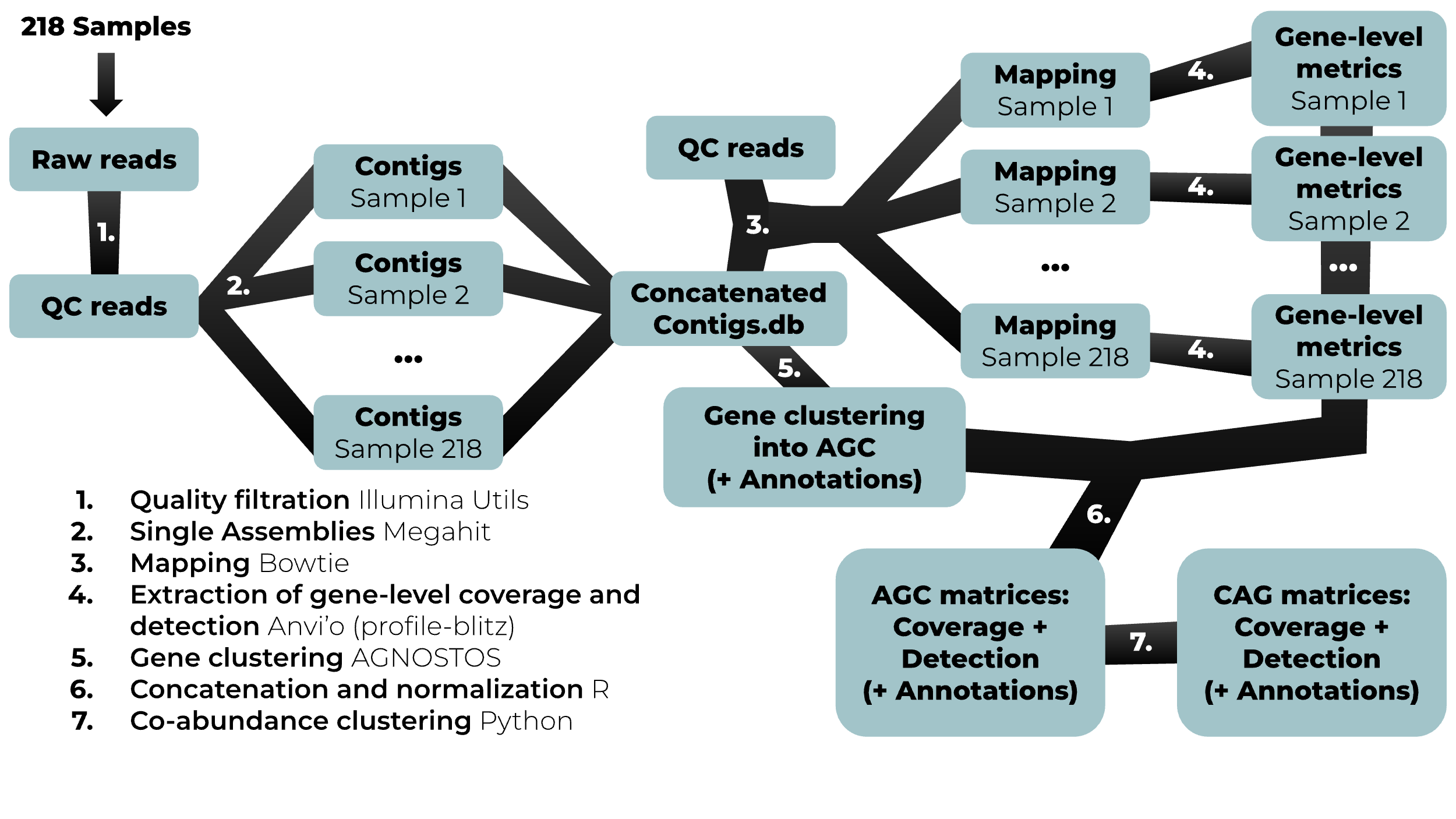
**Supplementary Materials**

SUPPLEMENTARY FIGURES AND TABLES

|  |  |  |
| --- | --- | --- |
| **Resource** | **Type** | **Link** |
| Raw reads | Nucl. Fastq.gz | ENA |
| Metagenomic assemblies | Nucl. Fasta | [IFREMER](https://sextant.ifremer.fr/geonetwork/srv/api/records/9c786963-e6a5-4a1c-95f8-1c6ab8a52d2b) |
| Unigenes catalog | Nucl. Fasta | [Zenodo](https://zenodo.org/records/14181291?preview=1&token=eyJhbGciOiJIUzUxMiIsImlhdCI6MTczMzQwOTA3OCwiZXhwIjoxNzUxMjQxNTk5fQ.eyJpZCI6ImY1ZWQ5ZDZmLTM5YjUtNDYwZS1hYWU5LWI3MjA5ZGUyYjViNyIsImRhdGEiOnt9LCJyYW5kb20iOiIxMDk2NTViZjQwNDJjNTQ4ZDQ5NjU0MzE0MzAxMjcxNSJ9.gvOin3q86LW8Vl2RT3TohYWrUrgCSyzvxnhxlDtuKReXF4Q1khAt8mhrChJUsaqFPmL9H4JK7OXnwTRgDHmdgw) |
| Polar-specific ORF catalog | Nucl. Fasta | [Zenodo](https://zenodo.org/records/14181291?preview=1&token=eyJhbGciOiJIUzUxMiIsImlhdCI6MTczMzQwOTA3OCwiZXhwIjoxNzUxMjQxNTk5fQ.eyJpZCI6ImY1ZWQ5ZDZmLTM5YjUtNDYwZS1hYWU5LWI3MjA5ZGUyYjViNyIsImRhdGEiOnt9LCJyYW5kb20iOiIxMDk2NTViZjQwNDJjNTQ4ZDQ5NjU0MzE0MzAxMjcxNSJ9.gvOin3q86LW8Vl2RT3TohYWrUrgCSyzvxnhxlDtuKReXF4Q1khAt8mhrChJUsaqFPmL9H4JK7OXnwTRgDHmdgw) |
| SO-RGC (unigenes from 0.2-3µm) | Nucl. Fasta | [Zenodo](https://zenodo.org/records/14181291?preview=1&token=eyJhbGciOiJIUzUxMiIsImlhdCI6MTczMzQwOTA3OCwiZXhwIjoxNzUxMjQxNTk5fQ.eyJpZCI6ImY1ZWQ5ZDZmLTM5YjUtNDYwZS1hYWU5LWI3MjA5ZGUyYjViNyIsImRhdGEiOnt9LCJyYW5kb20iOiIxMDk2NTViZjQwNDJjNTQ4ZDQ5NjU0MzE0MzAxMjcxNSJ9.gvOin3q86LW8Vl2RT3TohYWrUrgCSyzvxnhxlDtuKReXF4Q1khAt8mhrChJUsaqFPmL9H4JK7OXnwTRgDHmdgw) |
| AGC catalog | TSV file | [Zenodo](https://zenodo.org/records/14181291?preview=1&token=eyJhbGciOiJIUzUxMiIsImlhdCI6MTczMzQwOTA3OCwiZXhwIjoxNzUxMjQxNTk5fQ.eyJpZCI6ImY1ZWQ5ZDZmLTM5YjUtNDYwZS1hYWU5LWI3MjA5ZGUyYjViNyIsImRhdGEiOnt9LCJyYW5kb20iOiIxMDk2NTViZjQwNDJjNTQ4ZDQ5NjU0MzE0MzAxMjcxNSJ9.gvOin3q86LW8Vl2RT3TohYWrUrgCSyzvxnhxlDtuKReXF4Q1khAt8mhrChJUsaqFPmL9H4JK7OXnwTRgDHmdgw) |
| CAGs catalog | TSV file | [Zenodo](https://zenodo.org/records/14181291?preview=1&token=eyJhbGciOiJIUzUxMiIsImlhdCI6MTczMzQwOTA3OCwiZXhwIjoxNzUxMjQxNTk5fQ.eyJpZCI6ImY1ZWQ5ZDZmLTM5YjUtNDYwZS1hYWU5LWI3MjA5ZGUyYjViNyIsImRhdGEiOnt9LCJyYW5kb20iOiIxMDk2NTViZjQwNDJjNTQ4ZDQ5NjU0MzE0MzAxMjcxNSJ9.gvOin3q86LW8Vl2RT3TohYWrUrgCSyzvxnhxlDtuKReXF4Q1khAt8mhrChJUsaqFPmL9H4JK7OXnwTRgDHmdgw) |
| AGC-level coverage matrix | TSV file | [Zenodo](https://zenodo.org/records/14181291?preview=1&token=eyJhbGciOiJIUzUxMiIsImlhdCI6MTczMzQwOTA3OCwiZXhwIjoxNzUxMjQxNTk5fQ.eyJpZCI6ImY1ZWQ5ZDZmLTM5YjUtNDYwZS1hYWU5LWI3MjA5ZGUyYjViNyIsImRhdGEiOnt9LCJyYW5kb20iOiIxMDk2NTViZjQwNDJjNTQ4ZDQ5NjU0MzE0MzAxMjcxNSJ9.gvOin3q86LW8Vl2RT3TohYWrUrgCSyzvxnhxlDtuKReXF4Q1khAt8mhrChJUsaqFPmL9H4JK7OXnwTRgDHmdgw) |
| AGC-level detection matrix | TSV file | [Zenodo](https://zenodo.org/records/14181291?preview=1&token=eyJhbGciOiJIUzUxMiIsImlhdCI6MTczMzQwOTA3OCwiZXhwIjoxNzUxMjQxNTk5fQ.eyJpZCI6ImY1ZWQ5ZDZmLTM5YjUtNDYwZS1hYWU5LWI3MjA5ZGUyYjViNyIsImRhdGEiOnt9LCJyYW5kb20iOiIxMDk2NTViZjQwNDJjNTQ4ZDQ5NjU0MzE0MzAxMjcxNSJ9.gvOin3q86LW8Vl2RT3TohYWrUrgCSyzvxnhxlDtuKReXF4Q1khAt8mhrChJUsaqFPmL9H4JK7OXnwTRgDHmdgw) |
| AGC-level coverage matrix with detection and non-zero variance thresholds applied | R object | [Zenodo](https://zenodo.org/records/14181291?preview=1&token=eyJhbGciOiJIUzUxMiIsImlhdCI6MTczMzQwOTA3OCwiZXhwIjoxNzUxMjQxNTk5fQ.eyJpZCI6ImY1ZWQ5ZDZmLTM5YjUtNDYwZS1hYWU5LWI3MjA5ZGUyYjViNyIsImRhdGEiOnt9LCJyYW5kb20iOiIxMDk2NTViZjQwNDJjNTQ4ZDQ5NjU0MzE0MzAxMjcxNSJ9.gvOin3q86LW8Vl2RT3TohYWrUrgCSyzvxnhxlDtuKReXF4Q1khAt8mhrChJUsaqFPmL9H4JK7OXnwTRgDHmdgw) |
| CAG-level coverage matrix for Free-living size-fractions | TSV file | [Zenodo](https://zenodo.org/records/14181291?preview=1&token=eyJhbGciOiJIUzUxMiIsImlhdCI6MTczMzQwOTA3OCwiZXhwIjoxNzUxMjQxNTk5fQ.eyJpZCI6ImY1ZWQ5ZDZmLTM5YjUtNDYwZS1hYWU5LWI3MjA5ZGUyYjViNyIsImRhdGEiOnt9LCJyYW5kb20iOiIxMDk2NTViZjQwNDJjNTQ4ZDQ5NjU0MzE0MzAxMjcxNSJ9.gvOin3q86LW8Vl2RT3TohYWrUrgCSyzvxnhxlDtuKReXF4Q1khAt8mhrChJUsaqFPmL9H4JK7OXnwTRgDHmdgw) |
| CAG-level coverage matrix for >3µm size-fractions | TSV file | [Zenodo](https://zenodo.org/records/14181291?preview=1&token=eyJhbGciOiJIUzUxMiIsImlhdCI6MTczMzQwOTA3OCwiZXhwIjoxNzUxMjQxNTk5fQ.eyJpZCI6ImY1ZWQ5ZDZmLTM5YjUtNDYwZS1hYWU5LWI3MjA5ZGUyYjViNyIsImRhdGEiOnt9LCJyYW5kb20iOiIxMDk2NTViZjQwNDJjNTQ4ZDQ5NjU0MzE0MzAxMjcxNSJ9.gvOin3q86LW8Vl2RT3TohYWrUrgCSyzvxnhxlDtuKReXF4Q1khAt8mhrChJUsaqFPmL9H4JK7OXnwTRgDHmdgw) |
| Annotations table (functional and taxonomic) | TSV file | [Zenodo](https://zenodo.org/records/14181291?preview=1&token=eyJhbGciOiJIUzUxMiIsImlhdCI6MTczMzQwOTA3OCwiZXhwIjoxNzUxMjQxNTk5fQ.eyJpZCI6ImY1ZWQ5ZDZmLTM5YjUtNDYwZS1hYWU5LWI3MjA5ZGUyYjViNyIsImRhdGEiOnt9LCJyYW5kb20iOiIxMDk2NTViZjQwNDJjNTQ4ZDQ5NjU0MzE0MzAxMjcxNSJ9.gvOin3q86LW8Vl2RT3TohYWrUrgCSyzvxnhxlDtuKReXF4Q1khAt8mhrChJUsaqFPmL9H4JK7OXnwTRgDHmdgw) |
| CAGs of interest taxonomy | TSV files | [Zenodo](https://zenodo.org/records/14181291?preview=1&token=eyJhbGciOiJIUzUxMiIsImlhdCI6MTczMzQwOTA3OCwiZXhwIjoxNzUxMjQxNTk5fQ.eyJpZCI6ImY1ZWQ5ZDZmLTM5YjUtNDYwZS1hYWU5LWI3MjA5ZGUyYjViNyIsImRhdGEiOnt9LCJyYW5kb20iOiIxMDk2NTViZjQwNDJjNTQ4ZDQ5NjU0MzE0MzAxMjcxNSJ9.gvOin3q86LW8Vl2RT3TohYWrUrgCSyzvxnhxlDtuKReXF4Q1khAt8mhrChJUsaqFPmL9H4JK7OXnwTRgDHmdgw) |
| CAGs of interest functional enrichment | TSV files | [Zenodo](https://zenodo.org/records/14181291?preview=1&token=eyJhbGciOiJIUzUxMiIsImlhdCI6MTczMzQwOTA3OCwiZXhwIjoxNzUxMjQxNTk5fQ.eyJpZCI6ImY1ZWQ5ZDZmLTM5YjUtNDYwZS1hYWU5LWI3MjA5ZGUyYjViNyIsImRhdGEiOnt9LCJyYW5kb20iOiIxMDk2NTViZjQwNDJjNTQ4ZDQ5NjU0MzE0MzAxMjcxNSJ9.gvOin3q86LW8Vl2RT3TohYWrUrgCSyzvxnhxlDtuKReXF4Q1khAt8mhrChJUsaqFPmL9H4JK7OXnwTRgDHmdgw) |

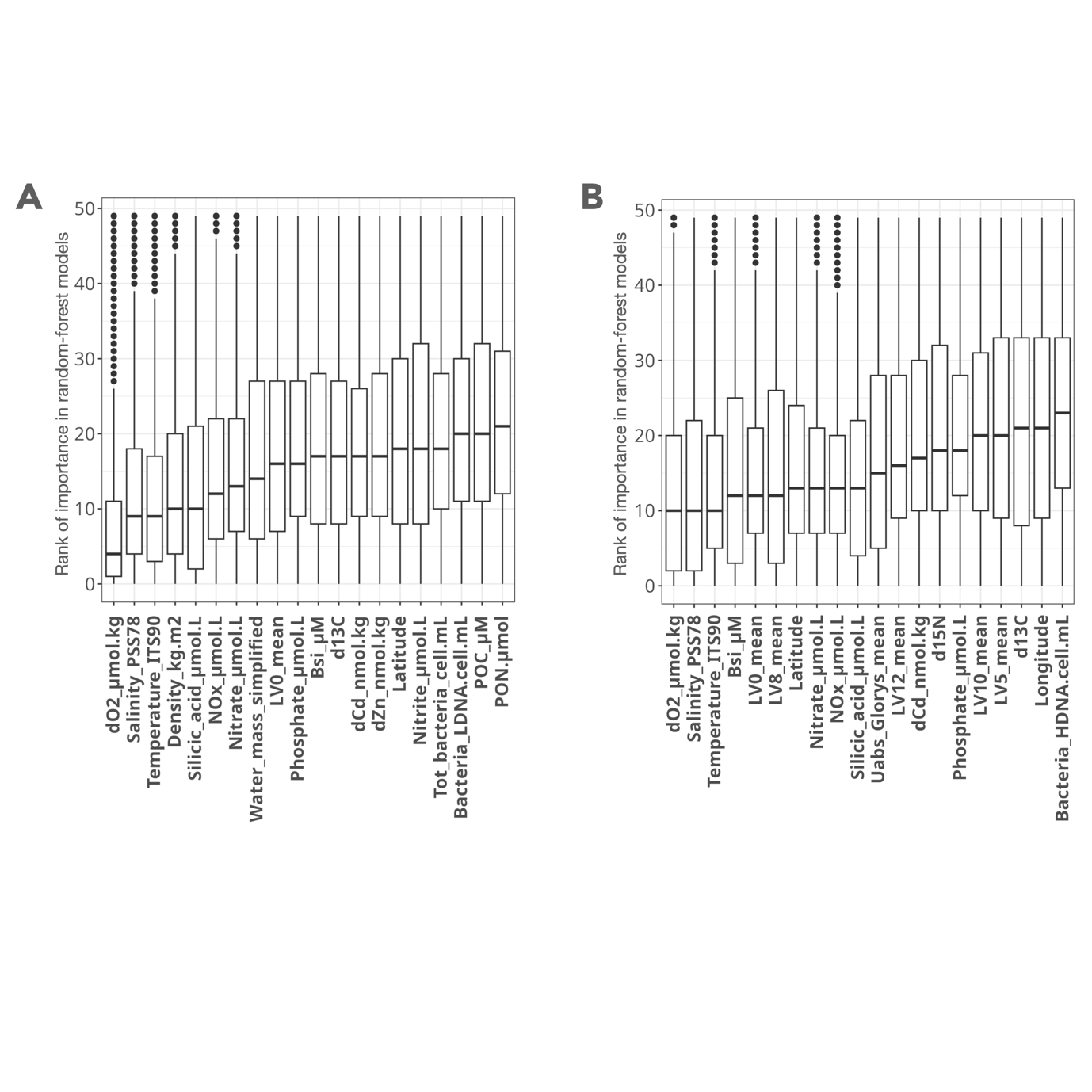
*Table S1: List of resources made publicly available through this project.*



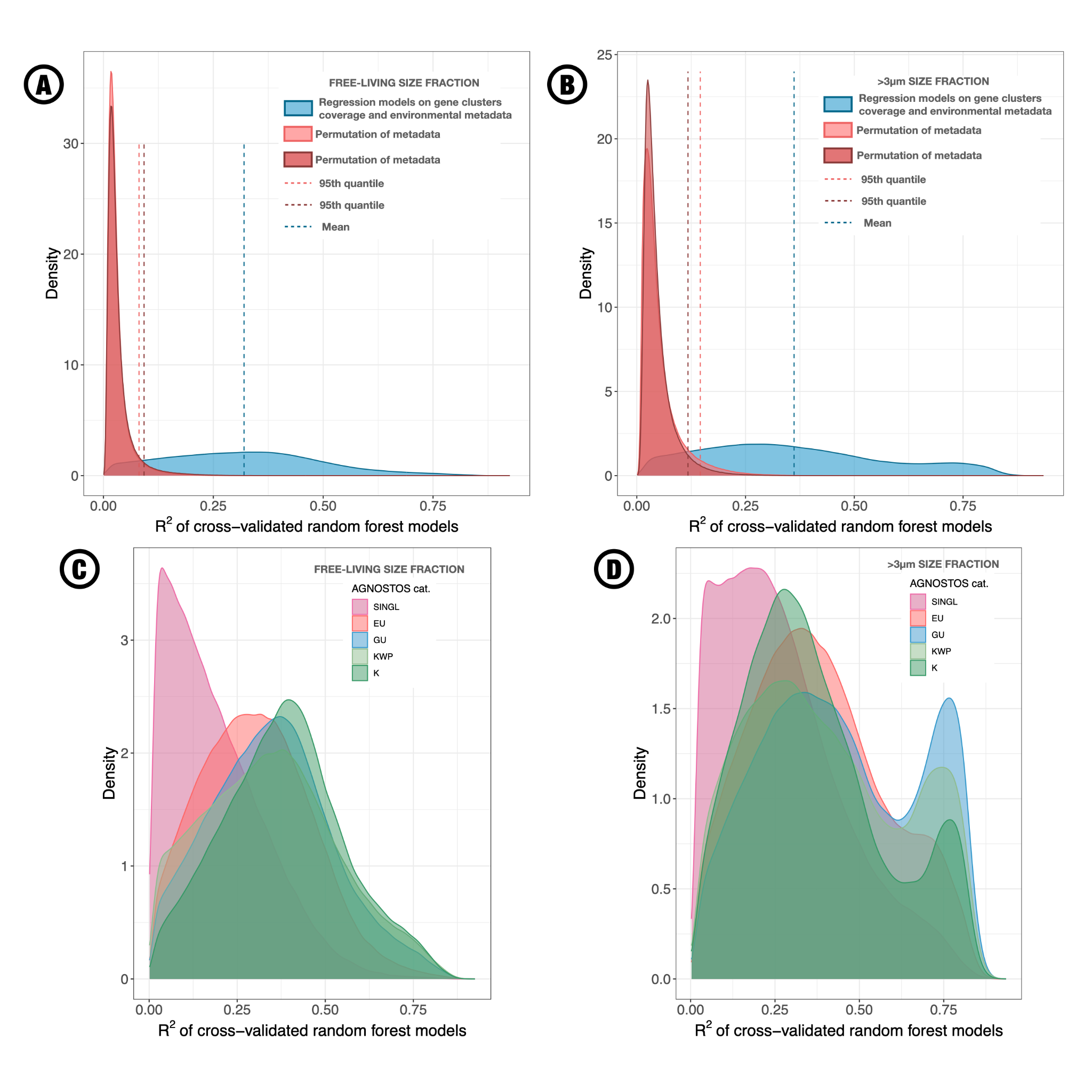
*Figure S1: Diagram of bioinformatics steps implemented to produce AGC and CAG-level matrices from raw metagenomic reads. Steps 1. to 4. are implemented in our NOEMIE Nextflow pipeline, now publicly available at* [*https://gitlab.ifremer.fr/bioinfo/workflows/noemie*](https://gitlab.ifremer.fr/bioinfo/workflows/noemie)*.*



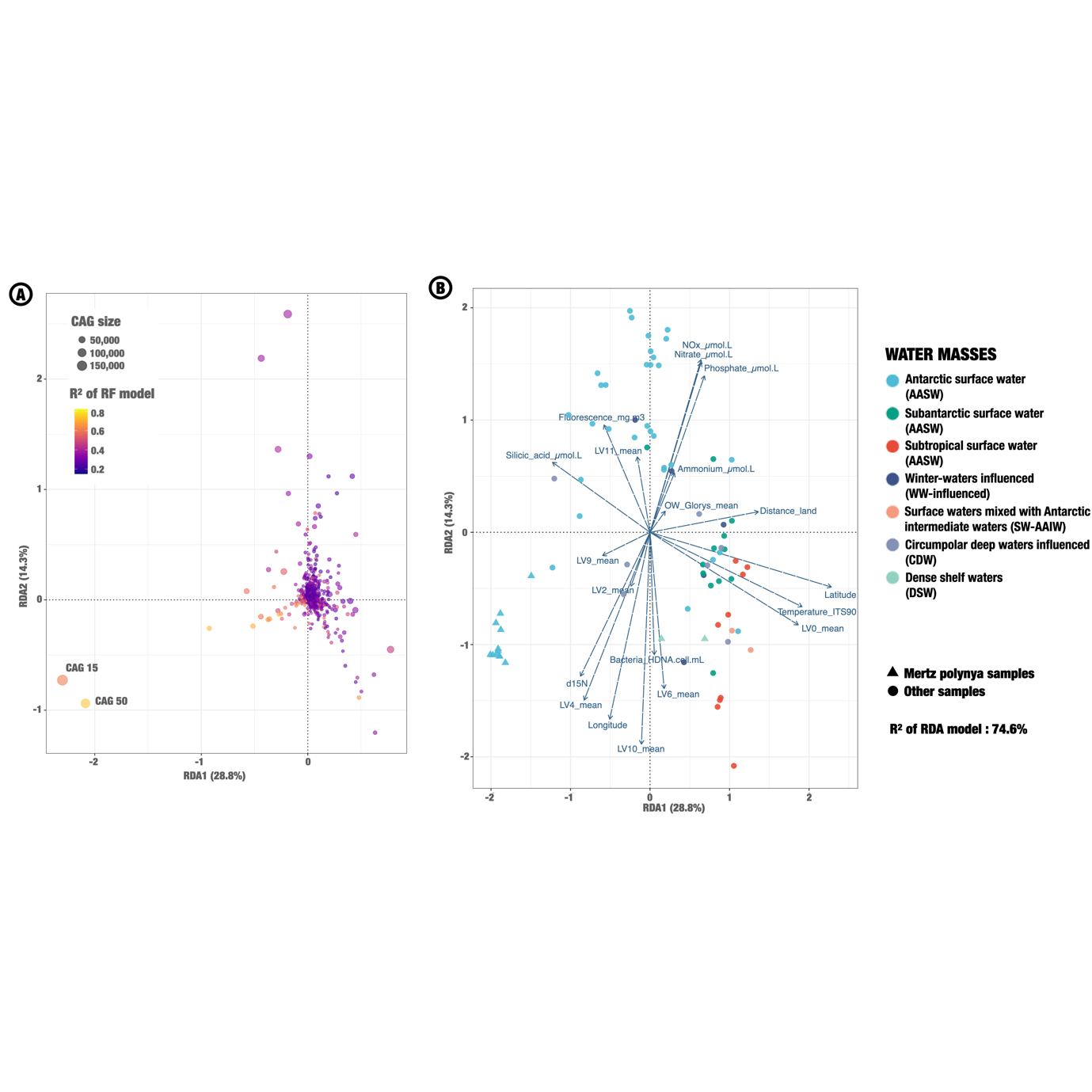
*Figure S2: AGNOSTOS gene clusters (AGC) based collector curves across the 218 ACE metagenomes, drawn using different thresholds on detection. Detection is the proportion of nucleotide in a given ORF mapped with a coverage of at least 1X. As described in the methods, AGC-level detection in a given sample 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.*



*Figure S3: 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 free-living (A) and >3µm (B) size-fractions. Ranked importance reflects the position of importance in predictions, from 1 (the most important predictor) to 50 (the least important).*



*Figure S4: 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 free-living (A) and >3µm (B) size-fractions. Cross-validation R2 values ranged from 0.2% to 92.5% (mean = 32.0%, median = 31.3%) across 1,906,624 models for the free-living size fraction, and from 0.3% to 93.5% (mean = 36.1%; median = 33.4%) across 2,437,988 models for the >3 µm 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 free-living size fraction (resp. >3µm), 95% of models built on permuted data showed R2 values below 8.1% and 9.2% (resp. 11.8% and 14.6%) in each permutation.*



*Figure S5: Redundancy analysis of CAGs abundance in response to environmental metadata in the >3µm 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 free-living 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 ponds1. Our results point towards a dominance of autotrophic ammonia oxydizing micro-organisms in the deep Southern Ocean, which has been described before through amplicon data2. No CAG from the >3µm 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 >3µm 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 >3µm 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 free-living 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 free-living 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.

*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 strong3 (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 Emiliania 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. 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).

2. 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).

3. 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).