*-dominated bloom. Interestingly, our Mertz-associated CAGs were similar to genomic markers of a polynya in the Amundsen Sea dominated by *P. antarctica*²⁸, suggesting bacterial functional redundancy in polynyas independently from the dominant phytoplankton lineage. Given the remarkable activity and contribution of these coastal systems, Southern Ocean polynya diversity – both organismal and process-wise – constitutes a key investigation topic for future polar research.

Better understanding adaptation to polar environments in planktonic biomes was a major objective of this study. We found this was best exemplified by gene clusters showing different latitudinal niches and functional profiles despite all being associated with the ubiquitous *Pelagibacter*. *Pelagibacter* subclades adapted to SO's extreme conditions have been observed through amplicon sequencing off the Kerguelen Islands⁵², while *Pelagibacter* genomes assembled from Arctic metagenomes contained polar-specific gene content, the vast majority of which coded for poorly characterized proteins⁵³. Our results suggest a genomic adaptation of *Pelagibacter* across oceanographic fronts transitioning from subtropical surface waters to Antarctic surface waters, even including the specific conditions of a polynya bloom: strong competition for nutrients, organic matter, and trace metals. We notably show that a series of ABC transporters, including one associated with nickel and one associated with zinc, as well as glycine betaine, an osmoprotectant for which *Pelagibacter* has high affinity and multifunctional transporters⁵⁴, are only detected in warmer, more saline waters. Conversely, *pyridine nucleotide-disulfide oxidoreductase*, which could help against oxidative stress, and *LysM*, which allows cells to bind to glycans, are associated with oxygenated, high nitrite, and particulate organic matter waters. Interestingly, the NikM subunit was also identified as specific to CAG22, but all ORFs annotated as NikM by eggNOG matched the DUF4198 domain, suggesting they do not directly transport nickel⁵⁵. DUF4198 has been identified as involved in the response to iron limitation in transcriptomics studies of marine⁵⁶, freshwater⁵⁷, and pathogenic⁵⁸ bacteria. Our results then suggest that this domain could be a marker of adaptation to trace metals limitation in *Pelagibacter*, and push for its functional characterization. We highlight distinct functional repertoires across CAGs related to various SAR11 strains within the different biomes of the SO, notably suggesting a polar adaptation of clade Ia.1.1 strains in relation with trace metals and organic matter availability. A genome-centric, strain-resolved population genomics analysis of ACE metagenomes could thus provide valuable insights into SAR11 Southern Ocean-adapted ecotypes.

By compiling catalogs of contigs, unigenes, AGC and CAGs from across the SO, we provide a robust basis for any future polar and/or global-scale meta-omics investigation (Table S1). Doing so, we address a critical gap in the metagenomes currently available in public databases¹⁴. We notably provide the Southern Ocean Reference Gene Catalog (SO-RGC), focused on the 0.2–3 µm size fraction and complementary to the OM-RGC^{21,22}, and a catalog of polar-specific ACE ORFs, i.e., detected in at least one Arctic sample while being absent from non-polar Tara Oceans samples. Overall, our results advocate for the development of regional-scale descriptions and models of planktonic diversity in the Southern Ocean, distinguishing coastal and off-shore systems, and implementing the specific response of prokaryotes

to localized eukaryotic blooms. Our statistical results suggest that our gene catalog, combined with extensive environmental and biogeochemical monitoring, could lead to correlative models of gene abundance at SO scale, offering new tools to predict the future of this rapidly changing ecosystem. It is worth noting that the ACE campaign ran from spring to late summer and some of the variability observed could be temporal, as illustrated by the strong seasonal dynamics of viral communities of Marguerite Bay⁵⁹. The genomic content of microbial populations in the dark winter of the SO remains to be unraveled. Existing Antarctic time-series (e.g., the Palmer LTER⁶⁰ or the Rothera time-series⁶¹) should thus be complemented by genomic time series to provide valuable insights into seasonal cycles and enhance our ability to monitor and predict the impact of climate change on Southern Ocean microbial communities.

Methods

A list of all publicly available resources is available in Table S1.

Sampling and sequencing protocols

218 samples for metagenomics analyses were collected at 34 stations during the ACE campaign in the Austral summer 2016–2017. 197 of the 218 samples, thereafter called CTD samples, were collected from Niskin bottles during rosette upcast and separated into three size fractions (0.2–3 µm, 3–200 µm, and 0.2–40 µm). The remaining 21 samples, thereafter called UDW samples, correspond to water pumped directly from the surface and separated into the same three size fractions. Samples were sent for DNA extraction and shotgun sequencing to Genoscope, the French National Platform for DNA Sequencing, following protocols used by *Tara* expeditions⁶². Briefly, after filter cryogrinding, DNA was extracted using total RNA/DNA Purification and Nucleospin RNA/DNA Buffer Set (MACHEREY-NAGEL). Metagenomic libraries were prepared using the Illumina kit according to manufacturer instructions. DNA libraries were sequenced on a Novaseq 4000 instrument, with a target of 100 M paired-end reads per library (2 × 150bp; 500 bp insert size).

Environmental metadata compilation

The ACE campaign hosted 22 scientific projects encompassing biology, oceanography, climatology, glaciology, and biochemistry. For each CTD sample, all available metadata from the corresponding cast and depth were retrieved from [SPI-ACE repository](#). Similarly, metadata from each pumping event were retrieved for UDW samples, but considering the limited number of sequenced UDW samples and the lack of homogeneity in measured variables across CTD and UDW samples, these metadata and their corresponding samples could not be used in statistical investigations based on environmental variables (i.e., random forest models, RDA). For 10 of the 21 UDW samples, surface CTD metadata from the same sampling event were available, enabling us to incorporate these samples in statistical investigations along with CTD samples, while the remaining 11 UDW samples could not be considered. Up to 56 variables were retrieved per CTD samples, including basic physico-chemical variables (e.g., temperature, salinity, nutrients, depth), trace metals concentrations (e.g., dissolved Fe, Cu, Ni, Zn), isotopes (e.g., ¹³C, ¹⁵N) and pigment-based measures (e.g., concentrations of cyanobacteria, diatoms or haptophytes). Variables measured in less than half of the samples were dropped for further statistical explorations, leading to the selection of 33 variables. In addition to these data retrieved in situ, physical variables were calculated at each sampling using a Lagrangian approach and an integration time of 10 days. These included current velocity, Okubo-Weiss (a proxy of eddy presence) and Lagrangian betweenness (a proxy of bottleneck presence which has been related to biodiversity⁶³). Fourteen latent variables computed through a sparse PCA for each ACE station to summarize the global biogeochemical context²⁰ were added to the metadata set. Please refer to the original Landwehr et al.²⁰ study for a

full description of each latent variable. Finally, each sample was associated to a Longhurst biogeographical province based on its coordinates, and to a water mass based on temperature-salinity-oxygen diagrams computed from CTD downcast profiles (Fig. 3A). When needed, missing values in the CTD metadata set were imputed using the *k*-nearest-neighbors approach encoded in the *caret* R package⁶⁴, with the default value of *k* = 5. For a full list of available metadata variables, a precise description of their compilation and of their pre-processing, please refer to this GitHub repository: [ACE_gene_centric_scripts](#).

Assembly of metagenomic short reads and the profiling of resulting contigs

Short-reads were quality-filtered using the Minoche⁶⁵ approach implemented in *illumina-utils* v2.3 with default parameters, and sample-by-sample assemblies were obtained from MEGAHIT v1.2.9⁶⁷. The 68,074,004 contigs from the 218 single assemblies were concatenated into a FASTA file from which a single *anvi'o* contigs database, hereafter called the ACE Contigs-DB, was generated using the program *anvi-gen-contigs-database* as implemented in *anvi'o* v8⁶⁸. During the generation of the ACE Contigs-DB open-reading frames were detected in all contigs using Prodigal v2.6.3⁶⁹ which resulted in 175,336,776 non-dereplicated ORFs that represented the raw ACE gene catalog for downstream analyses. To estimate the fraction of eukaryotic organisms sampled, *Phyloflash* v3.4⁷⁰ was used on quality-filtered reads. Considering that some samples were dominated by eukaryotes, it is likely that some contigs in the ACE contigs database are from eukaryotic organisms. To assess this likelihood, *Eukrep* v0.6.71 (West et al.) and *Whokaryote*⁷² v0.0.1 were used to try to detect eukaryotic contigs. However, only 2,343,800 contigs (3.4%) were classified as eukaryotic by both tools, clearly underestimating the eukaryotic fraction of contigs. The annotation of these contigs using the UniRef90 database and MMSeqs2 v14.7e284⁷³ demonstrated the presence of 229,179 (9.8%) potential false positives annotated as bacteria. We thus decided to keep all contigs in the database for the rest of the pipeline, while tagging the ones identified as eukaryotic by *EukRep* as potentially eukaryotic.

All contigs were annotated taxonomically using *Kraken2*⁷⁴ v2.1.3 and the Genome Taxonomy Database (GTDB)²⁶. *GeNomad* was used to detect and annotate viral contigs⁴⁰. Subsets of contigs of particular interest were further annotated: contigs containing ORFs found in CAGs of interest were annotated using MMSeqs v14.7e284⁷³ and the UniRef90 database. Contigs annotated to *Mimiviridae* by *geNomad* were ran through a specific bioinformatic pipeline: for each contig, proteins were predicted using *prodigal* v2.670 and aligned using *diamond*⁷⁵ v2.18 (“-ultra-sensitive” option, and percent identity of at least 30%) against a global protein resource of manually characterized and curated MAGs from the sunlit ocean and corresponding to bacterial, archaeal, eukaryotic and plastid populations^{76,77} as well as *Nucleocytoviricota*⁷⁸ and *Mirusviricota*⁴¹. Contigs were assigned to *Nucleocytoviricota* if at least 25% of the corresponding proteins had the best hit for a reference *Nucleocytoviricota* protein, and if this percentage was above that of any other high-rank taxonomic category (*Bacteria*, *Archaea*, *Eukarya*, *plastids*, *Mirusviricota*). The same rules were applied for contig assignment to *Mirusviricota*. In addition, contigs were assigned to a taxonomic level when at least 50% of their proteins had a best match for the same one.

To connect GTDB annotations of SAR11 species to the currently used clade nomenclature, we integrated the selected GTDB representative genomes into a phylogenomic analysis together with those used in Freil et al.⁷⁹. The analysis was achieved in *anvi'o* v8⁸⁰ using the concatenated ribosomal proteins described in the *Bacteria_71* model in the platform (*Ribosomal_L1*, *Ribosomal_L13*, *Ribosomal_L14*, *Ribosomal_L16*, *Ribosomal_L17*, *Ribosomal_L18p*, *Ribosomal_L19*, *Ribosomal_L2*, *Ribosomal_L20*, *Ribosomal_L21p*, *Ribosomal_L22*,

Ribosomal_L23, *Ribosomal_L27*, *Ribosomal_L27A*, *Ribosomal_L28*, *Ribosomal_L29*, *Ribosomal_L3*, *Ribosomal_L32p*, *Ribosomal_L35p*, *Ribosomal_L4*, *Ribosomal_L5*, *Ribosomal_L6*, *Ribosomal_L9_C*, *Ribosomal_S10*, *Ribosomal_S11*, *Ribosomal_S13*, *Ribosomal_S15*, *Ribosomal_S16*, *Ribosomal_S17*, *Ribosomal_S19*, *Ribosomal_S2*, *Ribosomal_S20p*, *Ribosomal_S3_C*, *Ribosomal_S6*, *Ribosomal_S7*, *Ribosomal_S8*, *Ribosomal_S9*, *ribosomal_L24*). The final tree was generated with *FastTree* v2.1.11⁸¹.

Generation and annotation of Southern Ocean's microbial reference gene catalog

Open-reading frames were detected in all contigs using Prodigal v2.6.3⁶⁹. We clustered the 175,336,776 ORFs from the raw ACE gene catalog at 95% similarity and 90% coverage using CD-Hit V4.8.1⁸², to produce unigenes comparable to those of the OM-RGC computed from Tara Oceans and Tara Polar Circle expeditions. The 89,739,060 unigenes produced constitute the full ACE reference gene catalog (ACE-RGC). The ACE-RGC was annotated with *EggNOG-mapper* v2.1.8⁸³ and *KOFamSCAN* v1.3.0⁸⁴. To allow easier usage in conjunction with the OM-RGC, in which only the 0.2–3 µm size fraction is included, the SO-RGC was defined as the unigenes from the ACE-RGC that contained at least one ORF detected in a contig assembled in the 0.2–3 µm size fraction. Finally, to produce deep homology clusters, the AGNOSTOS clustering pipeline²⁵ was used on the raw ACE gene catalog to produce 30,123,228 AGNOSTOS gene clusters (AGC), of which 765,003 were discarded as low quality. The 29,358,225 good quality AGC were classified in 4 categories based on their PFAM annotation and their similarity with the members of the AGNOSTOS-DB: Known (K), Known without PFAM (KWP), Genomic unknown (GU; genes of unknown function yet found in a genomic context—MAG, SAG, isolate genome...) or Environmental unknown (EU; genes of unknown function never observed in a genomic context). For a detailed description of these categories and of the methodology for clustering and annotating within the AGNOSTOS pipeline, please refer to Vanni et al.²⁵. Note that the AGC we use in this study are issued from an AGNOSTOS-based clustering and annotation of ACE ORFs, and not to an integration of ACE ORFs within the public AGNOSTOS gene database. AGC-level *EggNOG* and *KEGG* annotations were defined as the modal value from the annotations of all cluster's members.

Computation of gene- and cluster-level coverage and detection

Quality-filtered short reads were mapped on the ACE contigs DB to produce contigs-level coverage and detection (percentage of the contigs covered at least at 1X) profiles, using *Bowtie2* v2.4.5⁸⁵, in competitive mode (i.e., no multi-mapping) with equivalent mapping scores across different references distributed at random. Gene-level metrics were obtained for all the raw ACE gene catalog through the program *anvi-profile-blitz* () implemented in *anvi'o*^{68,80} for this project. By deriving gene-level metrics from the larger genomic context afforded by contigs, rather than using read recruitment to individual gene sequences, we were able to (1) avoid bell-shaped coverage signal that would dwindle around ORF extremities, (2) avoid mapping errors due to assembled sequences being removed from the reference during pre-mapping de-replication, and (3) use exhaustive contigs-level metrics to build direct links between gene-level results obtained in this study and MAGs-level results potentially obtained from the same metagenomic assemblies. The coverage values reported from *anvi-profile-blitz* were expressed per base-pair, i.e., normalized by gene length. Outputs from all samples were then concatenated into a coverage matrix and a detection matrix of each 218 columns and 175,336,776 lines where each line represented an individual ORF.

The coverage of each unigene was defined as the sum of the per-base pair coverages from all members of its dereplication cluster. Similarly, per-base pair coverages of all members of each AGNOSTOS

cluster were summed to obtain AGC-level coverages. To avoid false-positive coverage values due to mapping mistakes and read dilution across conserved domains, a threshold of detection was applied at the cluster level. Detection at the cluster level was defined as $\text{Detection}_{\text{Cluster}} = \max(\text{Detection}_{\text{Cluster members}})$.

Increasing the threshold of detection at cluster level caused both the mean slope of the collector curve and the amount of undetected AGNOSTOS clusters to increase (Fig. S3). A flat collector curve is likely to be the result of false positives, considering the many singletons that are likely to be rare, but a high number of undetected clusters is likely to be due to false negatives, since their sequences should be present at least in the samples in which they were assembled. We then decided to use a threshold of 60% detection, as it was the highest threshold allowing to detect more than 95% of AGNOSTOS clusters in at least one sample. To apply this threshold, all AGC-level coverage values corresponding to AGC-level detection scores below 60% were turned to 0.

Pre-processing and normalization of cluster-level abundances

After applying the 60% detection threshold, all remaining coverage values were rounded to the nearest integer. The whole AGNOSTOS cluster-level coverage matrix was then normalized using the relative log expression method from the DESeq2 R package⁸⁶. The relative log expression normalization method was identified as one of the most adapted to metagenomics-based microbiome studies⁸⁷.

Highlighting the originality of Southern Ocean's microbial genes

Protein sequences from the ACE-RGC were clustered with those of the OM-RGC and the NMPFamsDB at 30% similarity threshold, with a coverage threshold of 80%. Clusters were separated in three categories: pure ACE when only composed of sequences assembled in ACE samples, pure TARA/NMPFam when only composed of sequences from the OM-RGC or NMPFamDB, respectively, and mixed for the rest. Clusters were further characterized based on their members' origin of assembly, mainly distinguishing sunlit (<150 m) and dark (>150 m) samples as well as the different size fractions.

To better estimate the global presence of genes from the ACE-RGC, short reads from 134 samples from Tara Oceans and Tara Polar Circle expeditions corresponding to the 0.2–3 μm size fraction were quality-filtered and mapped on the ACE contigs DB using the protocol described in *Computation of gene- and cluster-level coverage and detection*. Description of the Tara samples is available in Salazar et al.²¹.

Identifying environmental drivers of gene-clusters distribution at Southern Ocean's scale

For further biogeographical explorations, the global matrix was split into two parts, the small size fraction part corresponding to 0.2–3 and 0.2–40 μm filter sizes, and the large one corresponding to the 3–200 μm size fraction. The pooling of 0.2–3 and 0.2–40 μm size fractions was motivated by (1) the small amount of 0.2–40 μm samples compared to the two other size fractions and (2) the high similarity in k-mer composition between samples from the two size fractions, in opposition to >3 μm samples (Fig. S2). As stated in the *Environmental metadata compilation* section, UDW samples were removed from the biogeographical investigations due to differences in environmental metadata availability. Finally, clusters showing near-zero variance abundance profiles were removed from each matrix using the pre-Process function from the Caret R Package⁶⁴. The near zero variance definition was set at a minimal threshold of 20% unique abundance values and a maximal ratio of 95–5 between the most abundant and second-most abundant values. All near-zero variance clusters had null abundance values in more than 80% of samples, so in practice this step removed rare AGC.

Identification of AGNOSTOS clusters highly linked with the environment

A random forest regression model was fitted for each cluster of the small and large size fraction matrices that passed the near zero variance threshold. Normalized coverage values were used as interest variables, and the 50 environmental variables from the CTD metadata as predictors. For each model, the number of predictors tried at each split was optimized between 5 and 8 (default being the rounded down square root of the total number of predictors), the number of trees was set at 501 and other parameters were left at default in the ranger function from the ranger R package⁸⁸. Each model went through 3 repetitions of 4-fold cross-validation using the train function from the Caret package. Variable importance, based on permutations, and adjusted cross-validation R^2 values from each selected model were retrieved. Density of R^2 values were drawn for the small and large size fraction results, separately. To estimate a threshold of R-squared at which it is unlikely that the link between coverage and metadata could be observed by chance, the same runs of random-forest models were computed on four matrices with randomly permuted rows, two of the small-size fraction matrices and two of the large ones. Based on the 95th centile value for each set of permutations, R^2 thresholds were set at 10% for the small size fraction matrix and 15% for the large one. AGNOSTOS clusters meeting these thresholds were defined as highly linked with the environment (env-AGC).

Grouping of co-abundant AGNOSTOS clusters

To further reduce the dimensionality of the two matrices of interest without removing any env-AGC, the approach described by Minot and Willis³⁴ was used to group them into groups of Co-Abundant env-AGC (CAG). This approach, based on the Approximate Nearest Neighbor heuristic, allows to cluster millions of genes/gene clusters into co-abundant groups with limited computer power and in reasonable time. The clustering Python scripts available at <https://github.com/FredHutch/find-cags> were used with default parameters independently on the small and large size fractions env-AGC matrices. CAG-level coverage matrices were created by summing coverage across all members of a CAG.

Constrained ordination and further investigation of CAGs

Redundancy analysis was fitted on the Hellinger-transformed small and large size fraction CAG-level matrices, using the vegan R package⁸⁹. Again, coverage values were used as interest variables, and environmental variables as explanatory variables. Both analyses were significant (ANOVA, p -value = 0.001 for both small and large), allowing us to go further by selecting environmental variables through a two-directional stepwise selection based on the Akaike Information Criterion (AIC). The selected models were again significant (ANOVA, p -value = 0.001 for both small and large). For each model, CAGs appearing on the extremities of axes 1–5 were individually selected to be analyzed in depth. Taxonomic annotations of genes within each CAG were retrieved through the annotation of their contigs of origin, obtained from both MMSeqs2 taxonomic annotation tool with the UniRef90 database as reference⁷³ and Kraken2⁷⁴ with the Genome Taxonomy Database (GTDB)²⁶ (see *Assembly of metagenomic short reads and the profiling of resulting contigs*). Since it took 15–20 h to annotate splits of 25,000 contigs using MMSeqs2 on 24 CPUs with 80 Gb of memory, only a selection of CAGs of interest were annotated this way, while all contigs were annotated through Kraken and GTDB. In addition, genes in CAGs of particular interest were annotated based on the presence of their contig of origin in MAGs from the ACE MAGs database (Pommellec, Faure, Noel, Cormier, Eren, Fernandez-Guerra, Vanni, Fourquez, Houssais, Da Silva, Gavery, Perdureau, Labadie, Wincker, Poulain, Hassler, Lin, Cassar, Maignien, in prep.), allowing

precise taxonomic annotations for all genes found in MAGs. To estimate a potential functional enrichment in a CAG, AGNOSTOS cluster-level EggNOG and KEGG annotations were retrieved for all members of the CAG, and compared to the rest of AGNOSTOS cluster-level annotations through a one-sided Fisher test. Obtained *p*-values were corrected for multiple testing using the Benjamini-Hochberg method, and *p*-value threshold was set to 0.01 for enrichment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The 218 metagenomic samples generated in this study have been deposited in the European Nucleotide Archive (ENA) database, under the project code PRJEB83083: <https://www.ebi.ac.uk/ena/browser/view/PRJEB83083>. A correspondence table of all samples linking their ACE unique ID, BioSample code, ENA run and experiment codes has been deposited on [Zenodo](https://zenodo.org/record/10000000). All datasets computed from the raw metagenomics data and used to produce results in this study have been deposited in [Zenodo](https://zenodo.org/record/10000000) and [Figshare](https://figshare.com). Detailed list of links to the different raw and processed data is available in Table S1. Source data used to produce all Figures are provided with this paper. Source data are provided with this paper.

Code availability

The scripts used to produce the results of this study are available at https://github.com/EmileFaure/ACE_gene_centric_scripts⁹⁰. The different steps to go from quality-filtered reads to gene-level per-base coverage and detection matrices were integrated into a Nextflow workflow available at <https://gitlab.ifremer.fr/bioinfo/workflows/noemie>. The AGNOSTOS clustering workflow is available at <https://github.com/functional-dark-side/agnostos-wf>. The CAGs clustering workflow is available here: <https://github.com/FredHutch/find-cags>.

References

- Gray, A. R. The four-dimensional carbon cycle of the Southern Ocean. *Annu. Rev. Mar. Sci.* **16**, 163–190 (2024).
- Hassler, C. S., Sinoir, M., Clementson, L. A. & Butler, E. C. V. Exploring the Link between Micronutrients and Phytoplankton in the Southern Ocean during the 2007 Austral Summer. *Front. Microbiol.* **3**, 202 (2012).
- Tagliabue, A. et al. The integral role of iron in ocean biogeochemistry. *Nature* **543**, 51–59 (2017).
- Deppeler, S. L. & Davidson, A. T. Southern Ocean phytoplankton in a changing climate. *Front. Mar. Sci.* **4**, 40 (2017).
- Hauck, J. et al. On the Southern Ocean CO₂ uptake and the role of the biological carbon pump in the 21st century. *Glob. Biogeochem. Cycles* **29**, 1451–1470 (2015).
- Christaki, U. et al. Seasonal microbial food web dynamics in contrasting Southern Ocean productivity regimes. *Limnol. Oceanogr.* **66**, 108–122 (2021).
- Landa, M., Blain, S., Christaki, U., Monchy, S. & Obernosterer, I. Shifts in bacterial community composition associated with increased carbon cycling in a mosaic of phytoplankton blooms. *ISME J.* **10**, 39–50 (2016).
- Charlson, R. J., Lovelock, J. E., Andreae, M. O. & Warren, S. G. Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature* **326**, 655–661 (1987).
- Teng, Z.-J. et al. Biogeographic traits of dimethyl sulfide and dimethylsulfoniopropionate cycling in polar oceans. *Microbiome* **9**, 207 (2021).
- Doré, H. et al. Differential global distribution of marine picocyanobacteria gene clusters reveals distinct niche-related adaptive strategies. *ISME J.* **17**, 720–732 (2023).
- Faure, E., Ayata, S.-D. & Bittner, L. Towards omics-based predictions of planktonic functional composition from environmental data. *Nat. Commun.* **12**, 4361 (2021).
- Frémont, P. et al. Restructuring of plankton genomic biogeography in the surface ocean under climate change. *Nat. Clim. Change* **12**, 393–401 (2022).
- Sunagawa, S. et al. Tara oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **18**, 428–445 (2020).
- Laiolo, E. et al. Metagenomic probing toward an atlas of the taxonomic and metabolic foundations of the global ocean genome. *Front. Sci.* **1**, 1038696 (2024).
- Paoli, L. et al. Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).
- Cao, S. et al. Structure and function of the Arctic and Antarctic marine microbiota as revealed by metagenomics. *Microbiome* **8**, 47 (2020).
- Sun, Y., Debeljak, P. & Obernosterer, I. Microbial iron and carbon metabolism as revealed by taxonomy-specific functional diversity in the Southern Ocean. *ISME J.* **15**, 2933–2946 (2021).
- Debeljak, P., Toulza, E., Beier, S., Blain, S. & Obernosterer, I. Microbial iron metabolism as revealed by gene expression profiles in contrasted Southern Ocean regimes. *Environ. Microbiol.* **21**, 2360–2374 (2019).
- Dutta, A. et al. Depth drives the distribution of microbial ecological functions in the coastal western Antarctic Peninsula. *Front. Microbiol.* **14**, 1168507 (2023).
- Landwehr, S. et al. Exploring the coupled ocean and atmosphere system with a data science approach applied to observations from the Antarctic Circumnavigation Expedition. *Earth Syst. Dyn.* **12**, 1295–1369 (2021).
- Salazar, G. et al. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell* **179**, 1068–1083.e21 (2019).
- Sunagawa, S. et al. Structure and function of the global ocean microbiome. *Science* **348**, 1261359–1261359 (2015).
- Acinas, S. G. et al. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Commun. Biol.* **4**, 1–15 (2021).
- Pavlopoulos, G. A. et al. Unraveling the functional dark matter through global metagenomics. *Nature* **622**, 594–602 (2023).
- Vanni, C. et al. Unifying the known and unknown microbial coding sequence space. *eLife* **11**, e67667 (2022).
- 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).
- Iudicone, D. et al. Water masses as a unifying framework for understanding the Southern Ocean Carbon Cycle. *Biogeosciences* **8**, 1031–1052 (2011).
- Kolody, B. C. et al. Overturning circulation structures the microbial functional seascape of the South Pacific. *Science* **389**, 176–182 (2025).
- Abell, G. C. J. & Bowman, J. P. Ecological and biogeographic relationships of class Flavobacteria in the Southern Ocean. *FEMS Microbiol. Ecol.* **51**, 265–277 (2005).
- Wilkins, D., van Sebille, E., Rintoul, S. R., Lauro, F. M. & Cavicchioli, R. Advection shapes Southern Ocean microbial assemblages independent of distance and environment effects. *Nat. Commun.* **4**, 2457 (2013).
- Li, Q., England, M. H., Hogg, A. M., Rintoul, S. R. & Morrison, A. K. Abyssal ocean overturning slowdown and warming driven by Antarctic meltwater. *Nature* **615**, 841–847 (2023).
- Delmont, T. O., Eren, A. M., Vineis, J. H. & Post, A. F. Genome reconstructions indicate the partitioning of ecological functions inside a phytoplankton bloom in the Amundsen Sea, Antarctica. *Front. Microbiol.* **6**, 1090 (2015).

33. Roda-Garcia, J. J., Haro-Moreno, J. M. & López-Pérez, M. Evolutionary pathways for deep-sea adaptation in marine planktonic Actinobacteriota. *Front. Microbiol.* **14**, 1159270 (2023).
34. Minot, S. S. & Willis, A. D. Clustering co-abundant genes identifies components of the gut microbiome that are reproducibly associated with colorectal cancer and inflammatory bowel disease. *Microbiome* **7**, 110 (2019).
35. Mock, T. et al. Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. *Nature* **541**, 536–540 (2017).
36. Bertrand, E. M. et al. Phytoplankton–bacterial interactions mediate micronutrient colimitation at the coastal Antarctic sea ice edge. *Proc. Natl. Acad. Sci. USA* **112**, 9938–9943 (2015).
37. Xue, C. et al. Polysaccharide utilization by a marine heterotrophic bacterium from the SAR92 clade. *FEMS Microbiol. Ecol.* **97**, fiab120 (2021).
38. Kappelmann, L. (*Meta*-)genomic Analysis of the Diversity and the Carbohydrate Degradation Potential of the SAR92 Clade during a Diatom-induced Bacterioplankton Bloom. Doctoral dissertation, University of Bremen, Bremen/Germany (2013).
39. Kim, S.-J. et al. Genomic and metatranscriptomic analyses of carbon remineralization in an Antarctic polynya. *Microbiome* **7**, 29 (2019).
40. Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* **42**, 1303–1312 (2024).
41. Gaïa, M. et al. Mirusviruses link herpesviruses to giant viruses. *Nature* **616**, 783–789 (2023).
42. Holst, F. et al. Helixer: ab initio prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. *Nat. Methods*. <https://doi.org/10.1038/s41592-025-02939-1> (2025).
43. Ibarbalz, F. M. et al. Global trends in marine plankton diversity across kingdoms of life. *Cell* **179**, 1084–1097.e21 (2019).
44. Sussfeld, D. et al. New groups of highly divergent proteins in families as old as cellular life with important biological functions in the ocean. *Environ. Microbiol.* **20**, 65 (2025).
45. van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).
46. Hu, M. et al. Evaluation of large language models for discovery of gene set function. *Nat. Methods* **22**, 82–91 (2025).
47. Mizuno, C. M., Rodriguez-Valera, F. & Ghai, R. Genomes of planktonic acidimicrobiales: widening horizons for marine actinobacteria by metagenomics. *mBio* **6**, 10–1128 (2015).
48. Liu, J. et al. Oceanospirillales containing the DMSP lyase DddD are key utilizers of carbon from DMSP in coastal seawater. *Microbiome* **10**, 110 (2022).
49. Ye, N. et al. The role of zinc in the adaptive evolution of polar phytoplankton. *Nat. Ecol. Evol.* **6**, 965–978 (2022).
50. Meng, L. et al. Genomic adaptation of giant viruses in polar oceans. *Nat. Commun.* **14**, 6233 (2023).
51. Wilkins, D. et al. Biogeographic partitioning of Southern Ocean microorganisms revealed by metagenomics. *Environ. Microbiol.* **15**, 1318–1333 (2013).
52. Dinasquet, J., Landa, M. & Obernosterer, I. SAR11 clade microdiversity and activity during the early spring blooms off Kerguelen Island, Southern Ocean. *Environ. Microbiol. Rep.* **14**, 907–916 (2022).
53. Kraemer, S., Ramachandran, A., Colatriano, D., Lovejoy, C. & Walsh, D. A. Diversity and biogeography of SAR11 bacteria from the Arctic Ocean. *ISME J.* **14**, 79–90 (2020).
54. Noell, S. E. & Giovannoni, S. J. SAR11 bacteria have a high affinity and multifunctional glycine betaine transporter. *Environ. Microbiol.* **21**, 2559–2575 (2019).
55. Rodionov, D. A., Hebbeln, P., Gelfand, M. S. & Eitinger, T. Comparative and functional genomic analysis of prokaryotic nickel and cobalt uptake transporters: evidence for a novel group of ATP-binding cassette transporters. *J. Bacteriol.* **188**, 317–327 (2006).
56. Manck, L. E., Espinoza, J. L., Dupont, C. L. & Barbeau, K. A. Transcriptomic study of substrate-specific transport mechanisms for iron and carbon in the marine copiotroph *Alteromonas macleodii*. *mSystems* **5**, e00070-20 (2020).
57. da Silva Neto, J. F., Lourenço, R. F. & Marques, M. V. Global transcriptional response of *Caulobacter crescentus* to iron availability. *BMC Genom.* **14**, 549 (2013).
58. Huo, C. et al. The transcriptomic and bioinformatic characterizations of iron acquisition and heme utilization in avibacterium *Paragallinarum* in response to iron-starvation. *Front. Microbiol.* **12**, 610196 (2021).
59. Piedade, G. J. et al. Seasonal dynamics and diversity of Antarctic marine viruses reveal a novel viral seascape. *Nat. Commun.* **15**, 9192 (2024).
60. Smith, R. C. et al. The Palmer LTER: a long-term ecological research program at Palmer Station, Antarctica. *Oceanography* **8**, 77–86 (1995).
61. Venables, H. et al. Sustained year-round oceanographic measurements from Rothera Research Station, Antarctica, 1997–2017. *Sci. Data* **10**, 265 (2023).
62. Alberti, A. et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci. Data* **4**, 170093 (2017).
63. Ser-Giacomi, E. et al. Lagrangian betweenness as a measure of bottlenecks in dynamical systems with oceanographic examples. *Nat. Commun.* **12**, 4935 (2021).
64. Kuhn, M. Building predictive models in R using the caret Package. *J. Stat. Softw.* **28**, 1–26 (2008).
65. Minoche, A. E., Dohm, J. C. & Himmelbauer, H. & others. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.* **12**, R112 (2011).
66. Eren, A. M., Vineis, J. H., Morrison, H. G. & Sogin, M. L. A filtering method to generate high quality short reads using illumina paired-end technology. *PLoS ONE* **8**, e66643 (2013).
67. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
68. Eren, A. M. et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
69. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 1 (2010).
70. Gruber-Vodicka, H. R., Seah, B. K. B. & Priesse, E. phyloFlash: rapid small-subunit rRNA profiling and targeted assembly from metagenomes. *mSystems* **5**, e00920-20 (2020).
71. West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C. & Banfield, J. F. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.* **28**, 569–580 (2018).
72. Pronk, L. J. U. & Medema, M. H. Whokaryote: distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. *Microb. Genom.* **8**, 000823 (2022).
73. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
74. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
75. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
76. Delmont, T. O. et al. Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genom.* **2**, 100123 (2022).
77. Jamy, M. et al. Identification of a deep-branching lineage of algae using environmental plastid genomes. *Nat. Commun.* **17**, 662 (2026).

78. Mihara, T. et al. Linking virus genomes with host taxonomy. *Viruses* **8**, 66 (2016).
79. Freil, K. C. et al. New SAR11 isolate genomes and global marine metagenomes resolve ecologically relevant units within the Pelagibacterales. *Nat. Commun.* **17**, 328 (2026).
80. Eren, A. M. et al. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* **6**, 3–6 (2021).
81. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
82. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinform. Oxf. Engl.* **22**, 1658–1659 (2006).
83. Huerta-Cepas, J. et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–D293 (2016).
84. Aramaki, T. et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
85. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
86. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
87. Lloréns-Rico, V., Vieira-Silva, S., Gonçalves, P. J., Falony, G. & Raes, J. Benchmarking microbiome transformations favors experimental quantitative approaches to address compositionality and sampling depth biases. *Nat. Commun.* **12**, 3562 (2021).
88. Wright, M. N. & Ziegler, A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77**, 1–17 (2017).
89. Oksanen, J. et al. vegan: Community Ecology Package. R package version 2.4–3. (2017).
90. Faure, E. & Delpech, L. M. Water mass specific genes dominate the Southern Ocean microbiome. ACE gene centric scripts. <https://doi.org/10.5281/zenodo.17961605> (2025).
91. NOAA National Centers for Environmental Information. ETOPO 2022 15 Arc-Second Global Relief Model. NOAA National Centers for Environmental Information <https://doi.org/10.25921/FD45-GT74> (2022).
92. Vihtakari, M. *ggOceanMaps: Plot Data on Oceanographic Maps Using 'Ggplot2'*. (2022).
93. Henry, T. et al. Physical and biogeochemical oceanography data from Conductivity, Temperature, Depth (CTD) rosette deployments during the Antarctic Circumnavigation Expedition (ACE). Zenodo <https://doi.org/10.5281/zenodo.3813646> (2020).

Acknowledgements

The Antarctic Circumnavigation Expedition was funded by the Swiss Polar Institute and Ferring Pharmaceuticals. This work is dedicated to the memory of the late David Walton, Chief Scientist of the ACE cruise. The authors thank all onboard expedition members for their support. This study was founded through an Agence Nationale de la Recherche grant 18-CE02-0024 to LM, and supported by France Génomique (ANR-10-INBS-09-08). C.H. was supported by the Swiss National Science Foundation (PPOOP2_166197) and is currently supported by the Ferring Pharmaceutical—Margaretha Kamprad Chair in environmental sciences attributed to Prof. J. Chappelaz. N.-C. was supported by the “Laboratoire d’Excellence” LabexMER (ANR-10-LABX-19) and cofunded by a grant

from the French government under the program “Investissements d’Avenir”. The authors thank Alberto Baudena from the Consiglio Nazionale Delle Ricerche, Istituto di Scienze Marine (CNR-ISMAR), Lerici (SP), Italy, who provided the Lagrangian data, and Stéphane Pesant from the European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge (UK), who helped preparing the public release of the data.

Author contributions

L.M., N.C., and Y.L. conceived and planned the sampling. L.M., N.C., Y.L., M.F., M.N.H., and C.H. collected the samples. C.H., N.C., Y.L., L.M., E.F., L.M.D., and M.N.H. produced, gathered, and organized the environmental metadata. C.S., F.G., A.P., K.L., P.W., and J.P. were in charge of the molecular work and sequencing. E.F., L.M., J.P., C.N., and A.C. developed the bioinformatics pipeline to create and annotate the gene matrices, with help and inputs from AME. A.F.G. and C.V. ran the AGNOSTOS clustering. UG took care of giant virus-specific annotations. E.F. developed the data analysis pipeline and produced all figures and results, under the supervision of L.M., with help from L.M.D. and J.P. E.F. wrote the manuscript, with contributions from L.M., N.C., Y.L., A.M.E., C.H., U.G., M.N.H., M.F., A.F.G., L.M.D., and J.P. All co-authors validated the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-026-69584-w>.

Correspondence and requests for materials should be addressed to Emile Faure, Nicolas Cassar or Loïs. Maignien.

Peer review information *Nature Communications* thanks Shane Hogle and the other anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026

¹Ifremer, BEEP, Univ Brest, Plouzané, France. ²Centre National de la Recherche Scientifique, UMR 7144 Adaptation and Diversity in the Marine Environment, Station Biologique, Sorbonne Université, Roscoff, France. ³IFREMER, IRSI – Service de Bioinformatique (SeBiMER) Plouzané, Plouzané, France. ⁴Littoral Environnement et Sociétés (LIENSs), UMRi 7266 CNRS - La Rochelle Université, La Rochelle, France. ⁵Department of Biology, École Normale Supérieure de Lyon, Lyon, France. ⁶Helmholtz Institute for Functional Marine Biodiversity, Oldenburg, Germany. ⁷Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany. ⁸Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, Oldenburg, Germany. ⁹Marine 'Omics Bridging Group, Max Planck Institute for Marine Microbiology, Bremen, Germany. ¹⁰Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA, USA. ¹¹Centre for Ancient Environmental Genomics, Globe Institute, University of Copenhagen, Copenhagen, Denmark. ¹²Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark. ¹³MARUM Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. ¹⁴Australian Antarctic Program Partnership, Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, TAS, Australia. ¹⁵Laboratoire d'Océanographie et du Climat (LOCEAN), CNRS-Sorbonne Université, Paris, France. ¹⁶Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France. ¹⁷Genoscope, Institut François Jacob, CEA, Université Paris-Saclay, Evry, France. ¹⁸Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GO-SEE, CNRS, Paris, France. ¹⁹Swiss Polar Institute, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. ²⁰School of Architecture, Civil, and environmental engineering, Smart Environmental Sensing in Extreme Environments, ALPOLE, Ecole Polytechnique Fédérale de Lausanne, Sion, Switzerland. ²¹Institute of Earth Sciences, University of Lausanne, Lausanne, Switzerland. ²²Department of Life Sciences, Texas A&M University – Corpus Christi, Corpus Christi, TX, USA. ²³Division of Earth and Climate Sciences, Nicholas School of the Environment, Duke University, Durham, NC, USA. ²⁴CNRS, IRD, Ifremer, LEMAR, Université de Brest, Plouzané, France. ✉ e-mail: emile.faure@sb-roscoff.fr; Nicolas.Cassar@duke.edu; lois.maignien@univ-brest.fr