Water mass specific genes dominate the Southern Ocean microbiome

Emile Faure

emile.faure@univ-brest.fr

Station Biologique de Roscoff, Sorbonne Université

Jolann Pommellec

Université de Bretagne Occidentale

Cyril Noel

IFREMER

Alexandre Cormier

IFREMER

Lisa-Marie Delpech

Université de La Rochelle https://orcid.org/0009-0006-0154-0211

Murat Eren

Helmholtz institute for functional marine biodiversity

Antonio Fernandez-Guerra

Lundbeck GeoGenetics Centre, The Globe Institute, University of Copenhagen https://orcid.org/0000-0002-8679-490X

Chiara Vanni

Max Planck Institute for Marine Microbiology https://orcid.org/0000-0002-1124-1147

Marion Fourquez

MIO, Aix Marseille Université

Marie-Noëlle Houssais

Sorbonne Université

Corinne Da Silva

Commissariat à l'Energie Atomique

Frederick Gavory

GENOSCOPE, CEA

Aude Perdereau

GENOSCOPE, CEA

Karine Labadie

GENOSCOPE, CEA

Patrick Wincker

Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université d’Evry, Université Paris-Saclay, https://orcid.org/0000-0001-7562-3454

Julie Poulain

Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France

Christel Hassler

Ecole Polytechnique Fédérale de Lausanne

Yajuan Lin

Texas A&M Corpus Christi https://orcid.org/0000-0002-9057-9321

Nicolas Cassar

Duke University

Lois Maignien

Department of Biology, University of Southern Denmark, 5230 Odense M, Denmark

Article

Keywords:

Posted Date: January 3rd, 2025

DOI: https://doi.org/10.21203/rs.3.rs-5608865/v1

Additional Declarations: There is **NO** Competing Interest.

Water mass specific genes dominate the Southern Ocean microbiome

Emile Faure[1,2], Jolann Pommellec[1], Cyril Noel[3], Alexandre Cormier[3], Lisa-Marie Delpech[1,4,5], Murat A Eren[6,7,8,9,10], Antonio Fernandez-Guerra[11,12], Chiara Vanni[13], Marion Fourquez[14], Marie-Noëlle Houssais[15], Corinne da Silva[16], Frederick Gavory[16], Aude Perdereau[16], Karine Labadie[16], Patrick Wincker[17,18], Julie Poulain[17,18], Christel Hassler[19,20,21], Yajuan Lin[22,23,24], Nicolas Cassar[23,24*], Loïs Maignien[1,10*]*

1 - Univ Brest (UBO), CNRS, IFREMER, Laboratoire de Microbiologie des Environnements Extrêmes, Plouzané, 29280, France

2 - Sorbonne Université, Centre National de la Recherche Scientifique, UMR 7144 Adaptation and Diversity in the Marine Environment, Station Biologique, Roscoff 29680, France

3 - IFREMER, IRSI – Service de Bioinformatique (SeBiMER) Plouzané, 29280 France

4 - Littoral Environnement et Sociétés (LIENSs), UMRi 7266 CNRS - La Rochelle Université, La Rochelle 17000, France

5 - Department of Biology, École Normale Supérieure de Lyon, Lyon 69007, France

6 - Helmholtz Institute for Functional Marine Biodiversity, 26129 Oldenburg, Germany

7 - Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, 27570 Bremerhaven, Germany

8 - Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, 26129 Oldenburg, Germany

9 - Marine 'Omics Bridging Group, Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany

10 - Bay Paul Center, Marine Biological Laboratory, Woods Hole, 02543 MA, USA

11 - Centre for Ancient Environmental Genomics, Globe Institute, University of Copenhagen, Copenhagen, Denmark

12 - Lundbeck Foundation GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark

13 - MARUM Center for Marine Environmental Sciences, University of Bremen,28359 Bremen, Germany

14 - Aix Marseille Univ., Université de Toulon, CNRS, IRD, MIO UMR 110, Marseille 13288, France.

15 - Laboratoire d'Océanographie et du Climat (LOCEAN), CNRS-Sorbonne Université, Paris, France

16 - Genoscope, Institut François Jacob, CEA, Université Paris-Saclay, Evry, 91057, France.

17 - Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, 91057, France.

18 - Research Federation for the Study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GO-SEE, CNRS, Paris, France.

19 - Swiss Polar Institute, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

20 - School of Architecture, Civil, and environmental engineering, Smart Environmental Sensing in Extreme Environments, ALPOLE, Ecole Polytechnique Fédérale de Lausanne, Sion, Switzerland.

21 - Institute of Earth Sciences, University of Lausanne, Lausanne, Switzerland.

22 - Department of Life Sciences, Texas A&M University – Corpus Christi, 78412 Corpus Christi, Texas, USA

23 - Division of Earth and Climate Sciences, Nicholas School of the Environment, Duke University, 27710 Durham, North Carolina, USA

24 - CNRS, Université de Brest, IRD, Ifremer, LEMAR, Plouzané, France

Corresponding authors: emile.faure@sb-roscoff.fr; lois.maignien@univ-brest.fr; Nicolas.Cassar@duke.edu

ABSTRACT

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 comprehensive SO 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 adaptations to organic matter consumption in the blooming Mertz polynya and temperature-dependent trace metal utilization by the ubiquist Bacteria Pelagibacter. Our work takes a step towards a more comprehensive understanding of SO's plankton ecology and evolution, capturing the current state of the unique microbial diversity in this rapidly changing Ocean.

INTRODUCTION

The Southern Ocean (SO) dominates other oceans in heat and carbon uptakes while being particularly exposed to climate change impacts[1]. It is mainly composed of high nutrient low chlorophyll waters (HNLC) where phytoplanktonic growth is limited by trace elements such as iron or manganese[2,3]. In presence of these elements, phytoplankton blooms can reach concentrations of 10^8 cells per liter[4], playing a key role in carbon sequestration through the biological pump[5]. 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]. 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 driving effect of environmental conditions on planktonic communities[8–11]. However, only two sampling locations in the SO were included in recent efforts to compute global genes and genomes catalogs[12,13], 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 first polar gene catalog in 2020, showing the high prevalence of polar specific genes in the SO[14]. 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 changes[1].

The Antarctic Circumnavigation Expedition (ACE) circumnavigated the Southern Ocean during the 2016-2017 austral summer, producing an unprecedented amount of physical, chemical and biological observations[15]. Analyzing 218 metagenomes, we increase by an order of magnitude the number of SO samples ever considered in a meta-omics study to present the first SO-specific gene catalog (Fig. 1). Building on the seminal work of previous global metagenomics efforts[16–18], we first demonstrate the broadscale uniqueness of the SO compared to other oceans, before diving into its regional variability. To exemplify the uniqueness of biomes in the SO, we focus on the genomic signature of specialist species occurring in the Mertz polynya, before using SAR11 as a case study of genomic adaptations to polar conditions in a ubiquitous taxon.

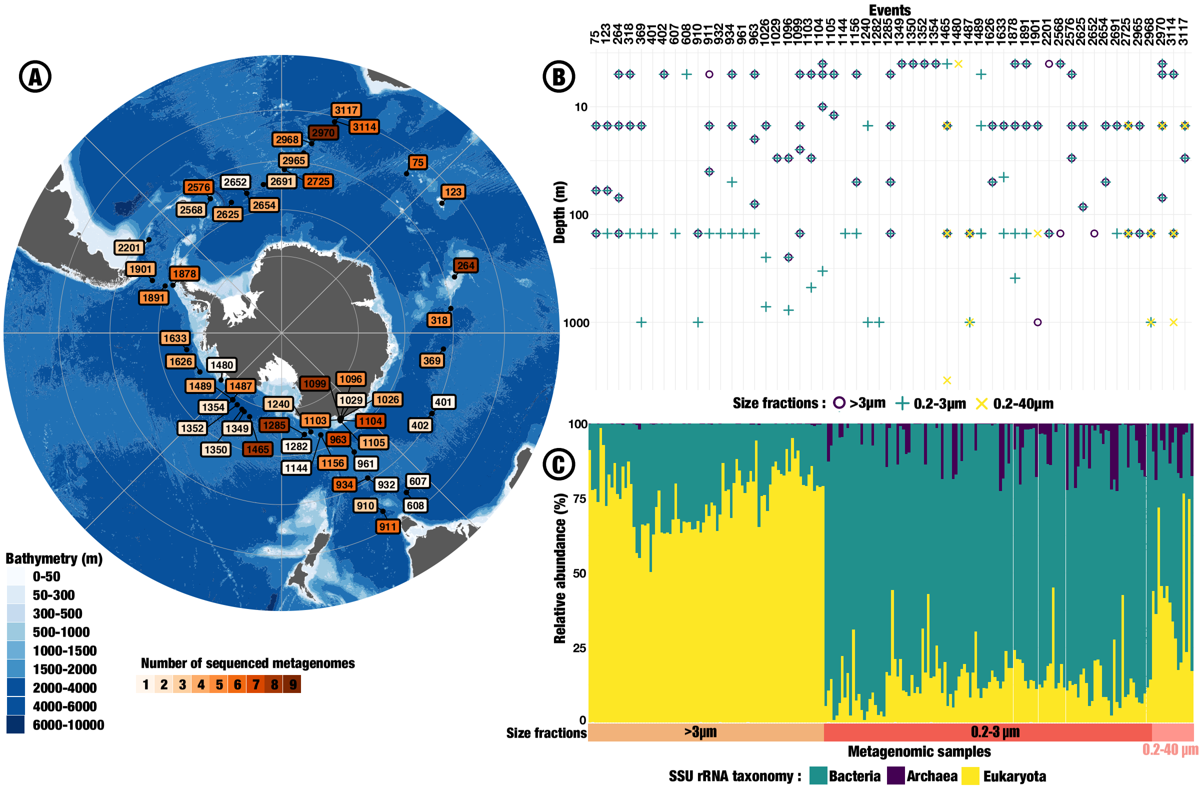

Figure 1: Overview of the ACE metagenomics dataset. (A) Map of CTD downcast events on which metagenomics samples were taken, colored according to the number of samples taken on the cast. (B) Depth and size fraction chart of all ACE metagenomes. Each vertical line corresponds to an event as pictured on 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 reconstruction from metagenomics short reads. Samples were separated according to their size fraction on the x axis, as indicated on the bottom layer.

RESULTS

Broadscale novelty of Southern Ocean's microbial genes

An unsuspected genomic diversity at SO scale

Individual assemblies of the 218 metagenomes (Fig.1, Fig.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 catalog[16] (OM-RGC-v2) at thresholds of 30% similarity and 80% coverage in amino acid sequence (Fig. 2A). This number remained at 37.9% accounting only for ACE unigenes assembled from the 0.2-3µm size fraction in the sunlit layer (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.

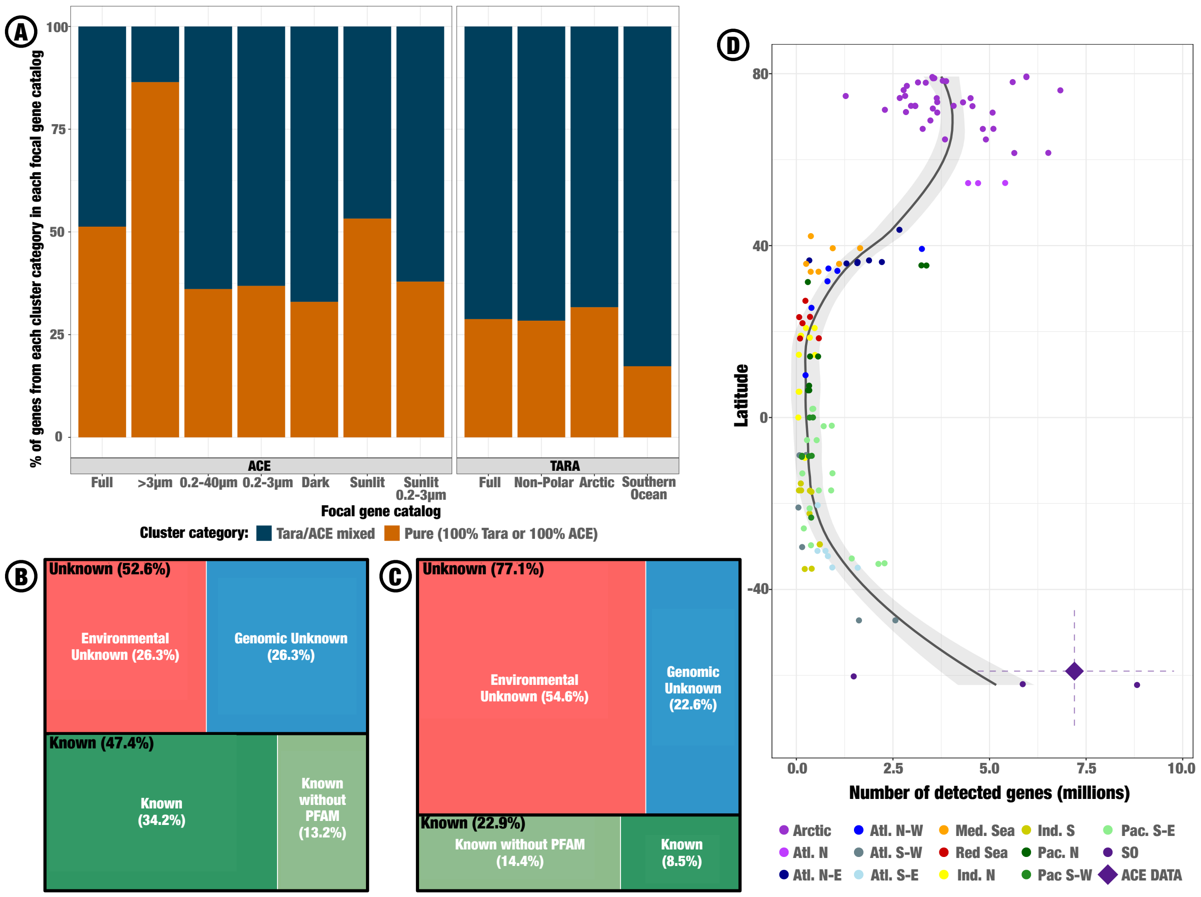

Figure 2: Novelty of Southern Ocean microbial genes. (A) Distribution of genes from either the ACE unigene catalog (left box) or the OM-RGC v2 (right box) into pure or mixed gene clusters. Genes from both catalogs were clustered at 30% similarity and 80% coverage thresholds in amino acid sequences, then 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 1X 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 was fitted to the number of detected genes in Tara samples, not taking into account ACE samples.*

We further explored distant gene homology using AGNOSTOS[19]. 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) 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,2C). The asymptotic nature of collector curves drawn at AGC level suggests that the ACE AGC catalog covers most of SO genomic diversity (Figure S2).

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 isolation[14]. To quantify this bipolar pattern among ACE genes, we mapped 134 *Tara Oceans* (TO) and *Tara Polar Circle* (TPC) metagenomes covering most subtropical and arctic oceanic regions onto ACE contigs. The resulting detection matrix shows 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, while it was of 3,956,716 in TPC samples, illustrating the high level of endemicity of the SO despite similarities in gene content between poles. This mean dropped to 334,896 ORFs for non-polar TO 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 TO 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 4,314 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).

The SO hosts a diversity of unique microbial biomes shaped by oceanographic fronts and phytoplankton blooms

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 free-living (0.2-40 + 0.2-3 µm) and >3µm size fractions considering their different taxonomic profiles (Figure 1C). We focused exclusively on AGCs with non-repeated coverage values in at least 20% of samples (1,906,624 and 2,437,988 clusters in the free-living and >3µm size fractions, resp.), avoiding rare AGCs as well as AGCs with uniform distribution across samples.

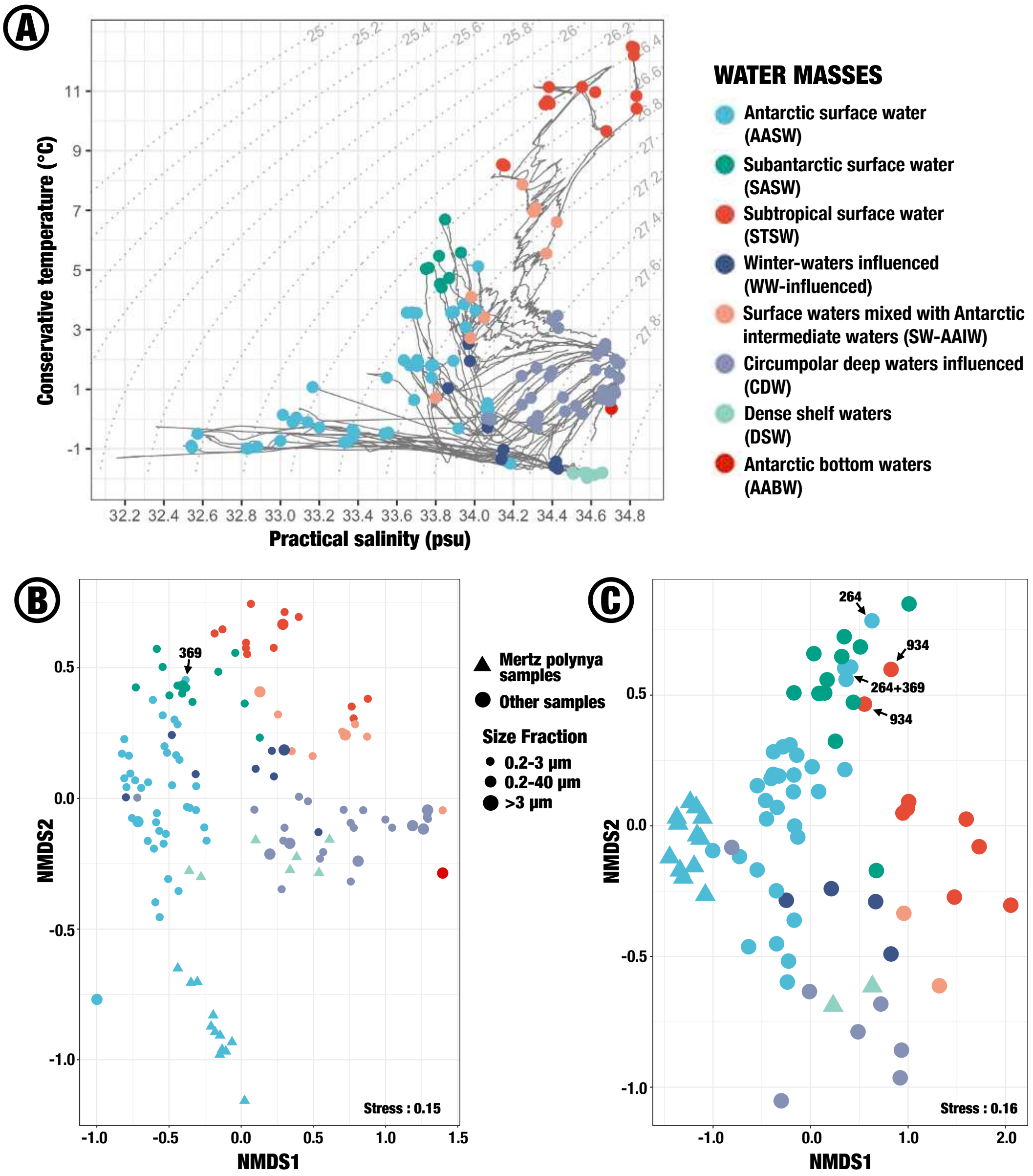

Figure 3: Microbial genes assemblages of the Southern Ocean are water mass specific. (A) Temperature – salinity diagram based on ACE downcast CTD data[20]. Each grey line corresponds to a CTD cast. Dots correspond to depths at which seawater was sampled for metagenomic libraries construction, colored according to their attributed water masses. Dotted lines in the background correspond to isopycnals. (B,C) NMDS computed on AGC abundance matrices of free-living (B) and >3µm (C) size fractions, colored according to their water masses using the same color legend as in (A). Positions of samples in B and C are only determined by their composition in AGC. The event numbers of samples taken above 150m 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

The classification of our samples in water masses based on temperature-salinity-oxygen diagrams was the best grouping variable for predicting AGC abundance (Figure 3, S3). AGC assemblages were significantly distinct across water masses in the both size fractions (Figure 3B, PERMANOVA with 999 permutations, p-value<0.001). Sub-Antarctic surface waters (SASW) samples were separated from both Antarctic surface waters (AASW) samples and sub-tropical surface waters (STSW) samples, suggesting biogeographical barriers at both the Sub-Antarctic Front and the Polar Front. This was already observed for Flavobacteria[21] but lacked confirmation on broader taxonomic range[22]. All surface samples showing potential mismatches between their AGC assemblage and their attributed water masses came from events located on oceanographic fronts (Figure 3B,3C), suggesting a potential mixing of microbial assemblages at these fronts.

Samples from Circumpolar Deep Waters (CDW) 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 CDW and surface waters. Samples from Dense Shelf Waters (DSW) were mostly similar to CDW samples, except the two shallowest DSW samples which appeared closer to AASW in genomic composition for the free-living size fraction (Figure 3B). These results suggest a diminution of AGC diversity in deep water masses (see Supplementary Results). Still, the only Antarctic Bottom Water sample (AABW, 3460m depth) had the most extreme coordinate on the NMDS first axis (Figure 3B), suggesting a unique genomic composition. In light of this uniqueness and considering the projected decrease in AABW formation due to increasing influence of meltwater[23] from Antarctica, a better characterization of the functional roles from AABW microbial populations is urgently needed.

Identifying genomic markers following environmental gradients 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 free-living, 15% in the attached size fraction, Figure S4). 89.0% (resp. 82.1%) of the considered AGC were env-AGC in the free-living (resp. >3µm) 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[9,10]. To analyze env-AGC in a multivariate context, we grouped them into 156,671 and 28,756 co-abundant groups (CAGs)[24] in the free-living and >3µm size fractions, respectively. We then identified CAGs of interest associated with various biomes through a redundancy analysis (Figure 4). 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).

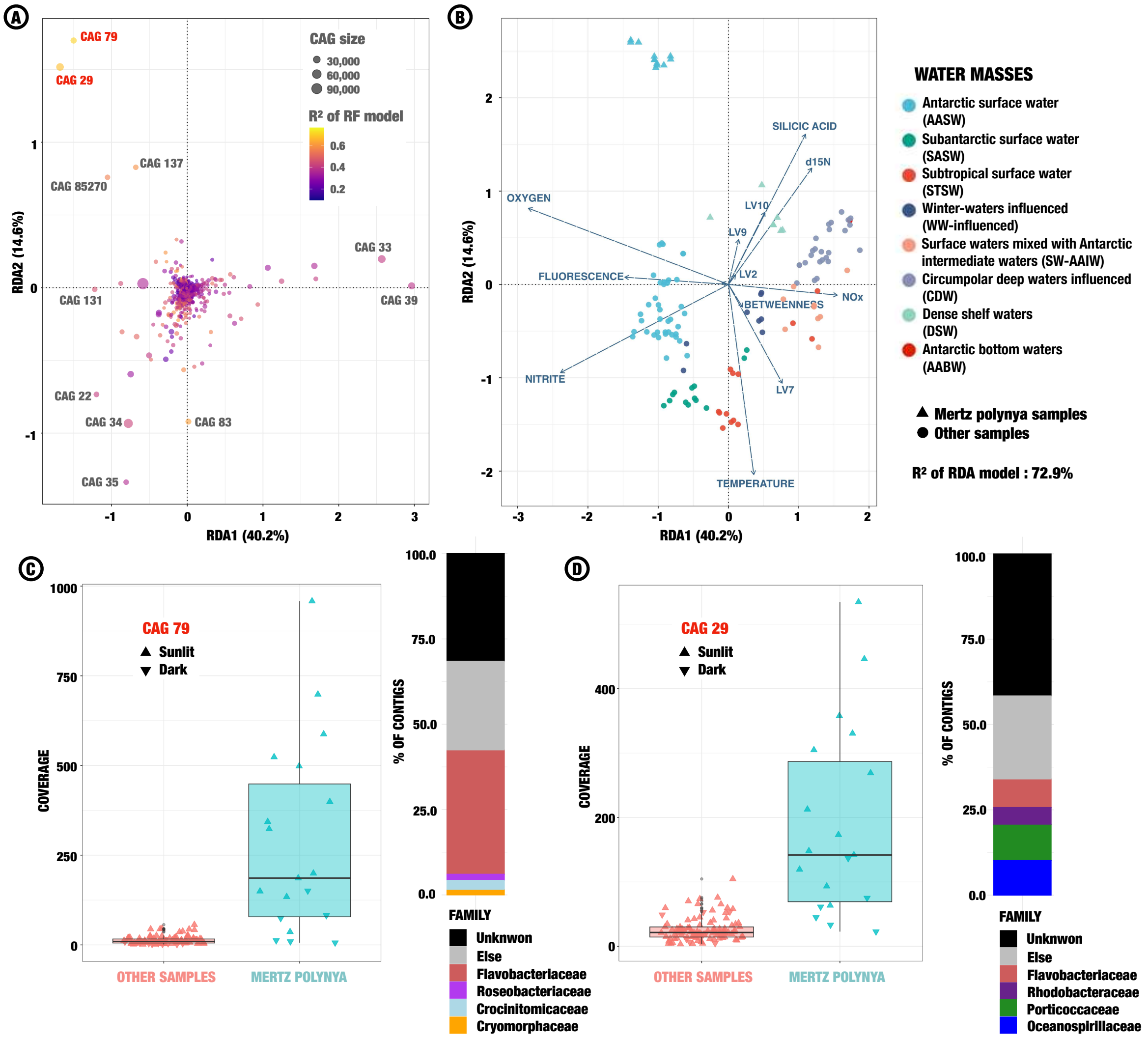

Figure 4: The response of co-abundant groups of env-AGCs (CAGs) to environmental gradients at SO scale highlights Mertz polynya's originality. Redundancy analysis of CAGs abundance in response to environmental variables, in the free-living size fraction (A,B). RDA triplot was separated in two parts for better readability, (A) showing the distribution of CAGs in the RDA space, colored according to their mean random forest R-squared (reflecting the predictability of their abundance using environmental data). The size of each dot corresponds to the size of the CAG, in number of env-AGC. The different CAGs of interest mentioned in this study are indicated with grey labels, while the two CAGs plotted in (C) and (D) are highlighted in red. (B) shows samples and environmental variables distribution in the same RDA space. Samples are colored according to their water mass. The first axis of the RDA opposed surface samples with high fluorescence and oxygen (RDA1<0) from deep samples showing high NOx concentrations (RDA1>0). The second axis was driven by temperature, opposing warm STSW samples (RDA2<0) from colder samples, and isolating all AASW samples from Mertz as an outlier group (RDA2>0). LV stands here for latent variables, corresponding to the ones described in Landwehr et al.[15]. LV2 is linked with cloud condensation, LV7 with seasonal signal, LV9 is linked to marginal sea ice zone and snowfall and LV10 to the dial cycle. A similar RDA triplot for the >3µm size fraction is presented in Figure S5. Mertz polynya's originality is further illustrated in (C) and (D), showing the abundance and taxonomy of the two CAGs most linked to it. Boxplots corresponding to each CAG's coverage are plotted at Mertz versus *in other samples, with each individual sample plotted as points shaped according to categorical depth: sunlit (150m and above) and dark (below 150m). Family-level taxonomic profiles are represented next to each boxplot, as estimated through contigs taxonomic*

annotations using UniRef90 as reference. For each CAG, only the four most abundant families are colored, the rest being aggregated as "Else".

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

Four CAGs from the free-living size fraction were enriched in Mertz samples (RDA2 > 0.5, Figure 4A): CAG 79, CAG 29, CAG 137 and CAG 85270 (Figure 4C,4D). They contained below 16% of *environmental unknowns (EU)* and more than 50% of *known (K)* and *known without PFAM (KWP)*, when 54.6% of all AGCs were annotated as EU. 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 the 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 metallo-protein were enriched in all four CAGs as well, including heme-binding proteins, M6 family metalloprotease or metal-dependent hydrolase. Finally, the four CAGs were enriched in phage integrase, and three out of four were enriched in phage plasmid primase P4, suggesting a strong phage presence in the Mertz polynya. We provide a complete list of enriched functions in each CAG of interest (Table S1).

Taxonomic profiles of all CAGs of interest were estimated based on contig-level taxonomic annotation using the UniRef90 database[25] as reference (Figure 4C,4D). To increase taxonomic precision, we also investigated annotations obtained on MAGs binned from the same metagenomic assemblies of ACE data (Pommellec et al., *in prep.*). ORFs from CAGs 79 and 85270 were mainly affiliated with the *Polaribacter* genus in both MAG-based and contigs-based annotations. In CAG 137, Rhodobacteraceae dominated both annotation types but Roseobacteraceae were absent from MAGs while abundant in contig-based annotations. In CAG 29, HTCC2207 SAR 92 was the most represented genus using MAGs-based annotation, matching the Porticoccaceae dominance in contigs-based annotations (Figure 4D). SAR 92 is a widely distributed oligotrophic clade known for its ability to consume polysaccharides in the epipelagic zone, notably through TonB-dependent receptors[26]. It has been associated with late stages of diatom-induced bacterioplankton blooms in the North Sea, uptaking and degrading specific polysaccharides including chrysolaminarin[27]. *Polaribacter* and SAR 92 have both been associated with Phaeocystis-produced chrysolaminarin degradation in the SO[28]. The second most represented genus in CAG 29 was ASP10-02a, which was identified as the main cobalamin (Vitamin B12) producer in a coastal area of the SO, playing a key role in primary production co-limitation by micronutrients[29].

Two CAGs from the >3µm size fraction were enriched in the Mertz polynya (Figure S5, discussed in Supplementary results). Both were linked to *Fragilariopsis cylindrus*, an indicator species of cold water evolutionarily adapted to the polar environment[30]. All CAGs enriched in Mertz polynya samples were thus linked with organisms specifically adapted to polar blooms conditions.

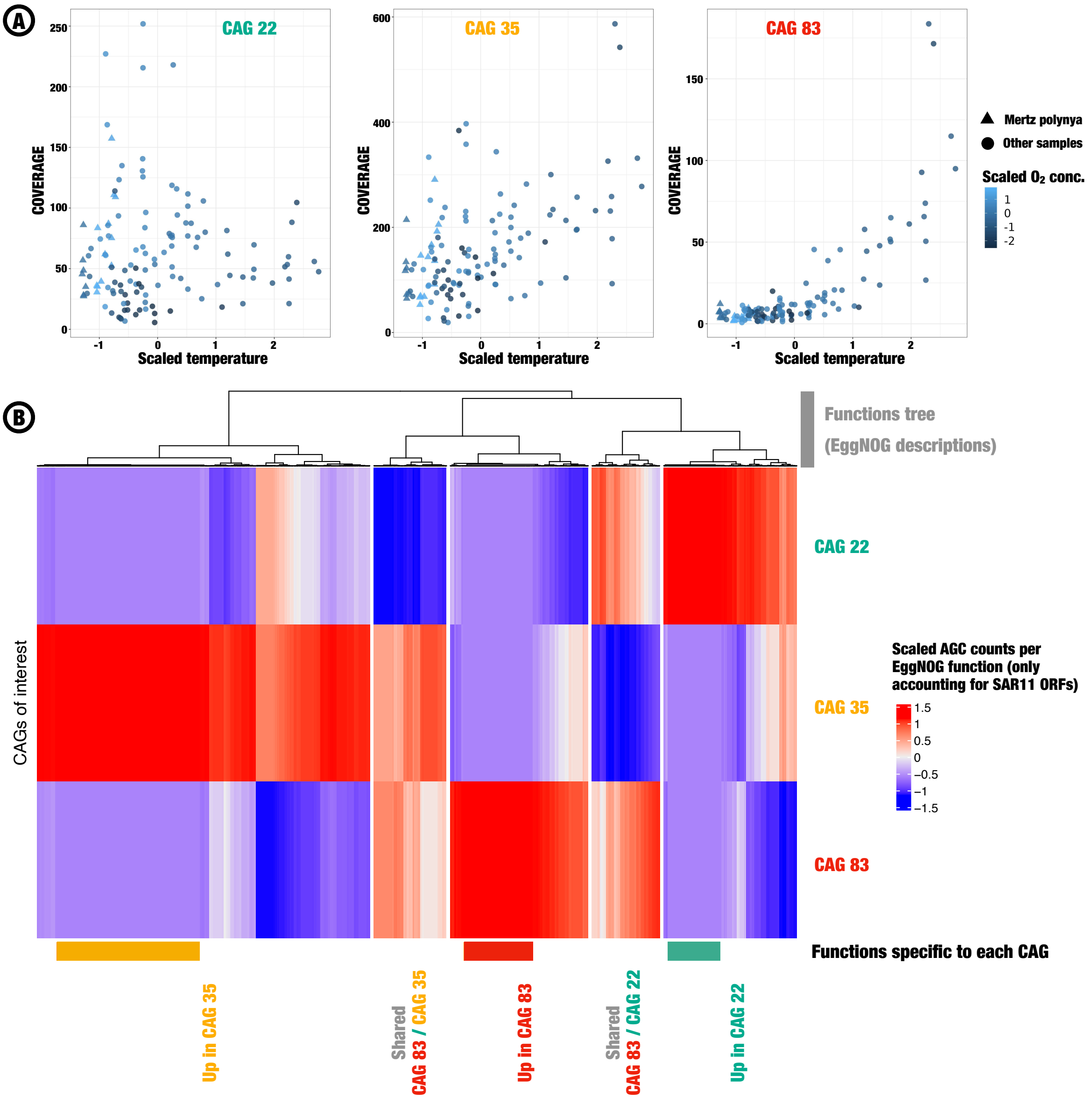

Figure 5: Functional changes across latitudes in SAR11-related CAGs. (A) Abundance of three CAGs dominated by Candidatus Pelagibacter *as a function of scaled temperature and oxygen. (B) Heatmap of scaled counts of AGCs per unique EggNOG functional annotations, taking only into account ORFs coming from contigs annotated as Candidatus* Pelagibacter*. The tree on top of the heatmap clusters each unique function according to its AGC count profile across CAGs, using euclidean distance and ward's D2 clustering method. The tree was manually cut to form 5 groups, differentiating functions shared by multiple CAGs from functions being more abundant in one CAG, as described by text labels below the heatmap. Functions found only in one of the three CAGs are highlighted with colored bars on the bottom layer of the heatmap.*

Global latitudinal shifts across gene groups illustrate adaptation to polar conditions in the ubiquitous SAR11

We identified three CAGs of interest with a distinct response to environmental data despite having similar taxonomic profiles (Figure 4A, 5). CAG 83 was specific to warm waters, with low abundance in all AASW samples supposing a latitudinal boundary at the polar front (Figure 5A). 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 AASW waters. CAGs 22, 35 and 83 were all dominated by Pelagibacteraceae, 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). To compare the functional annotations of SAR11 genes across the three CAGs, we extracted ORFs coming from SAR11 contigs and compared EggNOG annotations between them at AGC-level (Figure 5). CAGs 22 and 83 showed opposite functional profiles, *i.e.* functions highly present in one were rare or absent in the other, while CAG 35 shared functions with the two other CAGs, matching its ubiquitous distribution (Figure 5B). A total of 430 unique functions showed an increased presence in CAG22, among which 170 were only found in CAG22 (Figure 5B). The most observed was the transmembrane NikM subunit, transporting nickel, which was carried by SAR11 ORFs in 10 AGCs of CAG22 and none of CAG83 and CAG35. Among the other functions with increased detection in CAG22, many were related to trace metals (zinc, iron), sulfur (e.g. nucleotide-disulfide oxidoreductase, Sulfotransferase) and phosphorus cycles (e.g. polyphosphate pyrophosphohydrolases, metal-dependent phosphohydrolase). See Table S1 to access the complete list.

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. We illustrated how microbial gene assemblages from the SO are largely endemic and unknown, both at taxonomic and functional level. 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 >3µm size fraction, which should be treated with caution due to the difficulty of detecting good quality ORFs from eukaryotic contigs[31] (*cf* Methods). Yet, 53.3% of all singletons came from prokaryote-dominated samples of the free-living size fractions, and singletons from both size fractions were largely dominated by unknowns (83.8% in free-living, 93.2% in >3µm), suggesting a limited impact of eukaryotic contigs on our conclusions. Collector curves' asymptotic profiles were stronger when decreasing detection thresholds (Figure S2), 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 latitudes[32] 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 done through network-based methods allowing the characterization of distant and rare homologs[33].

Gene assemblages were structured by water mass at SO scale, supporting the observation that Processes leading to water mass formation and transport exert the strongest control on microbial community composition[22]. Our statistical approach allowed us to identify genes particularly abundant in the Mertz polynya, corroborating previous findings identifying polynya bacterial communities to be mostly heterotrophs exploiting residues from eukaryotic phytoplankton blooms[22], including taxa playing key roles in primary production limitation by iron and other micronutrients like cobalamin[29,34]. The Mertz polynya was iron-limited at the time of ACE sampling[15], and an investigation of metallo-proteins diversity in Mertz samples through specific annotation tools[35] could help to better identify the roles of prokaryotes in trace metal cycling in the context of diatom blooms. A previous study investigating a transect from Tasmania to Mertz identified a significant difference in genomic composition between samples taken at Mertz and samples taken above the Polar Front[36]. They highlighted the polar front as the main biogeographical boundary, acknowledging that the continental shelf could also explain the partitioning considering their lack of samples between the front and Mertz. Our results suggest a greater difference between populations of the polynya *versus* other AASW populations than between populations on both sides of the polar front, highlighting the uniqueness of SO's coastal biomes. Mertz being the only sampling location above the Antarctic shelf in our dataset, it is impossible to state if our observations could 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.

Phytoplankton dynamics in polynyas usually show dominance of either diatoms or Prymnesiophyceae, mainly *Phaeocystis antarctica*[37,38]. Global warming could be causing a shift from *P. antarctica* to diatom blooms in coastal polynyas[37,39], and the increased sinking rate of diatoms compared to Phaeocystis could impact carbon export[28]. *P. antarctica* did not appear as a significant contributor to any of the 6 CAGs identified as differentially abundant in the Mertz polynya, while it was the main contributor to CAG 34, abundant in waters with low silicic acid and moderately high temperature, far from the Antarctic coastline. Models predict the diatom-*Phaeocystis* competition to mainly depend on iron availability and light sensitivity[40]. A eukaryote-focused re-analysis of the key samples identified through our approach, *i.e.* using eukaryote specific gene-callers in combination with genomes from diatoms and *Phaeocystis* isolates, could lead to the detection of functional markers helping to decipher the mechanisms of the diatom-phaeocystis competition at genomic level. 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. An analysis of the bacterial transcriptional activity in multiple polynyas combined with measures of estimated carbon export[41] would lead to a better understanding of the impact of planktonic compositional switches on remineralization and sinking rates in polynyas, allowing for better predictions of their potential effect on the SO biological carbon pump in a context of global change.

In addition to gene clusters highlighting the functional and taxonomic uniqueness of SO's biomes, we identified gene clusters showing different latitudinal niches and functional profiles despite all being associated with the ubiquitous SAR11. SAR11 subclades adapted to SO's extreme conditions have been observed through amplicon sequencing off the Kerguelen Islands[42], while SAR11 genomes assembled from Arctic metagenomes contained polar-specific genes content, the vast majority of which coded for poorly characterized proteins[43]. Our results suggest a genomic adaptation of SAR11 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. SAR11 could thus play a role in trace metal cycling in SO polar conditions. A strain-resolved analysis of SAR11 genomes based on ACE metagenomes should provide unprecedented insights into SAR11 Southern Ocean adapted ecotypes.

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 Bay[44]. The genomic content of microbial populations in the dark winter of the SO remains to be described by future campaigns. Our results will soon be complemented by viral size fraction metagenomics and >3µm size fraction metatranscriptomics samples from the same campaign, which combined with our metagenomic assemblies should allow a better description of viral and eukaryotic functional adaptations in the SO, offering a holistic view of its unique genomic seascape.

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 databases[12]. 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[16,17], and a catalog of polar-specific ACE ORFs, *i.e.* detected in at least one Arctic sample while being absent from non-polar TOPC samples. Using these catalogs, we quantified the novelty of SO microbial genes, demonstrating their high endemicity. By linking gene-level abundance and environmental metadata, we were able to describe the biogeography of prokaryotes at SO-scale, identifying distinct gene assemblages in different water-masses and defining genomic markers of diverse biomes, from the blooming Mertz polynya to the Southernmost Sub-Tropical waters. Overall, our results advocate for the development of regional-scale descriptions and models of planktonic diversity in the Southern Ocean, distinguishing coastal and offshore 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. Existing Antarctic time-series (*e.g.*, the Palmer LTER[45] or the Rothera time-series[46]) 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.

MATERIAL AND 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* expeditions[47]. 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 100M paired-end reads per library (2x150bp; 500bp 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.* ^{13}C, ^{15}N) 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 biodiversity[48]. 14 latent variables computed through a sparse PCA for each ACE station to summarize the global biogeochemical context[15] were added to the metadata set. Please refer to the original Landwehr et al.[15] paper 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 (Figure 3A). When needed, missing values in the CTD metadata set were imputed using the k-nearest-neighbors approach encoded in the caret R package[49], 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 Minoche[50] approach implemented in illumina-utils[51] with default parameters, and sample-by-sample assemblies were obtained from MEGAHIT v1.2.9[52]. 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 v8[53]. During the generation of the ACE Contigs-DB open-reading frames were detected in all contigs using Prodigal v2.6.3[54] 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, especially in the size fraction >3 µm, Phyloflash v3.4[55] 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 v0.6.7[56] (West et al., 2018) and Whokaryote[57] were used to try and 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 MMSeqs v14.7e284[58] 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.

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

Open-reading frames were detected in all contigs using Prodigal v2.6.3[54]. The 175,336,776 non-dereplicated ORFs constitute the raw ACE gene catalog. Nucleotide sequences were then clustered at 95% similarity and 90% coverage using CD-Hit V4.8.1[59], 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 v2.1.8[60] and KOFamSCAN v1.3.0[61]. 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 coarser yet functionally homogeneous clusters, the AGNOSTOS clustering pipeline[19] 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 integrated 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.[19]. Please 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 (% of the contigs covered at least at 1X) profiles, using Bowtie2 v2.4.5[62], in competitive mode 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 (https://anvio.org/m/anvi-profile-blitz) implemented in anvi'o[53,63] 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 obtained in parallel work (Pommellec et al, in preparation). 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 cluster-level. Detection at 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 (Figure S2). 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 *rlog* method from the DESeq2 R package[64]. Relative log expression normalization method was identified as one of the most adapted to metagenomics-based microbiome studies[65].

Highlighting novelty in Southern Ocean's microbial genes

Protein sequences from the ACE-RGC were clustered with those of the OM-RGC at 30%, 50% and 80% similarity thresholds, with a fixed coverage threshold of 80%. Clusters were separated in three categories: pure ACE when only composed of sequences assembled in ACE samples, pure TARA when only composed of sequences from the OM-RGC, and mixed for the rest. Clusters were further characterized based on their members' origin of assembly, mainly distinguishing sunlit (<150m) and dark (>150m) 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.[16].

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 free-living part corresponding to 0.2-3 and 0.2-40 µm size fractions, and the >3µm part corresponding to the 3-200 µm size fraction. 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 preProcess function from the Caret R Package[49]. The near zero variance definition was set at a minimal threshold of 20% unique abundance values and a maximal ratio of 95 to 5 between the most abundant and second most abundant values.

Identification of AGNOSTOS clusters highly linked with the environment

A random forest regression model was fitted for each cluster of the free-living and >3µm 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 package[66]. 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-squared values from each selected model were retrieved. Density of R-squared values were drawn for free-living and >3µm 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 free-living matrices and two of the >3µm ones. Based on the 95th centile value for each set of permutations, R-squared thresholds were set at 10% for the free-living matrix and 15% for the >3µm 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 Willis[24] was used to group them into groups of Co-Abundant env-AGC (CAG). This approach, based around 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 free-living and >3µm 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 free-living and >3µm CAG-level matrices. 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 free-living and >3µm), 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 free-living and >3µm). For each model, CAGs appearing on the extremities of axis 1 to 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 through MMSeqs2 taxonomic annotation tool with the UniRef90 database as reference[58]. Since it took 15 to 20 hours to annotate splits of 25,000 contigs using 24 CPUs and 80 Gb of memory, only a selection of CAGs of interest were annotated this way. In addition, genes in CAGs were annotated based on the presence of their contig of origin in MAGs from the ACE MAGs database (Pommellec et al., 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.

Code and data availability

The 218 metagenomic samples were deposited on the European Nucleotide Archive (ENA), under the submission code ERA30995399. A correspondence table of all samples linking their ACE unique ID, BioSample code, ENA run and experiment codes is available on Zenodo. Links to all raw and processed data are available in Table S1. The scripts used to produce the results of this study are available at https://github.com/EmileFaure/ACE_gene_centric_scripts. 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.

Acknowledgements

The Antarctic Circumnavigation Expedition was funded by the Swiss Polar Institute and Ferring Pharmaceuticals. This work is dedicated to the memory of late David Walton, Chief scientist of the ACE cruise. Authors thank all onboard expedition members for their support. This study was founded through Agence Nationale de la Recherche grant 18-CE02-0024 to LM, and supported by France Génomique (ANR-10-INBS-09-08). CH was supported by the Swiss National Science Foundation (PP00P2_166197) and is currently supported by the Ferring Pharmaceutical – Margaretha Kamprad Chair in environmental sciences attributed to Prof. J. Chappelaz. NC 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". Authors thank Alberto Baudena from the Consiglio Nazionale Delle Ricerche, Istituto di Scienze Marine (CNR-ISMAR), Lerici (SP), Italy, who provided the Lagrangian data.

References

1. Gray, A. R. The Four-Dimensional Carbon Cycle of the Southern Ocean. Annu. Rev. Mar. Sci. **16**, 163–190 (2024).
2. 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).
3. Tagliabue, A. et al. The integral role of iron in ocean biogeochemistry. Nature **543**, 51–59 (2017).
4. Deppeler, S. L. & Davidson, A. T. Southern Ocean Phytoplankton in a Changing Climate. Front. Mar. Sci. **4**, (2017).
5. 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).
6. Christaki, U. et al. Seasonal microbial food web dynamics in contrasting Southern Ocean productivity regimes. Limnol. Oceanogr. **66**, 108–122 (2021).
7. 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).

8. Doré, H. et al. Differential global distribution of marine picocyanobacteria gene clusters reveals distinct niche-related adaptive strategies. ISME J. **17**, 720–732 (2023).

9. Faure, E., Ayata, S.-D. & Bittner, L. Towards omics-based predictions of planktonic functional composition from environmental data. Nat. Commun. **12**, 4361 (2021).

10. Frémont, P. et al. Restructuring of plankton genomic biogeography in the surface ocean under climate change. Nat. Clim. Change **12**, 393–401 (2022).

11. Sunagawa, S. et al. Tara Oceans: towards global ocean ecosystems biology. Nat. Rev. Microbiol. **18**, 428–445 (2020).

12. Laiolo, E. et al. Metagenomic probing toward an atlas of the taxonomic and metabolic foundations of the global ocean genome. Front. Sci. **1**, (2024).

13. Paoli, L. et al. Biosynthetic potential of the global ocean microbiome. Nature **607**, 111–118 (2022).

14. Cao, S. et al. Structure and function of the Arctic and Antarctic marine microbiota as revealed by metagenomics. Microbiome **8**, 47 (2020).

15. 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).

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

17. Sunagawa, S. et al. Structure and function of the global ocean microbiome. Science **348**, 1261359–1261359 (2015).

18. Acinas, S. G. et al. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. Commun. Biol. **4**, 1–15 (2021).

19. Vanni, C. et al. Unifying the known and unknown microbial coding sequence space. eLife **11**, e67667 (2022).

20. 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).

21. 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).

22. 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**, (2013).

23. 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).

24. 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).

25. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics **23**, 1282–1288 (2007).

26. Xue, C. et al. Polysaccharide utilization by a marine heterotrophic bacterium from the SAR92 clade. FEMS Microbiol. Ecol. **97**, fiab120 (2021).

27. Kappelmann, L. (Meta-)genomic Analysis of the Diversity and the Carbohydrate Degradation Potential of the SAR92 Clade during a Diatom-induced Bacterioplankton Bloom. (University of Bremen Bremen / Germany, 2013).

28. Kim, S.-J. et al. Genomic and metatranscriptomic analyses of carbon remineralization in an Antarctic polynya. Microbiome **7**, 29 (2019).

29. Bertrand, E. M. et al. Phytoplankton–bacterial interactions mediate micronutrient colimitation at the coastal Antarctic sea ice edge. Proc. Natl. Acad. Sci. **112**, 9938–9943 (2015).

30. Mock, T. et al. Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. Nature **541**, 536–540 (2017).

31. Holst, F. et al. Helixer–de novo Prediction of Primary Eukaryotic Gene Models Combining Deep Learning and a Hidden Markov Model. 2023.02.06.527280 Preprint at https://doi.org/10.1101/2023.02.06.527280 (2023).

32. Ibarbalz, F. M. et al. Global Trends in Marine Plankton Diversity across Kingdoms of Life. Cell **179**, 1084-1097.e21 (2019).

33. Sussfeld, D. et al. Network studies unveil new groups of highly divergent proteins in families as old as cellular life with important biological functions in the ocean. 2024.01.08.574615 Preprint at https://doi.org/10.1101/2024.01.08.574615 (2024).

34. 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).

35. Garber, A. I. et al. FeGenie: A Comprehensive Tool for the Identification of Iron Genes and Iron Gene Neighborhoods in Genome and Metagenome Assemblies. Front. Microbiol. **11**, 37 (2020).

36. Wilkins, D. et al. Biogeographic partitioning of Southern Ocean microorganisms revealed by metagenomics. Environ. Microbiol. **15**, 1318–1333 (2013).

37. Arrigo, K. R. & van Dijken, G. L. Phytoplankton dynamics within 37 Antarctic coastal polynya systems. J. Geophys. Res. Oceans **108**, (2003).

38. Schofield, O. et al. In situ phytoplankton distributions in the Amundsen Sea Polynya measured by autonomous gliders. Elem. Sci. Anthr. **3**, 000073 (2015).

39. Arrigo, K. R., Lowry, K. E. & van Dijken, G. L. Annual changes in sea ice and phytoplankton in polynyas of the Amundsen Sea, Antarctica. Deep Sea Res. Part II Top. Stud. Oceanogr. **71–76**, 5–15 (2012).

40. Nissen, C. & Vogt, M. Factors controlling the competition between &lt;i&gt;Phaeocystis&lt;/i&gt; and diatoms in the Southern Ocean and implications for carbon export fluxes. Biogeosciences **18**, 251–283 (2021).

41. Guidi, L. et al. Plankton networks driving carbon export in the oligotrophic ocean. Nature (2016).

42. 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).

43. 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).

44. Piedade, G. J. et al. Seasonal dynamics and diversity of Antarctic marine viruses reveal a novel viral seascape. Nat. Commun. **15**, 9192 (2024).

45. Smith, R. C. et al. The Palmer LTER: A Long-Term Ecological Research Program at Palmer Station, Antarctica. Oceanography **8**, 77–86 (1995).

46. Venables, H. et al. Sustained year-round oceanographic measurements from Rothera Research Station, Antarctica, 1997–2017. Sci. Data **10**, 265 (2023).

47. Alberti, A. et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. Sci. Data **4**, 170093 (2017).

48. Ser-Giacomi, E. et al. Lagrangian betweenness as a measure of bottlenecks in dynamical systems with oceanographic examples. Nat. Commun. **12**, 4935 (2021).

49. Kuhn, M. Building Predictive Models in R Using the caret Package. J. Stat. Softw. **28**, 1–26 (2008).

50. 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).

51. 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).

52. 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 btv033 (2015).

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

54. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11**, 1 (2010).

55. Gruber-Vodicka, H. R., Seah, B. K. B. & Pruesse, E. phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes. mSystems **5**, (2020).

56. 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).

57. Pronk, L. J. U. & Medema, M. H. Whokaryote: distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. Microb. Genomics **8**, 000823 (2022).

58. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. **35**, 1026–1028 (2017).

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

60. 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).

61. Aramaki, T. et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics **36**, 2251–2252 (2020).

62. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**, 357–359 (2012).

63. Eren, A. M. et al. Community-led, integrated, reproducible multi-omics with anvi'o. Nat. Microbiol. **6**, 3–6 (2021).

64. 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).

65. 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).

66. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J. Stat. Softw. **77**, (2017).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FaureetalWaterMassSpecificPolarGenesDominateTheSOMicrobiomeSupplementaryMaterials.docx