

Supplementary Materials

SUPPLEMENTARY FIGURES AND TABLES

Resource	Type	Link
Raw reads	Nucl. Fastq.gz	ENA
Metagenomic assemblies	Nucl. Fasta	IFREMER
Unigenes catalog	Nucl. Fasta	Zenodo
Polar-specific ORF catalog	Nucl. Fasta	Zenodo
SO-RGC (unigenes from 0.2-3µm)	Nucl. Fasta	Zenodo
AGC catalog	TSV file	Zenodo
AGC-level coverage matrix	TSV file	Zenodo
AGC-level detection matrix	TSV file	Zenodo
AGC annotations table (functional and taxonomic)	TSV files	Zenodo
AGC-level random forest outputs	TSV files	Zenodo
Metadata files	TSV files	Figshare
CAGs catalogs	TSV files	Figshare
CAG-level coverage matrix for small and large size-fractions	TSV files	Figshare
CAGs-level functional and taxonomic annotation tables	TSV files	Figshare
CAGs of interest functional enrichment	TSV files	Figshare

Table S1: List of resources made publicly available through this project.

		Description	Depth (Meters)			Temperature (°C)			Salinity (PSU)			Oxygen (μmol/kg)		
			min	mean	max	min	mean	max	min	mean	max	min	mean	max
Surface and subsurface waters	Antarctic Surface Waters (AASW)	Highly oxygenated, cold and fresh surface waters, sampled South of the Polar Front.	4	18	85	-1,49	1,10	5,12	32,54	33,52	34,18	287,4	326,7	378,6
	Sub-Antarctic Surface Waters (SASW)	Warmer, less oxygenated waters. Sampled North of the polar front, sometimes North of the Sub-Antarctic front, but colder than STSW.	4	33	150	4,41	5,06	6,69	33,75	33,83	33,93	277,4	287,3	292,8
	Sub-Tropical Surface Waters (STSW)	Warmer and saltier waters, sampled North of the subantarctic front, sometimes North of the Sub-Tropical front.	4	39	150	8,50	10,58	12,49	34,14	34,51	34,83	208,5	250,0	268,6
	Winter-Waters (WW)	Cold surface or subsurface (in between AASW and CDW) layer that persists from the previous winter.	50	143	150	-1,58	0,45	2,53	33,86	34,29	34,66	168,9	242,9	321,9
Intermediate waters	Antarctic Intermediate Waters mixed with surface waters (SW-AAIW)	Mix of AAIW (cold and fresh water mass) with the different surface water masses.	150	339	1000	0,71	5,00	7,87	33,80	34,16	34,42	185,0	253,8	309,3
Deep waters	Circumpolar Deep Waters (CDW)	Salty, low oxygen and relatively warm deep water mass.	150	867	1000	0,61	1,84	3,47	34,40	34,66	34,75	157,6	174,4	191,2
Bottom waters	Dense Shelf Waters (DSW)	Very cold and salty waters, formed at the shelf (mostly in polynyas).	150	460	772	-1,97	-1,82	-1,66	34,43	34,57	34,66	281,8	283,7	286,1
	Antarctic Bottom Waters (AABW)	Ocean bottom layer, dense and salty.	3460	3587	3800	0,35	0,36	0,36	34,70	34,70	34,70	199,6	200,4	201,1

Table S2: Main characteristics of water masses depicted in Figures 3 and 5. The values indicated for each variable correspond to minimum, mean and maximum across all CTD samples associated with each water mass. Positions of samples relative to hydrographic fronts are estimated using fronts positions given in Haumann et al. (2020)¹.

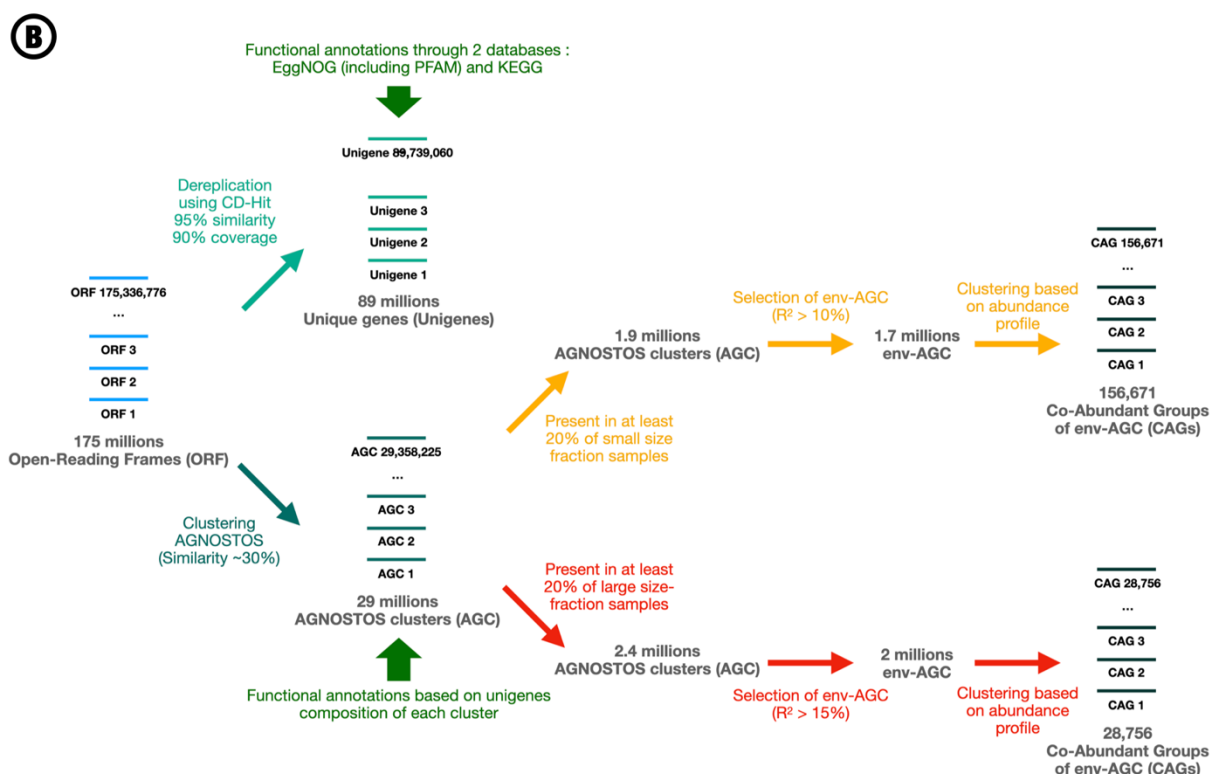


Figure S1: Diagrams of methodological steps implemented to produce AGC and CAG-level matrices from raw metagenomic reads. (A) focuses on the bioinformatics steps to go from raw reads to gene-level matrices. Steps 1. to 4. are implemented in our NOEMIE Nextflow pipeline, now publicly available at <https://gitlab.ifremer.fr/bioinfo/workflows/noemie>. (B) focuses on how individual ORFs were clustered into unigenes, AGC and CAGs.

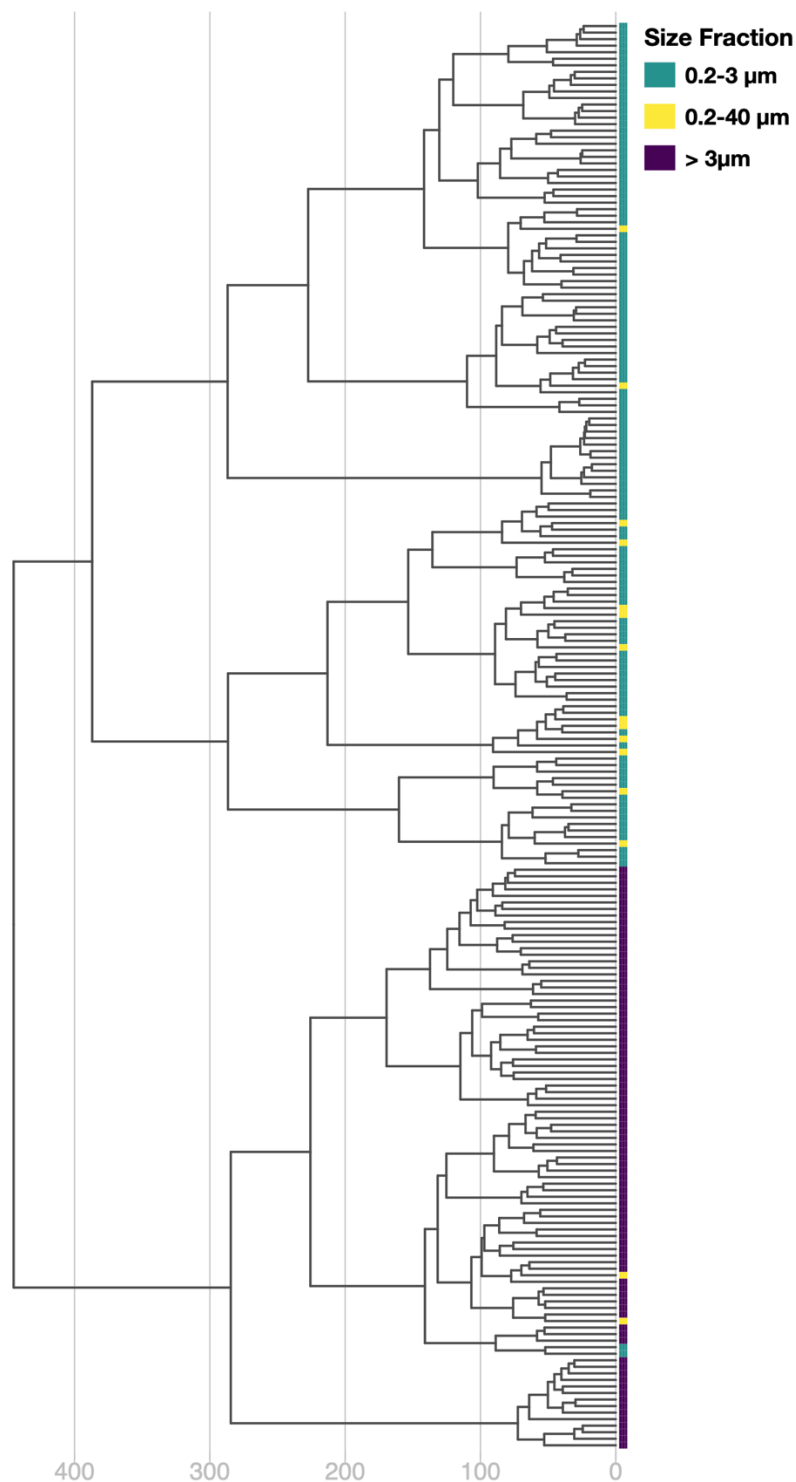


Figure S2: Hierarchical clustering of the 218 ACE metagenomes based on *k*-mer composition (Bray-Curtis distance, Ward D2 method, *k*-mer length of 21 base pair, achieved using SIMKA²). When cutting the tree in two, all but four samples from the small size-fraction (0.2-3 and 0.2-40 µm) cluster together, while the four remaining samples cluster with the large size fraction. 0.2-40 µm samples do not cluster together within the small size fraction, demonstrating the similarity between the two filter sizes and justifying their pooling into a single size fraction. Source data are provided as a Source Data file.

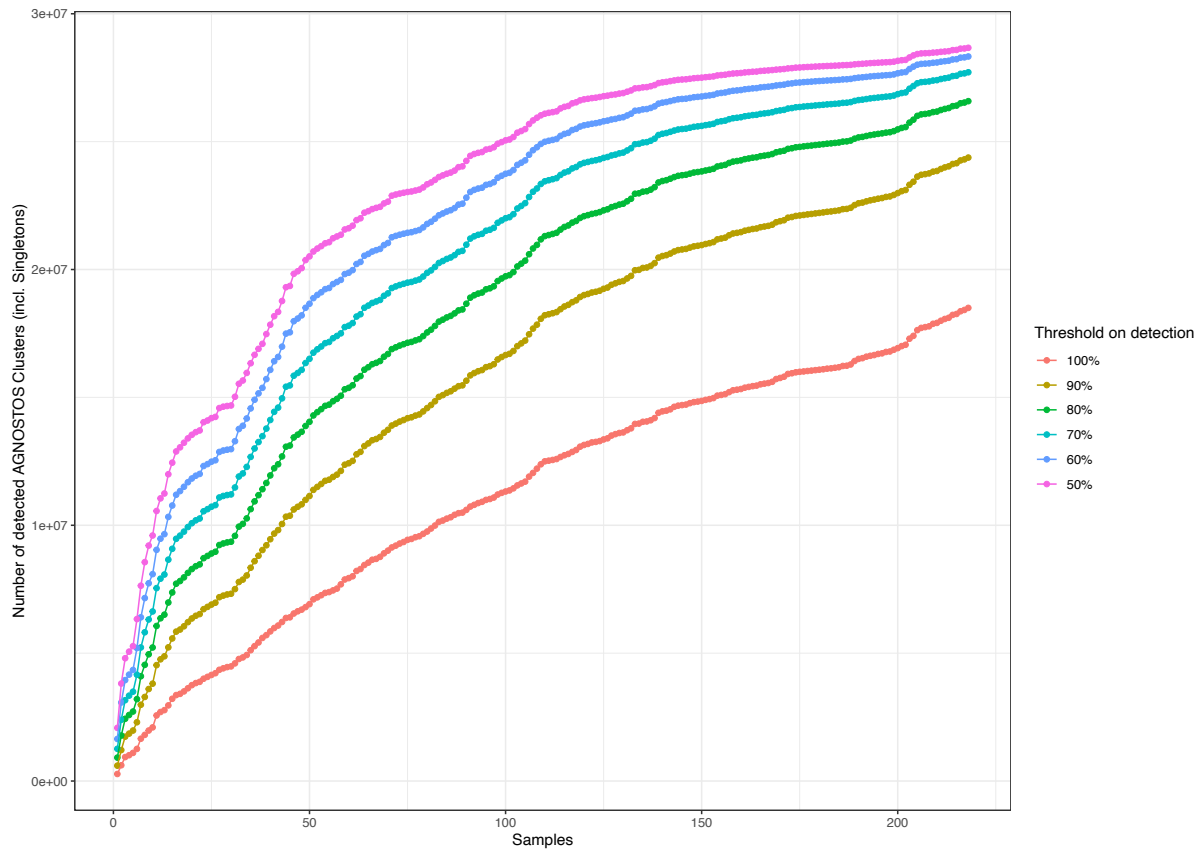


Figure S3: AGNOSTOS gene clusters (AGC) based collector curves across the 218 ACE metagenomes, drawn using different thresholds on detection. Detection is the proportion of nucleotide in a given ORF mapped with a coverage of at least 1X. As described in the methods, AGC-level detection in a given sample is defined as the maximum detection value across all ORFs constituting the AGC. Coverage values corresponding to detection values below threshold are then switched to 0. Higher thresholds thus guarantee that at least one ORF in the AGC was mapped across a significant part of its length by reads, but they provoke false negatives, as shown by the low maximum on the 100% detection threshold collector curve which misses about 10 million assembled gene clusters. In most metagenomics projects, collector curves are drawn without setting thresholds on coverage or detection, producing potentially misleading results. Source data are provided as a Source Data file.

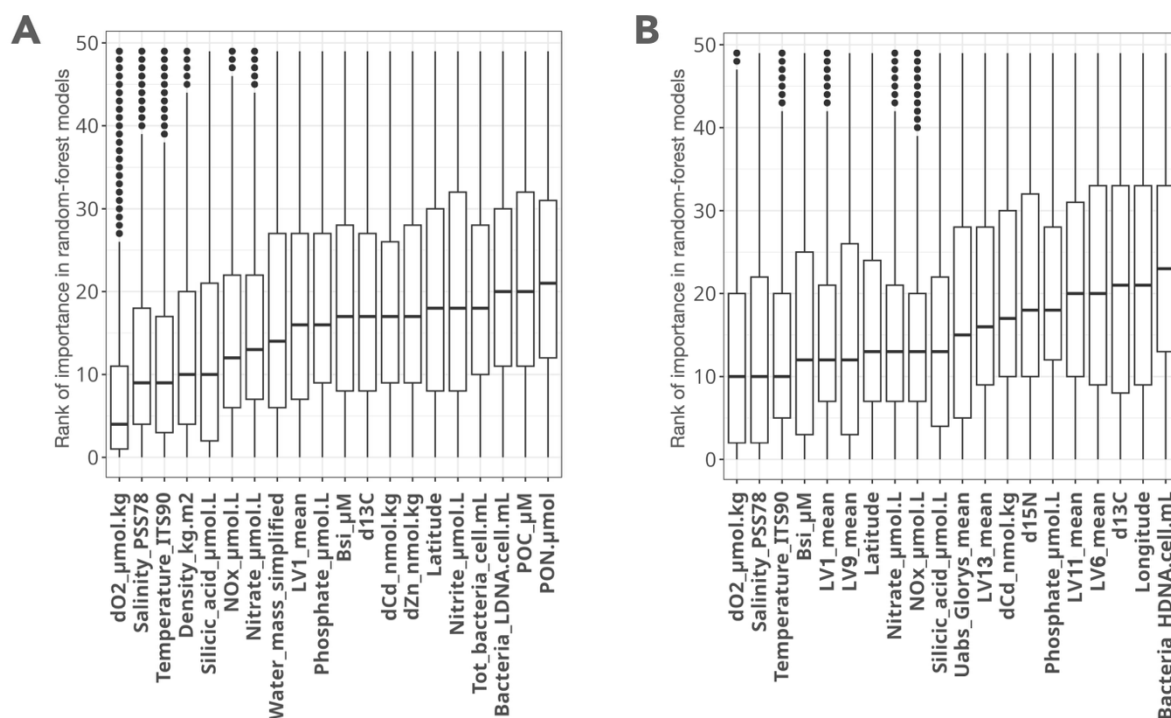


Figure S4: 20 best environmental predictors of AGC abundance at SO-scale. Boxplots of ranked importance across all random forest models are drawn for each predictor in the small (A) and large (B) size-fractions. Each boxplot is displaying the median, interquartile range (1st and 3rd quartile) and whiskers extending to most extreme values within 1.5x the interquartile range, with $n=1,906,625$ models (A) or $n=2,437,989$ models (B). Ranked importance reflects the position of importance in predictions, from 1 (the most important predictor) to 50 (the least important).

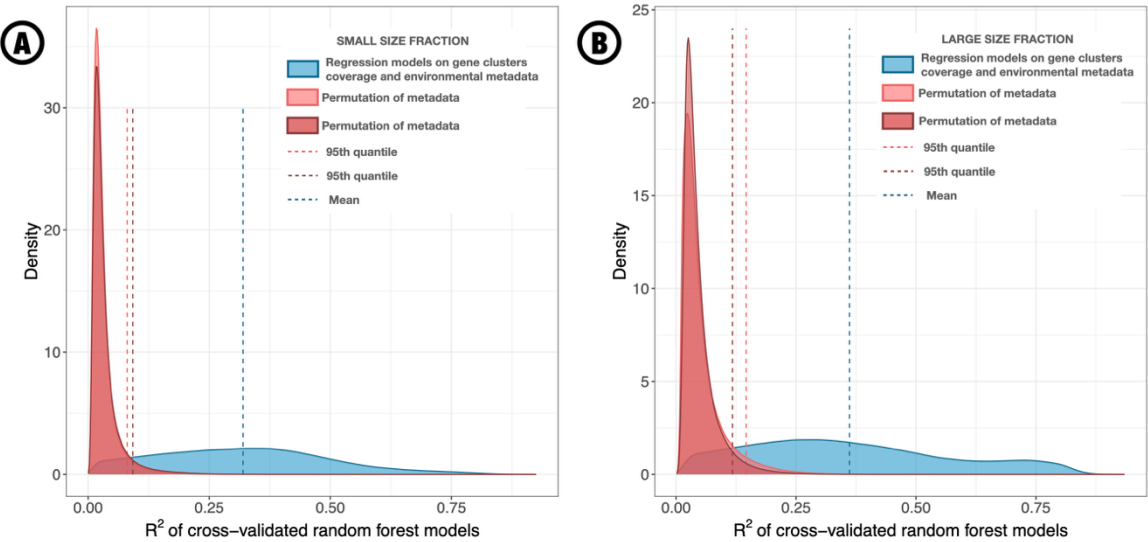


Figure S5: Gene clusters' abundance can be predicted from environmental context at Southern Ocean scale. The density of cross-validation R -squared is compared between models using environmental context to predict AGC abundance (blue), and the same models using randomly permuted environmental context (red). This is achieved for each AGC in the small (A, $n=1,906,625$ models) and large (B, $n=2,437,989$ models) size-fractions. Cross-validation R^2 values ranged from 0.2% to 92.5% (mean = 32.0%, median = 31.3%) across 1,906,624 models for the small size fraction, and from 0.3% to 93.5% (mean = 36.1%; median = 33.4%) across 2,437,988 models for the large size fraction (Figure S4). The same protocol was applied after randomly permuting rows in the environmental metadata, twice for each size fraction (Figure S4). For the small size fraction (resp. large), 95% of models built on permuted data showed R^2 values below 8.1% and 9.2% (resp. 11.8% and 14.6%) in each permutation.

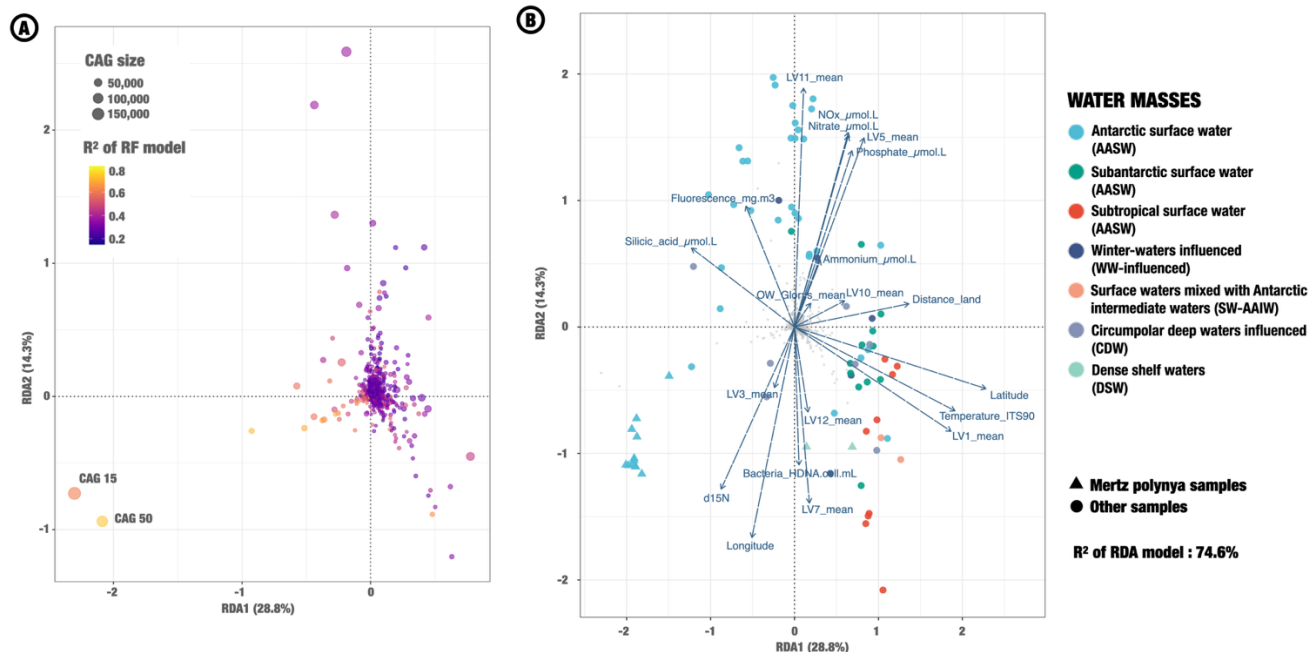


Figure S6: Redundancy analysis of CAGs abundance in response to environmental metadata in the large size fractions. RDA triplot was separated in two parts for better readability, A showing the distribution of CAGs in the RDA space, colored according to their mean random forest R -squared (reflecting 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 the main text are indicated with grey labels. (B) shows both samples and environmental variables distribution in the RDA space. Samples are colored according to their water mass of origin, as in Fig. 3. The first axis of the RDA opposed warmer, low latitude samples ($RDA1 > 0$) from colder, coastal Antarctic samples ($RDA1 < 0$). The second axis opposed rich samples ($RDA2 > 0$) from samples limited in nutrients ($RDA2 < 0$).

SUPPLEMENTARY RESULTS

Reduced amounts of unknowns in deep water masses

CAGs 39 and 33 were associated with deep samples in the small size fraction (RDA1>2, Figure 4A). Dominant families associated with these CAGs based on contigs annotation using UniRef90 were Pelagibacteraceae, Rhodospirillaceae, Nitrospinaceae, Planctomycetaceae, Nitrosopumilaceae and Acidimicrobiaceae. Contigs related to these CAGs were binned in MAGs mainly annotated as Archaea from the Thalassarchaeaceae, Poseidoniaceae and Nitrosopumilaceae families. CAG 39 and 33 were both significantly enriched in cytochrome C, as well as in multiple functional annotations linked with arylsulfatase, e.g. COG3119 Arylsulfatase A, arylsulfatase activity. The two best predictors of AGCs' abundance were oxygen and nitrite concentrations within CAG 39, oxygen and density within CAG 33 (NOx 5th, Nitrate 8th, Nitrite 9th). Arylsulfatase was identified as a marker of ammonia oxydizing archaea and ammonia oxydizing bacteria presence in aquaculture ponds³. Our results point towards a dominance of autotrophic ammonia oxydizing micro-organisms in the deep Southern Ocean, which has been described before through amplicon data⁴. No CAG from the large size fraction was found to be linked to deeper samples, with deep samples appearing mixed with surface ones on both RDA1 and RDA2 (Figure 4D). Most samples collected from the large size fraction at depths of 1000m or more did not reach the necessary threshold of DNA concentrations for extraction and sequencing following our methodological pipeline (Figure 1B). This suggests the absence or rarity of larger sized eukaryotic organisms in the mesopelagic zone, further confirmed by the position of the deepest samples on our RDA space.

CAG 39 contained 65.6% of known AGC and 32.9% of genomic unknown AGC, leaving only 1.5% of environmental unknown AGC over a total of 30,159. Similarly, only 5.8% of the 41,361 AGCs from CAG 33 were environmental unknowns, one order of magnitude below the mean over all AGCs (54.6%). Looking at other dimensions of the RDA, we identified three other CAGs (67400, 133918 and 558) with higher abundances in deep, nitrite limited waters, with respectively 3.4%, 4.3% and 8% of environmental unknowns. CAG 558, which had a slightly higher number of environmental unknowns compared to other deep CAGS, was significantly enriched in phage integrase and phage plasmid primase P4, suggesting it might include viral DNA which could increase its proportion of unknowns. Going back to Figure 2A, we observed that ACE genes assembled below the sunlit layer clustered better with Tara unigenes than ACE genes assembled from the sunlit layer. Taken together, these results suggest that organisms sampled at higher depth during the ACE campaign are more similar to organisms with genomic references in public databases than those sampled from the surface. Our hypothesis to explain this result is that deep environments are under more homogeneous selective pressures at global scale than surface ones (high pressure, no light, low yet stable temperature), leading similar organisms to dominate globally while the surface waters of the Antarctic are a unique ecosystem on earth showing high rates of local endemism.

More on eukaryotic CAGs from the Mertz polynya

In the large size fraction, CAGs 15 and 50 were outliers on the negative side of RDA1, indicating a strong association with Mertz polynya (Figure S5). Unlike the four CAGs associated with Mertz samples in the small size fraction, CAGs 15 and 50 contained a majority of unknowns AGC, reaching 90.1% of environmental and genomic unknowns in CAG 15. Only

11 eggNOG annotations were enriched in CAG 15, most of them linked with transcription. 147 eggNOG annotations were significantly enriched in CAG 50, including enzymes involved in the carbohydrate metabolism (e.g. sugar transporters, glycosyltransferase sugar binding region), in primary production (e.g., Chlorophyll A-B binding protein, ammonium transmembrane transporter activity), in iron metabolism (iron ion binding, 2OG-Fe(II) oxygenase superfamily) and in sulfur metabolism (e.g. Sulfotransferase family). CAGs 15 and 50 had respectively 56.5% and 88.2% of their contigs annotated to the Bacillariophyceae class, including 37.6% and 71.6% to *Fragilariopsis cylindrus*.

Two examples of CAGs corresponding to specialist species

Two of the CAGs showing clear responses to latitude dependent variables in the small size fraction, 131 and 34 (Figure 4), were linked with specialist species showing narrow realized niche. CAG 131 was polar-specific, abundant in oxygenated cold waters from AASW, nearly absent in STSW samples and in low abundance in SASW. Its contigs were mainly annotated as Rhodobacteraceae and Roseobacteraceae, highlighting the dominance of these clades in the coldest surface waters sampled during ACE campaign. CAG 34 was positively correlated to temperature until reaching values typical of STSW samples, in which it showed low coverage values, supposing a latitudinal barrier in its distribution at the Subantarctic front. CAG 34 was dominated by genes from small unicellular eukaryotes, mainly *Phaeocystis antarctica*, but also *Emiliania huxleyi*, *Chrysochromulina tobinii* and other haptophytes. Its position in the RDA thus reflects the combined effects of inter-species competition and environmental gradients on the realized niche of these eukaryotic species, absent from both diatoms-dominated coastal samples and warm oligotrophic conditions from STSW, on the two extremities of ACE's latitudinal range.

Phytoplankton dynamics in polynyas usually show dominance of either diatoms or Prymnesiophyceae, mainly *Phaeocystis antarctica*^{5,6}. Global warming could be causing a shift from *P. antarctica* to diatom blooms in coastal polynyas^{5,7}, and the increased sinking rate of diatoms compared to *Phaeocystis* could impact carbon export⁸. *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 sensitivity⁹. 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. An analysis of the bacterial transcriptional activity in multiple polynyas combined with measures of estimated carbon export¹⁰ 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.

Biogeographical outliers at Southern Ocean scale

In addition to CAGs with peripheral locations in the RDA space, we investigated CAGs with high predictability based on random forest regressions despite a central position on the RDA. These CAGs tended to be smaller in size, and abundant only in a restraint number of samples. For example, CAG 136 contained only 926 AGC, with a mean cross-validation R2 of 62.9%.

It displayed low abundances in all samples but 4, all from the sampling event 264, the only one at Kerguelen islands (Figure 1A). Similarly, CAG 177401 showed a mean R2 of 62.3% with only 139 AGC, and was abundant only at Crozet (event 123) and Kerguelen islands (event 264). Finally, CAG 73614 had a mean R2 of 60.7% over its 250 AGC, with higher abundances at Crozet, Kerguelen and Prince Edwards (event 75). The LV6 latent variable from Landwehr et al. (2021), reflecting the iron-fertilized biological productivity, was the best predictor of CAGs 177401 and 73614, while it was the fourth best for CAG 136. Enrichment statistics were not computed on these three CAGs due to the limited number of genes in each of the them, yet functions linked to iron displayed odd ratios above 50 in CAGs 136 (*Belongs to the iron ascorbate-dependent oxidoreductase family*) and 177401 (*COG1629 Outer membrane receptor proteins, mostly Fe transport*). Southern Ocean islands thus display specific genomic signatures, potentially linked with high biological activity in an iron fertilization context very different from Mertz polynya, where iron limitation is estimated to be strong¹¹ (LV8, reflecting iron-limited biological activity, peaks at Mertz and was the second-best predictor of CAG 136). A series of four CAGs showed particularly high abundances at a specific event, located in the Ross Sea gyre (event 1465), all in the 0.2-40 µm size fraction. These four CAGs all showed extremely high percentages of environmental unknowns across their AGC: 74.3% for CAG 11, 85.9% for CAG 307481, 96.6% for CAG 82 and 99.1% for CAG 196. The taxonomy estimates for these CAGs reveal communities of eukaryotes, dominated by genus *Emiliana* in CAG 82, *Symbiodinium* in CAGs 11 and 196, and *Fragilariopsis* in CAG 307481. The abundance profiles of these CAGs suggest the presence of specific eukaryotic communities in the Ross Sea gyre compared to other sampling areas of the transect, notably including high proportions of dinoflagellates. Again, our approach is too prokaryote-centered to explore the potential functional adaptations from these quasi-exclusively environmental unknown AGC. Their detection by our statistical approach still underlines the heterogeneity of genomic composition from prokaryotic communities at the scale of the Southern Ocean.

1. Haumann, F. A. *et al.* Physical and biogeochemical oceanography data from underway measurements with an AquaLine Ferrybox during the Antarctic Circumnavigation Expedition (ACE). Zenodo <https://doi.org/10.5281/zenodo.3660852> (2020).
2. Benoit, G. *et al.* Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Comput. Sci.* **2**, e94 (2016).
3. Dai, L. *et al.* Organic Matter Regulates Ammonia-Oxidizing Bacterial and Archaeal Communities in the Surface Sediments of Ctenopharyngodon idellus Aquaculture Ponds. *Front. Microbiol.* **9**, (2018).
4. Sow, S. L. S. *et al.* Biogeography of Southern Ocean prokaryotes: a comparison of the Indian and Pacific sectors. *Environ. Microbiol.* **24**, 2449–2466 (2022).
5. Arrigo, K. R. & van Dijken, G. L. Phytoplankton dynamics within 37 Antarctic coastal polynya systems. *J. Geophys. Res. Oceans* **108**, (2003).
6. Schofield, O. *et al.* In situ phytoplankton distributions in the Amundsen Sea Polynya measured by autonomous gliders. *Elem. Sci. Anthr.* **3**, 000073 (2015).
7. 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.*

212 *Oceanogr.* **71–76**, 5–15 (2012).

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

215 9. Nissen, C. & Vogt, M. Factors controlling the competition between
216 <i>Phaeocystis</i> and diatoms in the Southern Ocean and implications for carbon
217 export fluxes. *Biogeosciences* **18**, 251–283 (2021).

218 10. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean.
219 *Nature* **532**, 465–470 (2016).

220 11. Landwehr, S. *et al.* Exploring the coupled ocean and atmosphere system with a data
221 science approach applied to observations from the Antarctic Circumnavigation Expedition.
222 *Earth Syst. Dyn.* **12**, 1295–1369 (2021).
223