nature microbiology

ARTICLES

https://doi.org/10.1038/s41564-021-00979-9

Check for updates

Compendium of 530 metagenome-assembled bacterial and archaeal genomes from the polar Arctic Ocean

Marta Royo-Llonch[1], Pablo Sánchez[1], Clara Ruiz-González[1], Guillem Salazar[2], Carlos Pedrós-Alió[3], Marta Sebastián[1], Karine Labadie[4], Lucas Paoli[2], Federico M. Ibarbalz[5,6], Lucie Zinger[5], Benjamin Churcheward[7], Tara Oceans Coordinators*, Samuel Chaffron[7,8], Damien Eveillard[7,8], Eric Karsenti[5,8,9], Shinichi Sunagawa[2], Patrick Wincker[8,10], Lee Karp-Boss[11], Chris Bowler[5,8] and Silvia G. Acinas[1]✉

The role of the Arctic Ocean ecosystem in climate regulation may depend on the responses of marine microorganisms to environmental change. We applied genome-resolved metagenomics to 41 Arctic seawater samples, collected at various depths in different seasons during the *Tara* Oceans Polar Circle expedition, to evaluate the ecology, metabolic potential and activity of resident bacteria and archaea. We assembled 530 metagenome-assembled genomes (MAGs) to form the Arctic MAGs catalogue comprising 526 species. A total of 441 MAGs belonged to species that have not previously been reported and 299 genomes showed an exclusively polar distribution. Most Arctic MAGs have large genomes and the potential for fast generation times, both of which may enable adaptation to a copiotrophic lifestyle in nutrient-rich waters. We identified 38 habitat generalists and 111 specialists in the Arctic Ocean. We also found a general prevalence of 14 mixotrophs, while chemolithoautotrophs were mostly present in the mesopelagic layer during spring and autumn. We revealed 62 MAGs classified as key Arctic species, found only in the Arctic Ocean, showing the highest gene expression values and predicted to have habitat-specific traits. The Artic MAGs catalogue will inform our understanding of polar microorganisms that drive global biogeochemical cycles.

The Arctic is under increasing pressure from climate change and growing interests in economic opportunities (for example, natural resources like oil and gas, and tourism)[1]. Microorganisms are foundational to the marine food web; thus, we need to understand how they adapt and thrive, as well as forecast their fate in a future ocean impacted by anthropogenic change. In addition, the predicted invasion of the Arctic Ocean by species from lower latitudes due to temperature increases might alter the dynamics of the entire marine ecosystem, from microbes to large animals[2].

As an ecosystem, the Arctic Ocean is subject to extreme seasonal variations (that is, solar radiation, ice cover, temperature) and receives large inputs of fresh water rich in dissolved organic material from rivers and inflowing waters from the Pacific and Atlantic Oceans[3]. Thus, organisms inhabiting the upper water column have to adapt to a highly dynamic environment. Photosynthetic primary production occurs mostly during spring and summer with blooms forming under the ice cover and in the marginal ice zone[4]. Such blooms trigger a succession of bacterial populations, mostly heterotrophs from the phyla Bacteroidetes and Proteobacteria[5]. During winter, the lack of light makes productivity almost negligible, resulting in very low vertical carbon export from the surface layers[6]. Since photosynthesis is limited, heterotrophic bacteria and protists become the dominant players in the ecosystem[7]. During the polar night, prokaryotic mixotrophs and chemolithoautotrophs[8–11] increase in importance.

Geographically, the Arctic Ocean has been divided into eight regions of ecological significance[12]. The record of microbial diversity in most regions is generally limited to a few surveys and is largely based on PCR amplicon sequencing and other molecular approaches[13–16]. Other studies have also attempted to study functions such as nitrification[10,17], heterotrophy[18,19] or photoheterotrophy[20,21]. The uniqueness of polar microbial communities is now recognized[22,23] and recent technological advances, such as the reconstruction of genomes from metagenomes, allow one to go beyond the community level and explore the functional capabilities of specific taxa[23–25]. Nevertheless, our understanding of the complex Arctic ecosystem is limited by the lack of a thorough analysis of key active microbial players including their habitat range and metabolic potential in the Arctic Ocean.

[1]Department of Marine Biology and Oceanography, Institut de Ciències del Mar, Barcelona, Spain. [2]Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, Eidgenössische Technische Hochschule Zürich, Zürich, Switzerland. [3]Systems Biology Program, Centro Nacional de Biotecnología, Madrid, Spain. [4]Genoscope, Institut de Biologie François-Jacob, Commissariat à l'Energie Atomique, Université Paris-Saclay, Evry, France. [5]Institut de biologie de l'Ecole normale supérieure, Ecole normale supérieure, Centre National de la Recherche Scientifique, Institut National de la Santé et de la Recherche Médicale, Paris Sciences & Lettres Université Paris, Paris, France. [6]Universidad de Buenos Aires - CONICET, Centro de Investigaciones del Mar y la Atmósfera (CIMA), Buenos Aires, Argentina. [7]Laboratoire des Sciences du Numérique de Nantes, Centre National de la Recherche Scientifique, Université de Nantes, Nantes, France. [8]Research Federation (FR2022) Tara Oceans GO-SEE, Paris, France. [9]Directors' Research European Molecular Biology Laboratory, Heidelberg, Germany. [10]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France. [11]School of Marine Sciences, University of Maine, Orono, ME, USA. *A full list of members and their affiliations is present at the end of the paper. ✉e-mail: sacinas@icm.csic.es

The *Tara* Oceans Polar Circle expedition[26] circumnavigated the Arctic Ocean performing a holistic survey of its marine microbial diversity. In this study, we built 3,550 genomic bins using the 41 prokaryote-enriched metagenomes from photic to mesopelagic layers. These bins collectively constitute a large fraction of the Arctic prokaryotic diversity detected by metagenomics and metatranscriptomics. Within the pool of 3,550 genomic bins, 530 are of medium or high quality[27] (hereafter referred to as metagenome-assembled genomes (MAGs)). We performed an exhaustive Arctic ecogenomic analysis to study the most relevant uncultured Arctic prokaryotic MAGs, exploring their gene expression patterns, habitat preferences and metabolic potential. We identified key polar species by selecting those MAGs found exclusively in polar metagenomes and highly transcribed, to serve as a baseline for future monitoring of the state of the Arctic Ocean.

Results

Co-assembly and trends of prokaryotic Arctic bins. The 41 metagenomes (size fraction from 0.22 to 3 µm) cover a broad range of environmental and spatio-temporal conditions in the Arctic (Fig. 1a,b). We considered three different ocean layers (surface, subsurface chlorophyll maximum (SCM) or deep chlorophyll maximum (DCM) and mesopelagic) in five Arctic Ocean regions (four stations in the Atlantic Arctic, five stations in the Kara-Laptev Sea, four stations in the Pacific Arctic, one station in the Arctic Archipelago and four stations in the Baffin Bay and Davis strait) and two stations in the sub-Arctic North Atlantic. The sampling period encompassed spring, summer and autumn conditions with a wide range of water temperatures (from −1.7 to 11.1 °C), sea ice conditions and photoperiods.

We applied a new assembly strategy aimed at obtaining a less redundant set of bins with higher genome completeness. This method co-assembles samples that are most similar in their community composition (assessed with 16S miTags and non-metric multidimensional scaling (NMDS) with 100 iterations and a stress value of 0.08), followed by binning together all resulting contigs (Extended Data Fig. 1). This strategy produced 3,550 bins. According to a genome-based taxonomic classification[28], 1,834 bins were classified as bacteria and 146 as archaea. The remaining bins (1,570) could either be of eukaryotic or viral origin, or could not be classified due to a lack of single-copy core genes.

The complete set of 3,550 bins recruited 43.3% of Arctic metagenomic reads and 35.1% of North Atlantic metagenomes (Fig. 1d). In turn, a subset of 725 Arctic bins that fulfilled the quality standards used by Delmont et al.[29], recruited 23% of Arctic metagenomic reads (Supplementary Fig. 1). This is a threefold difference compared to the 6.84% read recruitment by the 892 MAGs generated in the Delmont et al.[29] study with *Tara* Oceans metagenomes, which excluded Arctic sampling[29]. Such differences could be due to the lower diversity reported in polar prokaryotic communities, compared to those from the temperate ocean[30].

Interestingly, mean metagenomic recruitments in the Arctic's mesopelagic were lower than in the photic layer (Fig. 1d), probably indicating that we are missing genomes from deep Arctic waters. In addition, the mean metagenomic read recruitment increased with depth in the temperate and Southern Ocean metagenomes, suggesting that some Arctic genomes may reach the mesopelagic layers of other latitudes through ocean circulation[22,31]. To obtain biogeographical patterns at a global scale, we used metagenomes and metatranscriptomes collected in all the oceanographic regions sampled by the *Tara* Oceans expedition[23]. Metatranscriptomic read recruitment in the photic layers of the Southern Ocean was fourfold that of temperate samples, suggesting a polar preference of certain bins and confirming bipolar expression patterns (Fig. 1d).

We detected a positive correlation between metagenomic and metatranscriptomic read recruitments by Arctic bins. Similar correlations have been found at the gene level in the eastern subtropical Pacific Ocean's microbiome[32], suggesting that, as may be expected, expression profiles depend on gene abundance[33]. The strength of the correlation decreased with depth and was stronger in polar samples compared to temperate latitudes (Supplementary Fig. 2), which could be associated with genomes that have been exported vertically from the photic zone and/or transported by deep ocean currents to/from more temperate latitudes. This result could be affected by a higher species richness of temperate mesopelagic waters compared to the Arctic, as well as differences in microeukaryote diversity[30]. Individual metatranscriptomic recruitments tended to be lower than metagenomic recruitments in temperate latitudes (Supplementary Fig. 2), suggesting a resting stage of polar prokaryotic cells in transit between polar habitats[34,35]. These results reinforce the polar habitat preference of a significant fraction of our *Tara* Arctic genomic dataset.

Following published quality thresholds[27], the 3,550 bins were classified into 3 quality groups based on genome completeness and quality values (Fig. 1b and Supplementary Table 1): 96 high-quality bins (manually curated with ≥90% completeness and <5% contamination); 434 medium-quality bins (with ≥50% completeness and <10% contamination); and 2,642 low-quality bins. The 530 high- and medium-quality bins with sufficient quality ratings were designated MAGs and are presented in this study as the Arctic MAGs catalogue. The MAGs catalogue reflects diversity across three seasons, late spring, summer and autumn, representing the most comprehensive resource of uncultured prokaryotic genomes from the Arctic Ocean to date.

Diversity, novelty and abundance of the Arctic MAGs catalogue. The Arctic MAGs catalogue consists of a high diversity of non-redundant MAGs. Only 8 combinations (0.006%) of the 140,185 genome pairs could be considered as closely related species, showing average nucleotide identities (ANIs) >96%[36] (Fig. 1c). Collectively, our analyses indicate that the Arctic MAGs represent consensus bacterial and archaeal genomes of 526 non-redundant species. Genomes were annotated using a genome-based phylogeny approach and the Genome Taxonomy Database (GTDB)[28,37].

Assembling conserved genes, such as the ribosomal operon, is a common challenge in MAG reconstruction[38]. In our study, only 27 MAGs (5%) contained full or partial 16S ribosomal RNA genes (Extended Data Fig. 2) that could not be annotated further than family level, reflecting that an important fraction of the microbial diversity in this ecosystem may have been consistently missed in previous studies. Therefore, we assessed their taxonomic annotation and novelty through a phylogenomics approach against a database that includes both cultured and uncultured taxa[28]. The Arctic MAGs catalogue included 472 bacteria and 58 archaea assigned to 21 known phyla (Fig. 2a), with >83% of unclassified genomes at the species level (Fig. 2b). This species novelty percentage is similar to the 81% found in a recent study that reconstructed MAGs from the Baltic Sea[39], calculated using the same methodology of our study. In the *Tara* Oceans MAG dataset[29], 44% of MAGs could not be annotated to any known species in the GTDB. This value, however, could be underestimated since the GTDB includes MAGs generated in a previous study from the *Tara* Oceans temperate metagenomes[40] (see the Supplementary Information on the taxonomical classification of Arctic MAGs for more details).

A significant positive correlation (two-sided Pearson's product-moment correlation) between whole-genome metagenomic and metatranscriptomic read recruitment in Arctic samples (Supplementary Fig. 3) was strong ($r > 0.7$) in 35% of the identified phyla (Crenarchaeota, Bacteroidota, Latescibacterota, Marinisomatota, Planctomycetota, Proteobacteria and Verrucomicrobiota), moderate (r between 0.5 and 0.7) in 15% of phyla (Acidobacteriota, Gemmatimonadota and Actinobacteriota) and weak (r between 0.3 and 0.5) in 15% of the detected phyla (Thermoplasmatota, Chloroflexota and Myxococcota).

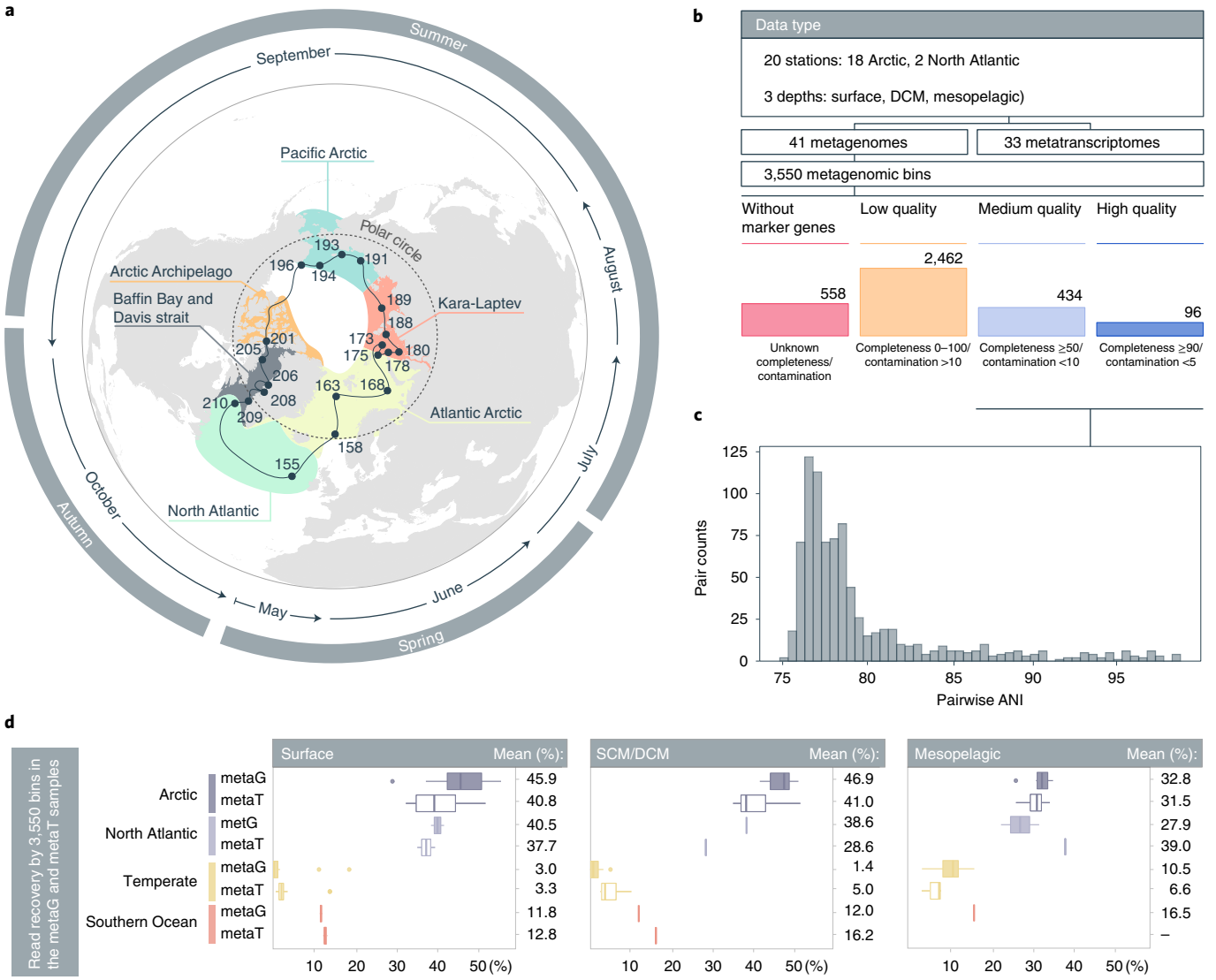


Fig. 1 | Metagenomic genome reconstruction of the *Tara* Oceans Polar Circle expedition. a, *Tara*'s trajectory and the stations at which metagenomes and metatranscriptomes were collected are shown. The coloured areas highlight the sampled regions: five Arctic regions and the sub-Arctic North Atlantic. The Polar Circle (66° N) is shown with a dashed line. The outer circles show the month and season of sampling during the circumnavigation, starting in May 2013. **b**, Outline of the polar metagenomics and metatranscriptomics dataset, the number of bins assembled from the metagenomic samples and their quality-based classification, measured by combining genome completeness and contamination. Only those 530 bins of medium and high quality were designated as MAGs. **c**, Pairwise ANI comparisons of 530 medium- and high-quality MAGs, showing that only 8 pairs could be considered the same species (ANI > 96%). **d**, Distribution of metagenomic (metaG; filled box plots) and metatranscriptomic (metaT; empty box plots) read recovery per sample by all reconstructed bins per sample. $n = 3{,}550$ bins examined over 68 metagenomic samples (37 samples from the *Tara* Oceans Polar Circle, 4 samples from the Southern Ocean and 27 from the *Tara* Oceans expedition) and 53 metatranscriptomic samples (33 samples from the *Tara* Oceans Polar Circle, 3 samples from the Southern Ocean and 17 from the *Tara* Oceans expedition). Samples are divided by layer (columns) and latitudinal range (purple boxes for the *Tara* Oceans Polar Circle, yellow boxes for the temperate samples from the *Tara* Oceans expedition and red boxes for the Southern Ocean samples from the *Tara* Oceans expedition). Data are shown as horizontal box plots (Tukey style): the lower (left) and upper (right) hinges correspond to the first and third quartiles (25th and 75th percentiles), the vertical line indicates the median and the whiskers indicate the lowest and highest points within 1.5× the interquartile ranges (IQRs) of the lower (first) or upper (third) quartile, respectively. Data beyond the end of the whiskers are outlying points and are plotted individually. The mean percentage of read recruitments per group of samples is indicated at the right of each plot.

The catalogue is predominated by rare Arctic taxa (Extended Data Fig. 3). The 12 most abundant MAGs, recruiting at least 200 reads per kilobase of genome per gigabase of metagenome (RPKG) belong to unknown species of the Bacteroidota, Actinobacteriota, Alphaproteobacteria, Gammaproteobacteria and SAR324 phyla.

The Arctic MAGs catalogue contains a set of diverse non-redundant Arctic genomes. We found a high degree of taxonomic novelty, with unknown lineages being representative of abundant and rare species in Arctic waters, active in terms of gene expression.

Metabolic potential and genomic expression among Arctic MAGs. *Prevalence of mixotrophy in Arctic prokaryotic genomes.* The greenhouse gas CO_2 is central in the global carbon cycle and the Arctic Ocean is considered a sink for atmospheric CO_2 (ref. [41]). Although primary production in Arctic waters is mainly

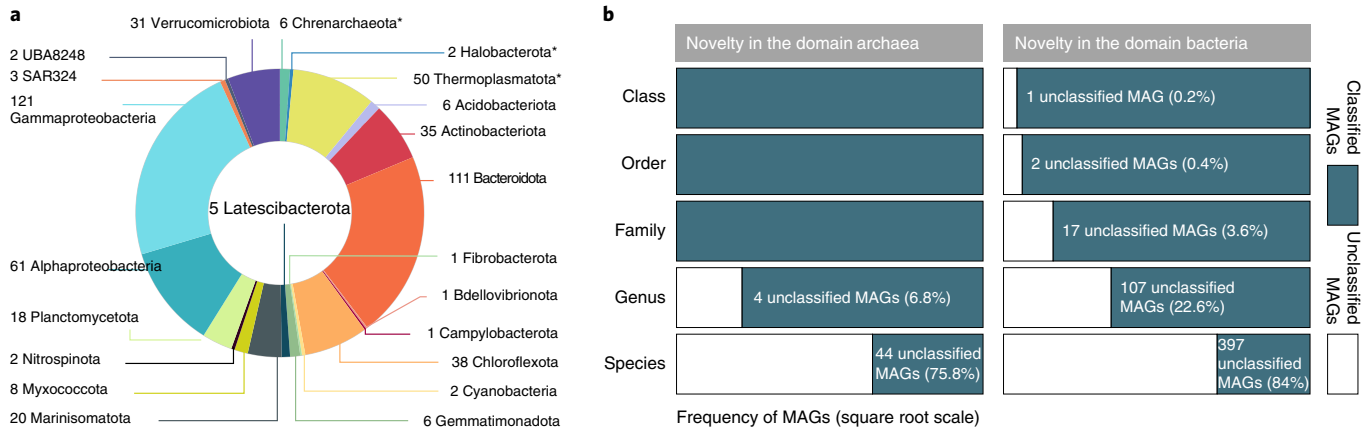


Fig. 2 | Taxonomical annotation and novelty of Arctic Ocean MAGs. a, Phylogenomics-based taxonomic classification of the 530 Arctic MAGs dataset at the phylum level (except for Proteobacteria that were split at the class level). Archaeal phyla are highlighted with an asterisk; annotations without an asterisk belong to the bacteria domain. **b**, Stacked bar plot for novelty quantification of the Arctic MAGs (*x* axis) at different taxonomic ranks (*y* axis). The taxonomically unclassified portion is depicted in white; the taxonomically classified portion is shown in blue. White labelling refers to the unclassified fraction. The frequency of MAGs on the *x* axis is shown on a square root scale.

performed by eukaryotic phytoplankton[12], inorganic carbon fixation by prokaryotes in the dark might be an important process, particularly during the polar night[10,11]. However, the relevance and ubiquity of different inorganic carbon fixation pathways across different Arctic regions, depths and seasons is unknown, as is the identity of the potential key players. In this study, we used a selection of 120 marker genes (Supplementary Table 2) representative of carbon fixation processes and energy metabolism and investigated their transcript abundance throughout the expedition.

Fifteen Arctic MAGs belonging to seven phyla contained RuBisCO (Kyoto Encyclopedia of Genes and Genomes (KEGG) K01601, K01602 or both), or RuBisCO and phosphoribulokinase (K00855) (Supplementary Table 3). Among them, we report RuBisCO-containing MAGs from the bacterial phyla Latescibacterota and UBA8248 (previously Tectomicrobia). Of these, we retrieved 14 RuBisCO large-chain sequences. These could be classified into phylogenetic groups corresponding to the RuBisCO 'forms' I, II, III-a, IV and IV-like defined in previous studies[42,43] (Fig. 3a).

RuBisCO forms I and II are directly involved in the autotrophic CO_2-fixing Calvin–Benson–Bassham pathway and were found in four MAGs. These showed whole-genome expression in the five Arctic regions during all seasons and at all depths. Nevertheless, expression of RuBisCO was found only for MAGs classified into form I (Fig. 3b) and was dominated by the *Synechococcus* MAG in the North Atlantic and Atlantic Arctic photic samples. For the rest of the sampling, the recruited transcripts belonged to Proteobacteria (Fig. 3c,d). Form II was detected in a new member of the family Thioglobaceae but was not transcribed. Since activity of the photosynthetic *Synechococcus* was only detected in the North Atlantic samples, genomic expression of form I RuBisCO MAGs in the Arctic suggests a larger contribution of chemoautotrophic processes.

RuBisCO form III-a was detected in a Crenarchaeota (UBA57 sp., previously known as Thaumarchaeota) and expressed in a mesopelagic late spring sample in the Atlantic Arctic (Fig. 3c,d). Archaea containing RuBisCO form III-a but lacking phosphoribulokinase, like this Arctic MAG, are proposed to be involved in a modified nucleotide scavenging pathway[44].

RuBisCO forms IV and IV-like do not perform CO_2 fixation and may be involved in methionine salvage, sulphur metabolism and D-Apiose catabolism[45]. These were found in eight MAGs (Fig. 3a); their transcription occurs mostly in summer and early autumn photic samples peaking in station TARA_180 in similar magnitude to Cyanobacteria's RuBisCO transcription in the North Atlantic.

These results indicate that at least 28% of RuBisCO-containing MAGs are putative autotrophs (forms I and II), prevalent in the Arctic Ocean and whose genomes are expressed across all regions, depths and seasons. RuBisCO-containing MAGs possessed multiple protein domains (15–134) annotated as ATP-binding cassette transporters suggesting a mixotrophic lifestyle. Mixotrophy is likely in the case of the photosynthetic *Synecochoccus* MAG, a lifestyle that has already been reported in other marine Cyanobacteria[46].

Mixotrophy has also been proposed to be relevant for specific Arctic heterotrophs, which can perform CO oxidation[47]. This process is suggested to serve as a supplemental energy source during organic carbon starvation[47] and is catalysed by carbon monoxide dehydrogenase form I (*cox* genes)[48]. To our knowledge, the potential for CO oxidation by prokaryotes in the Arctic Ocean has never been addressed. We found 114 MAGs (21.5%) transcribing the *coxL* (K03520) gene, of which 9 were expressing the CO-fixing *coxL* form I in all regions and seasons in the photic layer and in the North Atlantic-influenced mesopelagic (Extended Data Fig. 4a).

The transcription of key markers for alternative inorganic carbon fixation pathways, like the 3-hydroxypropionate bicycle (hereafter 3-HP) and the 3-hydroxypropionate/4-hydroxybutyrate cycle (hereafter 3-HP/4-HB), were also detected in certain MAGs; however, considering the lack of complementary genes for carbon fixation, their autotrophic capacity is putative (Supplementary Figs. 4 and 5).

Chemolithoautotrophic potential of Arctic prokaryotic genomes. We investigated the whole-genome recruitment of metatranscriptomic reads of five MAGs that contained reliable markers for the chemolithoautotrophic processes associated to ammonia and nitrite oxidation (Fig. 4). We could identify three ammonia-oxidizing archaea, two annotated as new *Nitrosopelagicus* spp. (containing 31% of the 3-HP/4-HB KEGG module, the full urease complex and, in MAG 1708, the ammonia-monooxygenase coding *amoA*) and one *Nitrosopumilus* sp. (containing 21% of the 3-HP/4-HB KEGG module and the full urease complex). One Alphaproteobacterium (*GCA-2728255* sp.) was classified as an ammonia-oxidizing bacterium and contained the characteristic nitrification marker hydroxylamine oxidoreductase. One *Nitrospinae* species, *LS-NOB* sp., was classified as a nitrite-oxidizing bacterium (NOB), with 35% of reverse tricarboxylic acid cycle module completeness, including ATP citrate lyase, the nitrite oxidoreductase and a nitrate/nitrite transporter. Their expression patterns showed an overall preference for

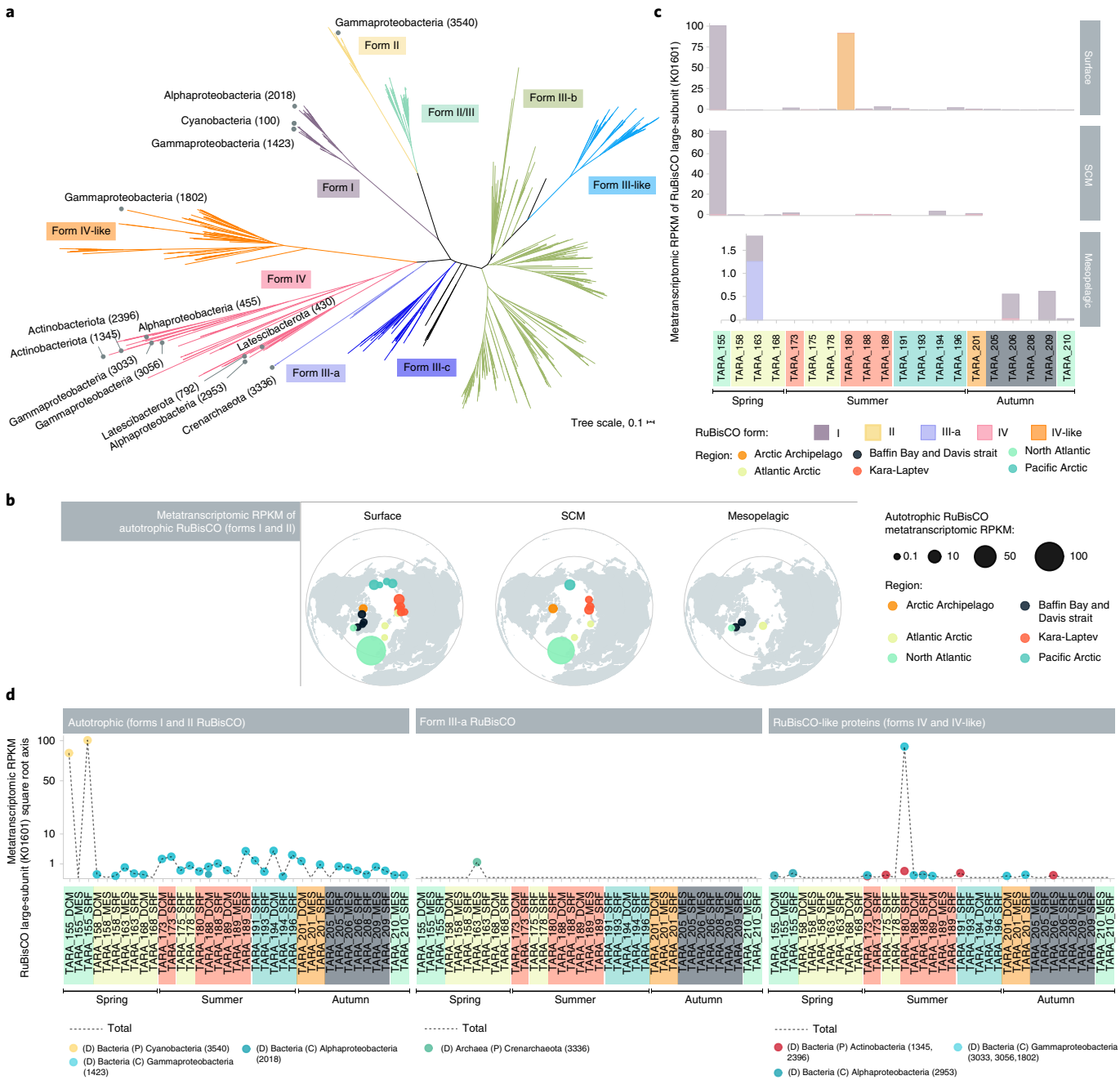


Fig. 3 | Potential autotrophy in RuBisCO-coding MAGs. a, Maximum-likelihood phylogenetic reconstruction of the 15 RuBisCO large-chain (K01601) amino acid sequences found in Arctic MAGs, coloured by RuBisCO form. **b**, Polar maps with the transcript abundance of RuBisCO forms I and II, involved in the Calvin cycle pathway (K01601), colour-coded by Arctic region. The size of the dots is proportional to the accumulated metatranscriptomic RPKMs. **c**, Stacked bar plot of metatranscriptomic RPKMs recruited by the *ruBisCO* gene, coloured by ruBisCO form. **d**, Metatranscriptomic RPKMs of *ruBisCO* genes collapsed by the phylum (or class in the case of Proteobacteria MAGs) of the genome they were found and separated by form. The black dashed line represents the total recruited metatranscriptomic RPKMs by RuBisCO in each sample; the numbers in parentheses in the legend display the MAG identification code.

mesopelagic depths, especially in the North Atlantic-influenced Arctic stations (in spring and autumn). Ammonia-oxidizing archaea and ammonia-oxidizing bacteria are also active in the Kara-Laptev's mesopelagic, the former recruiting more metatranscriptomic RPKGs, in agreement with previous results[17]. NOB expression is restricted to the photic zone in the North Atlantic (spring) and the mesopelagic zone in the Labrador Sea (North Atlantic) during autumn. To our knowledge, the *LS-NOB* sp. MAG is the first NOB found to be active in the region in both photic and aphotic layers.

Therefore, it appears that the set of Arctic MAGs consists of a majority of heterotrophic and mixotrophic organisms, with a few chemolithoautotrophs that are mostly active transcriptionally in the mesopelagic during spring and autumn. Future experimental validation is required to quantitatively confirm the relevance of these processes.

Ecological preferences and biogeographical patterns. The Arctic MAGs were used as reference genomes in the mapping of

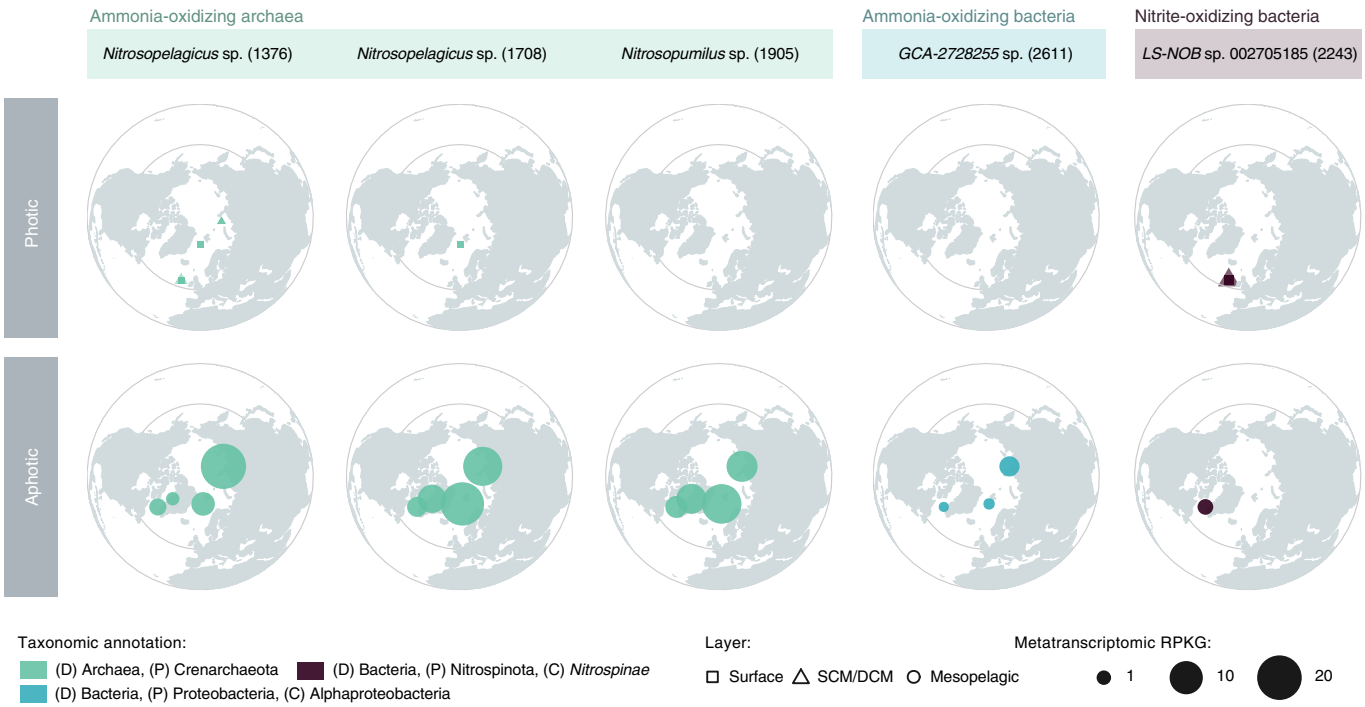


Fig. 4 | Chemolithoautotrophic Arctic Ocean MAGs. Five MAGs contained the specific marker genes to be putative chemolithoautotrophs in the Arctic Ocean. They are classified into ammonia-oxidizing archaea and bacteria and nitrite-oxidizing bacteria. Their metatranscriptomic recruitments in RPKGs is depicted by the size of the dots, while their shape indicates the water column layer. The colour of the dots depends on the taxonomic annotation of each MAG. The numbers in parentheses correspond to the MAG code.

metagenomic reads from 68 samples, representing all the polar and temperate oceanographic regions sampled by *Tara* Oceans. Bray–Curtis dissimilarities of MAG abundance composition showed that polar and temperate samples clustered separately (Fig. 5a) (NMDS with 100 iterations and 0.8 stress value, permutational multivariate analysis of variance (MANOVA) $R^2 = 0.14$, $P < 0.001$). This pattern is consistent with the 16S miTAG-based clustering of the global *Tara* Oceans. Within polar and non-polar samples, MAG assemblages were significantly structured by depth (permutational MANOVA $R^2 = 0.14$, $P < 0.001$) (Fig. 5a) in agreement with the reported vertical stratification of marine microbial communities[16,22,49]. This confirms both the unique diversity of prokaryotes in the Arctic and the presence of bipolar taxa, previously described in surveys based on PCR amplicon sequencing[22,50].

Even though non-detection of a taxon in metagenomic studies cannot be directly translated to its absence in the environment, we delineated the geographical distribution of individual MAGs using metagenomes representative of the global ocean and a stringent read mapping filtering of at least 20% of genome coverage. We found 153 MAGs (28.9%) detected exclusively in Arctic metagenomes and 23 (4%) showing a bipolar distribution (that is, recruiting reads only from the Arctic and Southern Ocean) (Fig. 5b). The somewhat higher proportion of bipolar MAGs compared to the 15% retrieved by 16S rRNA sequencing in a previous study[22] can be attributed to either methodological differences or the difference between sampling years. The bipolar subset of MAGs showed less diversity of phyla than other biogeographical categories, which is consistent with latitudinal diversity gradients[30], and lacked MAGs representative of Actinobacteriota and Verrucomicrobiota found in every other studied latitude (Fig. 5b).

On a pan-Arctic scale, we found that 4.9% of MAGs (26) are found in all five Arctic regions at high occupancy (≥90% occurrence in Arctic stations), of which only 1 is exclusively found in latitudes above the Arctic Polar Circle (MAG 2328, an unknown *Ascidiaceihabitans* sp. Alphaproteobacteria). Pan-Arctic prokaryotes in the catalogue are represented by Alphaproteobacteria and Bacteroidota, including a reduced number of Gammaproteobacteria and Actinobacteriota. Additionally, taxonomic diversity increases in the set of MAGs with a very limited distribution among Arctic regions (Extended Data Fig. 5a). We found that the Atlantic Arctic sampled during spring and the Baffin Bay and Davis strait region sampled during autumn, had the highest numbers of MAGs with a highly restricted distribution, detected only in one of these two regions (Extended Data Fig. 5b). It is important to note that the definition of pan-Arctic MAGs in this study can be biased towards taxa prone to thrive during spring and summer (the seasons when most of the sampling occurred), overlooking prokaryotes dominant during autumn and winter.

Nearly 60% of our MAG dataset is represented by polar-specific genomes (Fig. 5b). Genomes with an Arctic-only distribution were estimated to be significantly larger (2.9 mega base pairs (Mbp) on average) than those present in temperate latitudes (2.5 Mbp) (Dunnett–Tukey–Kramer pairwise multiple comparison test $P < 0.05$) (Fig. 5c). However, we did not find significant differences between their coding densities. These also showed lower potential minimum generation times and optimal growth temperatures (Dunnett–Tukey–Kramer pairwise multiple comparison test $P < 0.05$ in both analyses) (Extended Data Fig. 6a,b). Faster growth in environments with high resource availability (like polar regions during spring and summer) has been related to larger genomes and higher numbers of ribosomal gene copies[51]. Even though we did not find differences in ribosomal operon copy numbers between biogeographically distinct MAGs, due to their completeness values and the limitations of resolving ribosomal regions in de novo assemblies, we found that community-wide metagenomes from polar latitudes had higher numbers of rRNA operon copies than metagenomes from lower latitudes (Extended Data Fig. 6c).

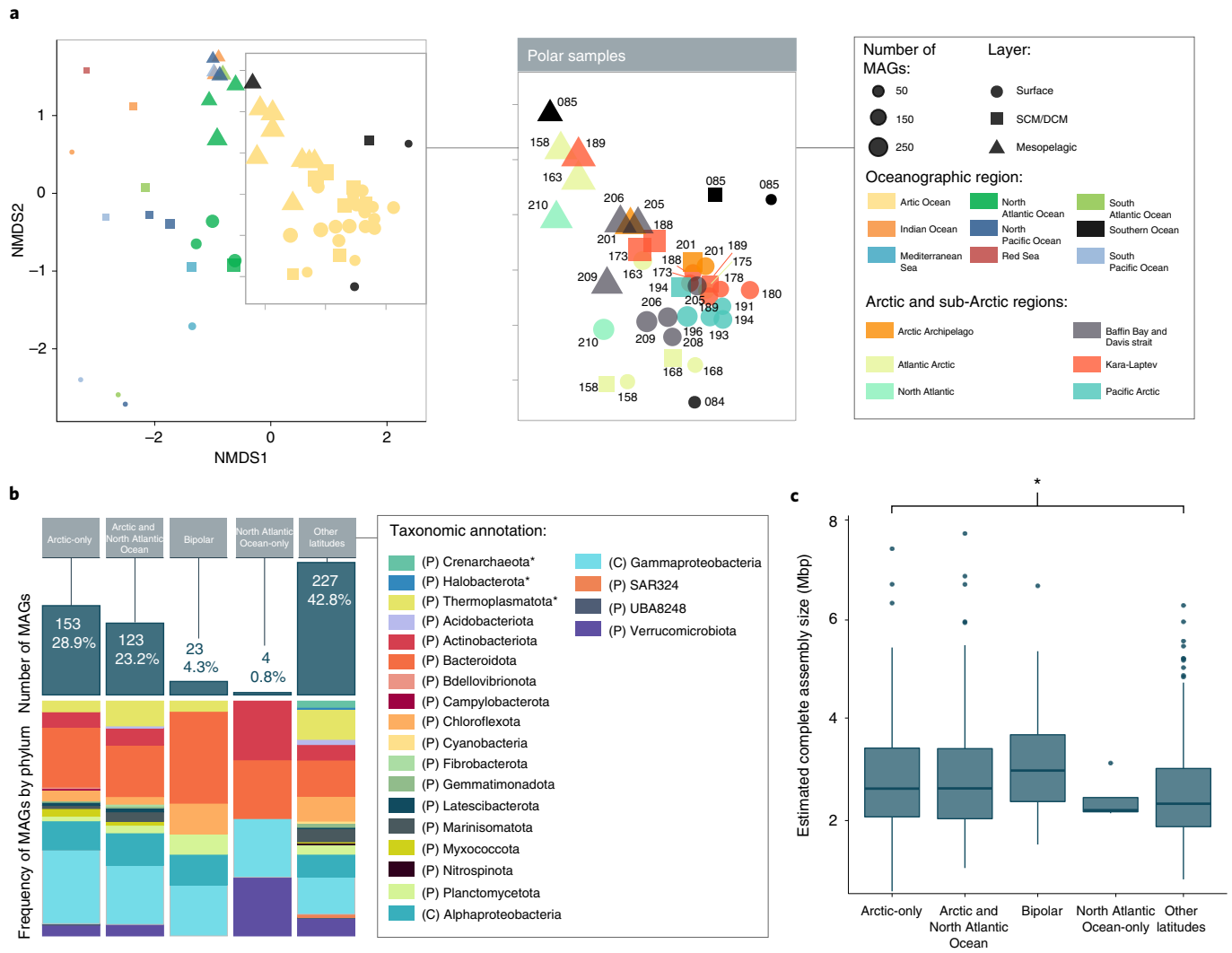


Fig. 5 | Composition and biogeography of the 530 Arctic microbial MAGs. a, NMDS ordination of metagenomic samples based on their composition of 530 Arctic MAGs. The shape defines the sample's layer in the water column (surface, DCM, mesopelagic) and the dot size represents the MAG richness (that is, the number of different MAGs) of the sample. The plot on the left shows all non-polar and polar (inside the square) samples, coloured by oceanographic region. Oceanographic regions are: Artic Ocean, Indian Ocean, Mediterranean Sea, North Atlantic Ocean, North Pacific Ocean, Red Sea, South Atlantic Ocean, Southern Ocean and South Pacific Ocean. The middle and right panels represent the same NMDS ordination of only polar samples colour-coded by season or Arctic/sub-Arctic region, respectively, and labelled with their station number. **b**, Biogeographical categorization of the 530 Arctic MAGs. The stacked bar plots represent the number of MAGs in each category, coloured by taxonomic annotation; the top bars represent the percentage within the medium- and high-quality Arctic MAGs dataset. **c**, Differences between complete assembly sizes from MAGs classified by their biogeographical categories ($n = 153$ MAGs classified as Arctic-only, 123 MAGs classified as Arctic and North Atlantic Ocean, 23 MAGs classified as bipolar, 4 MAGs classified as North Atlantic Ocean only and 227 MAGs classified as other latitudes). Data are shown as horizontal box plots (Tukey style): the lower (left) and upper (right) hinges correspond to the first and third quartiles (25th and 75th percentiles), the vertical line indicates the median and the whiskers indicate the lowest and highest points within 1.5× the IQRs of the lower (first) or upper (third) quartile, respectively. Data beyond the end of the whiskers are outlying points and are plotted individually. Statistical support was calculated using the two-sided Dunnett–Tukey–Kramer pairwise multiple comparison test adjusted for unequal variances and unequal sample sizes (Dunnett–Tukey–Kramer) and 95% CIs. The Dunnett–Tukey–Kramer test shows significant differences ($P < 0.05$) between MAGs specific to the Arctic and MAGs present at lower latitudes.

In summary, about 30% of our MAGs were exclusively present in Arctic regions. Their larger genomes, overall smaller minimum generation times and the increased numbers of ribosomal operon copies in polar metagenomes suggest that Arctic-specific MAGs could be associated with a copiotrophic lifestyle.

Disentangling generalist and specialist Arctic MAGs. We defined two subsets of MAGs based on their niche breadth: habitat generalists, evenly distributed; and habitat specialists, with an uneven distribution[52,53]. The latter are thought to be more sensitive to changes in environmental conditions[54] since they might have narrow environmental requirements. Generalists, on the other hand, are less dependent on environmental conditions, have a wide habitat tolerance and high functional plasticity[55]. In the current scenario of climate change, it is essential to identify which Arctic species may be more susceptible to environmental alterations.

We calculated the niche breadth of MAGs based on their abundance across the Arctic metagenomic dataset using the Levins niche breadth index[52] with random permutations, providing statistical support for the classification of MAGs into generalists or specialists

since it considers potential biases of metagenomic sequencing analyses[56]. For this study, each Arctic sample was considered as an individual habitat. Most MAGs (71%) could not be categorized into generalists or specialists, while 21% ($n = 111$) were habitat specialists and 7% ($n = 38$) were generalists (Extended Data Fig. 7a). The high contributions of specialists have been reported in other polar environmental extremes, such as coastal Antarctic lakes[57], or in highly productive marine sites compared to oligotrophic open ocean stations[58].

Since habitat generalists are likely to adapt to a broader range of habitats[55], we expected their complete genome size to be larger than that of habitat specialists. This difference was apparent but not statistically significant in the median genome size of MAGs that showed Arctic and North Atlantic distributions (Dunnett–Tukey–Kramer pairwise multiple comparison test; Extended Data Fig. 7b). Overall, the specialist MAG genome size was significantly lower than those of uncategorized MAGs (Dunnett–Tukey–Kramer pairwise multiple comparison test $P < 0.05$; Extended Data Fig. 7c).

While generalists were assigned to the bacteria phyla Actinobacteriota, Proteobacteria, Bacteroidota and Myxococcota (Extended Data Fig. 7b), specialists displayed a larger taxonomic diversity, including the archaeal phyla Thermoplasmatota and Crenarchaeota.

Interestingly, specialist and generalist MAGs recruit similar RPKGs in photic samples. In contrast, recruitment in the mesopelagic was much higher for specialists (Extended Data Fig. 7c, note the change of scale in the *y* axis). This difference might be explained by the wider range of estimated optimal growth temperatures of specialists compared to generalists (Extended Data Fig. 6e), as well as nutrient availability and niche compartmentalization in deeper waters, in contrast with the wider gradients in nitrate, temperature and salinity of the upper Arctic Ocean (Supplementary Table 4). Even though we did not find differences between specialists and generalists in relation to in situ seawater temperature (Extended Data Fig. 8a), we found that in the mesopelagic layers, generalist MAGs had maxima of metagenomic RPKG in warmer stations than specialists (Extended Data Fig. 8b) located in the Baffin Bay and Davis strait region. In the photic layers, temperature ranges were similar between both niche breadth groups (Extended Data Fig. 8b).

The comparison of the genetic content between niche breadth groups showed that all KEGG-annotated genes found in generalists were found in specialists, whereas 814 genes from specialists were not detected in generalists. This difference might result from the higher number of specialists, a higher functional redundancy in generalist MAGs, the incomplete nature of MAGs or a combination of all. Genes found only in specialists were typical of the aphotic ocean and were enriched functions related to genetic processing, energy metabolisms and environmental information processing (Extended Data Fig. 9).

Since the abundance or expression of microbial generalist or specialist MAGs as a whole could not be explicitly linked to any of the environmental variables tested (Supplementary Fig. 7), it is likely that community turnover in polar communities (suggested to drive changes in community gene expression in response to ocean warming[23]) could also transcend niche breadth.

Polar ocean key prokaryotic species. To define key prokaryotic species specific to polar regions and whose existence may be threatened by the expected changes in the polar environment, we examined MAGs that were detected only in polar metagenomes and showed higher genetic expression within their niche breadth category per sample. A total of 62 MAGs fell within these criteria (Fig. 6) (7 generalists, 25 specialists and 30 uncategorized), representing potential ecologically relevant taxa in the polar ecosystem that we advocate monitoring as a means to assess the health status of the Arctic Ocean.

For example, *Polaribacter* is one of the most common genera in polar waters and its bipolar distribution has been described previously[59]. It showed the highest number of MAGs with a bipolar distribution and highest expression in photic samples (Fig. 6). *Polaribacter* MAGs were assigned to both generalists and specialists. Interestingly, *Polaribacter* is predicted to be an ecologically central species in a polar cross-kingdom interactome, being one of the highly connected taxa[60] and ranking high according to the general keystone index (keystone index rank = 86 of 4,000)[61]. Another heterotrophic *Flavobacterium* (UA16 family) dominated gene expression in the mesopelagic, together with a MAG from the Myxococcota family UBA4427 (Fig. 6). Photic generalists were mostly heterotrophic but we also found a generalist annotated as Myxococcota thriving in the mesopelagic with a putative autotrophic metabolism.

The genetic expression of key polar specialists was not dominated by any particular taxonomic group (Fig. 6). We found potential for autotrophy (Calvin–Benson–Bassham cycle) in a Gammaproteobacterium (Methylophilaceae) and the 3-HP cycle in a new Gemmatimonadetes species and two Alphaproteobacteria (Rhodobacteraceae). Most metatranscriptomic recruitments in photic summer samples by key specialists were associated with Gammaproteobacteria (heterotrophic with denitrifying potential) and Alphaproteobacteria. In contrast, spring and autumn photic samples showed the highest expression values from Flavobacteria and Verrucomicrobia. In the mesopelagic, the specialists with the highest gene expression were MAGs from the phyla Verrucomicrobiota (spring), Chloroflexi (summer) and Marinisomatota (autumn).

For those MAGs uncategorized in terms of niche breadth, we found higher expression in surface waters through spring and summer associated with (1) those showing heterotrophic metabolism and (2) potentially chemolithoautotrophic or mixotrophic MAGs (Chloroflexi and Alpha- and Gammaproteobacteria). During these times, Thalassoarchaeaceae MGIIb and Chloroflexota were the most active in the mesopelagic. Autumn photic samples were clearly dominated by archaea MAGs from the family Poseidoniaceae MGIIa. *Oceanicoccus* Gammaproteobacteria were predominantly active in the SCM and a new family of Planctomycetota was most expressed in the mesopelagic (Fig. 6).

Overall, we uncovered a pool of 62 key polar MAGs (11% of the *Tara* Arctic MAG dataset), of which 7 were classified as habitat generalists and 25 as habitat specialists. While key polar generalists seem to be mostly heterotrophic, key polar specialists display a wider variety of metabolic markers, including autotrophic potential and denitrifying genes.

Conclusions

Modelling the impacts of climate change on the Arctic ecosystem requires knowledge of the key players, their dynamics, activity patterns and metabolic potential. Our finding that 83% of our Arctic Ocean MAGs represent previously unknown species shows evidence of how limited our knowledge is of prokaryotic communities in the Arctic Ocean.

Genome-resolved metagenomics in recent years has substantially expanded the tree of life[62]. Thousands of MAGs built from the global ocean[29,40,63,64] serve as a resource for microbial ecogenomic studies that shed light into new lineages with remarkable ecological impact. In this study, we present the most comprehensive dataset of uncultured prokaryotic genomes to date, providing a high number of new Arctic reference genomes including thousands of potentially non-prokaryotic bins for the exploration of keystone Arctic viral or eukaryotic genomes. Despite the wide seasonal and spatial gradient covered by the 2013 *Tara* Oceans Arctic expedition, the catalogue is representative of microbial diversity in the Arctic Ocean from spring to autumn. We encourage future studies to complement this catalogue with additional spatial and temporal coverages.

Fig. 6 | Expression patterns and metabolic potential of sentinel polar MAGs in the Arctic Ocean. The plot contains a selection of 62 MAGs, which are the most expressed per sample within the subset of polar-specific MAGs that are either generalists, specialists or uncategorized. The top tile plot represents which of the selected marker genes are encoded in each of these MAGs. The bottom heatmap represents the relative expression of each of these MAGs (*x* axis) in each sample (*y* axis). Recruitment normalizations were done for every niche breadth category. Samples on the *y* axis are coloured based on the Arctic region they belong to and the sampling season is indicated. MAGs on the *x* axis are coloured based on phylum; the numbers in parentheses correspond to the identification number of each MAG.

Our in-depth, genome-centric analysis of new lineages in the Arctic Ocean shows a prevalence of mixotrophic activity, while chemolithoautotrophy is mainly found in mesopelagic waters. Arctic MAGs show a tendency towards increased genome size and copiotrophy in MAGs present exclusively in the Arctic Ocean, compared to those also present at lower latitudes. Niche breadth analysis

of the Arctic Ocean MAGs has revealed a predominance of habitat specialists, which show higher abundances and gene expression in the mesopelagic, as well as a wider range of estimated optimal growth temperatures, compared to habitat generalists. Finally, those genomes with the highest genetic expression and so far found only in polar metagenomes were identified as key polar species of the Arctic's seawater ecosystem, some of which might be more susceptible to the effects of climate change due to their restricted niche breadth. The description of their functional capabilities and relevance in terms of genome expression is also key for future design of monitoring surveys, experiments and ecosystem models in this rapidly changing environment.

Methods

Sample and environmental data collection. As described previously[23], genetic and environmental data were collected during the *Tara* Oceans expedition (2009–2013), which includes the *Tara* Oceans Polar Expedition (2013). Polar stations had absolute latitudes above 64°. Sampling was conducted within the epipelagic (in the Arctic: surface, 5 m and SCM, 17–40 m; in the sub-Arctic North Atlantic, temperate and Southern Ocean latitudes: surface, 5 m and DCM, 30–120 m) and mesopelagic layer (in the Arctic: mesopelagic, 299–651 m; in the sub-Arctic North Atlantic, temperate and Southern Ocean latitudes: mesopelagic, 350–1,000 m). The sampling strategy and methodology have been described elsewhere[65]. Environmental data measured or inferred at the depth of sampling are published in the PANGAEA database (https://doi.org/10.1594/PANGAEA.875582).

Extraction and sequencing of DNA and complementary DNA. Metagenomic DNA and RNA were extracted from free-living prokaryote-enriched size fraction filters (0.2–3 µm) as described previously[66]. A detailed description of the DNA sequencing protocols is given in Salazar et al.[23].

Co-assembly, binning and curation. *Co-assembly.* Bins were generated from 41 *Tara* Oceans Arctic metagenomes, including 28 samples from the photic layer (20 from the surface, 7 from the SCM and 1 from the DCM), 9 from the mesopelagic layer and 4 integrated samples, in which waters from different layer were mixed. To maximize the recovery of environmental genomes from the dataset, we opted for an approach that involved the co-assembly of several samples together, hence increasing the sequencing depth for each co-assembly while keeping the computational needs attainable. The pools of samples to be co-assembled were chosen based on their taxonomic composition. Samples that clustered together in an NMDS based on 16S miTag abundance profiles were assembled jointly with megahit v.1.1.2 (--presets meta-large --min-contig-len 2000; Supplementary Table 5 and Extended Data Fig. 1)[67]. All assembled contigs were pooled together and de-replicated with CD-HIT-EST v.4.6.8-2017-0621, compiled from source with MAX_SEQ = 10000000, options -c 0.99 -T 64 -M 290000 -n 10 (ref. [68]), reducing the dataset from 3.95 to 1.91 M contigs.

Binning and curation. The input metagenome reads were back-mapped to the remaining contigs with Bowtie 2 v.2.3.2 (ref. [69]) with default options, keeping only mapping hits with quality >10 (SAMtools v.1.5; options -q 10 -F 4)[70,71]. Mapping hits were processed with jgi_summarize_bam_contig_depths from MetaBAT 2 v.2.12.1 (ref. [71]) with options --minContigLength 2000--minContigDepth 1 and then binned with MetaBAT 2 with default options.

The completeness and contamination of each bin, as well as a first estimation of their taxonomic classification, based on single-copy marker genes was assessed with CheckM v.1.0.11 (ref. [72]) using the lineage_wf workflow.

The 96 bins with an estimated completeness >95% and contamination <5% had their contigs reassembled in Geneious v.10.2.4 with a minimum overlap identity of 95%, maximum mismatches per read of 5, no minimum overlap and with no gap allowed options to find overlaps that allowed to reduce genome fragmentation; the results were curated manually. These were considered to be high-quality MAGs. Additionally, contigs of 434 bins with estimated genome completeness >50% and contamination <10% were also reassembled with CAP3 v.021015 (ref. [73]) with overlap length and percentage identity cut-offs of 25 bp and 95%, respectively. These were considered to be medium-quality MAGs.

All 3,550 genomes were given a numeric identifier, with the prefix 'TOA-bin-', which stands for *Tara* Oceans Arctic bin.

Taxonomic and functional annotation. All 3,550 bins were classified taxonomically with GTDB-Tk v.0.3.2 (ref. [37]) (GTDB release 89) using the classify_wf workflow. Genome completeness and contamination estimates were reassessed with CheckM as above. For those bins encoding the 16S rRNA gene, their taxonomic annotation was done using the SILVA 132 database and SINA aligner tool v.1.2.11 with a minimum of 50% of identity (higher thresholds could not classify the ribosomal genes) and last common ancestor algorithm (Supplementary Table 6).

Functional annotation of 530 MAGs, including gene prediction, transfer RNA, rRNA and CRISPR detection was done with Prokka v.1.13 (ref. [74]) using default options and the estimated domain classification from CheckM as the argument in the --kingdom option. Additionally, predicted coding sequences were annotated against the KEGG Orthology database (KEGG release 89.1)[75] with DIAMOND v.0.9.22 (ref. [76]) using options blastp -e 0.1--sensitive and against the Pfam database release 31.0 using HMMER v.3.1b2 (ref. [77]) and options --domtblout -E 0.1. Functional annotation of MAGs can be accessed in the Supplementary Information.

Genome redundancy analysis. ANI was calculated with FastANI v.1.2; default options[78] were estimated for each possible pair of MAGs with >50% of genome completeness and <10% of genome contamination to check whether the reconstructed genomes could belong to the same species (defined at >95% ANI). Since alignment fraction between genomes lower than 20% may provide spuriously large ANIs, the average amino acid identity, which considers only the fraction of orthologous genes, was also estimated (CompareM v.0.0.23 with default options; https://github.com/dparks1134/CompareM).

Read recruitment. *Selection and subsampling of samples.* The samples chosen for read recruitment included the 32 surface, SCM and mesopelagic metagenomes from the *Tara* Oceans Arctic Ocean sampling and the 5 metagenomes (surface, DCM and mesopelagic) obtained from the *Tara* Oceans sub-Arctic North Atlantic sampling (Figs. 1a,b), the 4 *Tara* Oceans metagenomes sampled in the Southern Ocean and a selection of 27 *Tara* Oceans expedition metagenomes from temperate latitudes (Supplementary Table 4 and Extended Data Fig. 10). These were selected based on their sequencing depth, which had to be at least as large as the smallest *Tara* Oceans Arctic metagenome, geographical location (covering the different oceans and seas sampled by the *Tara* Oceans expedition) and the coverage of different water layers. For the metagenomic samples selected, recruitment was also done with their available metatranscriptomes (33 from the *Tara* Oceans Arctic Stations, 3 from the Southern Ocean and 17 from the temperate ocean).

Paired-end libraries were used individually for fragment recruitment analysis after cleaning and a step of random subsampling. The latter was done with the DOE JGI's BBTools reformat.sh script v.38.08 (https://sourceforge.net/projects/bbmap/), selecting as the subsampling value the smallest sequencing depth of the *Tara* Oceans Arctic expedition meta-omic dataset (that is, 140,658,260 and 45,212,614 fragments for metagenomic and metatranscriptomic libraries respectively). Read length was 101 bp.

Competitive fragment recruitment analysis. Nucleotide-Nucleotide BLAST v.2.7.1+ was used to recruit metagenomic and metatranscriptomic reads similar to any of the 3,550 Arctic bins. BLAST is slower than other high-throughput aligners but allows for finer-tuned alignment parameters, plus it is the criterion standard against which all high-throughput aligners are compared. Recruitment was competitive, meaning that individual samples were aligned against the pooled contigs of all 3,550 bins. The BLAST alignment parameters were the following: -perc_identity 70, -evalue 0.0001. Only those reads with >90% coverage and mapping at identities equal to or higher than 95% were considered to be representative of the bin. In case of hits with the same e-value, larger bit-score or larger alignment length were used sequentially to choose the best hit. If ties persisted, the best hit was selected at random from the candidate reads. Best hits that corresponded to rRNAs (according to the Prokka annotation) were also discarded.

Detection and filtering of false-positive recruitments. Putative false-positive recruitments were detected and excluded considering their horizontal genomic coverage, which was calculated using the R package GenomicRanges v.1.34.0 (ref. [79]).

A minimum horizontal genomic coverage threshold was set testing the effect of different thresholds on the final number of bins recruiting (richness) and the number of samples in which they recruited (occurrence). The variation of species richness in each metagenome was tested for a range of increasing minimum horizontal genomic coverage thresholds (0, 0.1, 0.5, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 98, 100). Recruitments where the horizontal coverage was equal to or higher than the thresholds were considered true; those covering a smaller percentage of their genome than the cut-off value were discarded.

The number of species present in each metagenome decreased with the increase of minimum horizontal coverage, reaching an apparent saturation in richness when the minimum horizontal coverage was 20% for metagenomes from temperate latitudes (Supplementary Fig. 8).

Setting a horizontal genomic coverage threshold had an effect on the occurrence of each bin in the metagenomic samples. In all metagenomic datasets (Arctic, Southern Ocean and temperate), the distribution of occurrence versus mean abundance (RPKG) of bins stabilized when the minimum horizontal coverage was 10% or higher (Supplementary Fig. 9). Lower thresholds showed different patterns of distribution, increasing the number of higher occurrences at very low mean abundances (Supplementary Fig. 9). To date, there is no consensus about the minimum horizontal coverage thresholds to discard false mappings.

Based on our analyses, we chose 20% as the minimum horizontal genomic coverage to consider recruitments valid. The metagenomic read recruitments can be found in Supplementary Table 7 and the metatranscriptomic read recruitments are in Supplementary Table 8.

Abundance and distribution of bins. *Estimation of bin abundance and occurrence.* Only those read recruitments aligning with an identity ≥95% were considered to be representative of the bins. Recruitments passing the minimum horizontal genomic coverage threshold of 20% were considered to represent an actual presence of the bin in the sample. In comparison, those with a horizontal genomic coverage <20% were considered not representative of the bin and thus absent in the sample. Read recruitments were transformed to RPKGs. Metagenomic RPKGs are found in Supplementary Table 9 and metatranscriptomic RPKGs are found in Supplementary Table 10.

Distribution of communities based on medium- and high-quality MAG composition. The ordination of samples based on their MAG composition, with RPKG as an abundance estimate, was done with an NMDS approach using the function metaMDS from the vegan v.2.5-6 package in R. Permutational MANOVA was calculated using the function adonis (vegan package) with Bray–Curtis distances and 999 permutations.

Pan-Arctic categorization. MAGs were classified as pan-Arctic if they were present in at least 16 out of the 18 Arctic surface metagenomes and also in the 5 Arctic regions sampled.

Minimum generation time predictions. Estimations of minimum generation time and optimal growth temperature were performed for high-quality ($n=96$) and medium-quality MAGs ($n=434$) using Growthpred[80] and can be found in Supplementary Table 11. Growthpred relies on codon usage biases in highly expressed genes identified in genomes. First, for each MAG, we extracted the coding sequence using GffRead v.0.11.7 (ref. [81]). Growthpred then assessed the number of highly expressed genes present in the MAG. If this number was low (<10), then the MAG was discarded; otherwise, Growthpred estimated the optimal growth temperature (option -t) since we did not know it a priori. Usage codon bias was calculated using universal genetic code (option -c 0). Highly expressed genes (ribosomal protein genes) were retrieved from coding sequences using BLAST (option -r). Growthpred v.1.08 was used and a Snakemake pipeline is available at https://gitlab.univ-nantes.fr/combi-ls2n/growthsnake.

Statistical support for minimum generation time and optimal growth temperature differences between the different biogeographical groups was calculated using the Dunnett–Tukey–Kramer pairwise multiple comparison test adjusted for unequal variances and unequal sample sizes and 95% confidence interval (CI) with the DTK v.3.5 R package (https://cran.r-project.org/package=DTK).

Niche breadth and classification of MAGs as specialists or generalists. *Habitat specialist-generalist patterns in the Arctic Ocean.* Specialist-generalist classification of MAGs was based on Levins' index (B)[52]. To avoid sampling bias, the function spec.gen from the R package EcolUtils v.0.1 (https://github.com/GuillemSalazar/EcolUtils) was used to calculate B for 1,000 random permutations of the metagenomic RPKG table. It allowed us to categorize MAGs into generalists if the original B index was larger than its 95% CI or specialists if the original B index was smaller than its 95% CI. Since sampling occurred in a spatial and temporal gradient, each individual sample was considered as a habitat.

Statistical support for assembly size differences between the different niche breadth and biogeographical groups was calculated using the Dunnett–Tukey–Kramer pairwise multiple comparison test adjusted for unequal variances and unequal sample sizes and 95% CI with the DTK R package (https://cran.r-project.org/package=DTK).

Functional analysis of MAGs and transcript abundance. To explore the ubiquity of representative biogeochemical cycling metabolisms related to carbon, sulphur, nitrogen and methane, a selection of 120 marker genes (Supplementary Table 2) was searched in the Arctic MAG dataset; only those pathways with enough encoded markers were considered valid.

To estimate the transcript abundances of interesting marker genes, we filtered those metatranscriptomic read recruitments falling within the coordinates of the gene of interest (according to the quality standards explained in the read recruitment section). Read recruitments were normalized by gene size and sequencing depth using the reads per kilobase of transcript per million mapped reads (RPKM) unit.

Phylogeny of RuBisCO large-chain amino acid sequences. A total of 14 RuBisCO large-chain amino acid sequences were detected by their KEGG Orthology annotation (K01601) in Arctic MAGs. They were aligned against the RuBisCO large-chain reference alignment profile published by Jaffe et al.[42] and the RuBisCO large-chain sequences from heterotrophic marine Thaumarchaeota published by Aylward and Santoro[82] using Clustal Omega v.1.2.3 (default options and 100 iterations)[83]. Maximum-likelihood phylogenetic reconstruction was done using the Jones–Taylor–Thorton model with FastTree v.2.1.11 (default options)[84]. Phylogenetic tree editing was done in iTol v.6.3.2 (https://itol.embl.de)[85].

Definition of polar key Arctic MAGs. Those MAGs that showed metagenomic recruitment exclusively in polar samples were selected and classification as key polar genomes was done for the ones showing higher metatranscriptomic RPKGs per sample within each niche breadth category. For each sample and niche breadth category, all individual RPKGs were calculated relative to the highest in the sample; only those RPKG recruitments representative of at least 50% of the highest RPKG recruitment per sample were selected as key polar genomes and are shown in Fig. 6.

Estimation of ribosomal copy numbers in *Tara* Oceans samples from miTags. We assigned to each 16S miTag an estimated 16S rRNA gene copy number and calculated a mean value per sample weighted by the relative abundances. Prediction uncertainty of gene copy number for each 16S miTag depends on their phylogenetic distance to the closest complete genome and this parameter determines the mean degree of uncertainty per sample, also weighted by relative abundances. To evaluate the potential variation of the 16S rRNA gene copy number in the *Tara* Oceans samples, we made use of resources available in Louca et al.[86], which offered estimated values for SILVA accessions, plus an uncertainty measure based on their phylogenetic distance to the closest sequenced genome. In fact, this source of uncertainty prevents highly accurate estimations, for which results should be taken with caution[86]. Through the SILVA accessions of the *Tara* Oceans 16S miTags we calculated a mean copy number weighted by their relative abundance in all *Tara* Oceans epipelagic samples.

Data visualizations. All maps and data visualizations included in this manuscript have been generated with the R package ggplot2 v.3.3.2. Multi-panels and post-processing were done in Illustrator CC 2018 (Adobe).

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

Data availability

Accession numbers for the data used and generated in this study can be found in Supplementary Table 12, which includes the Arctic MAGs Catalogue and their functional annotation (European Bioinformatics Institute BioStudies ID: S-BSST451) and the co-assembly of metagenomic samples used to generate the metagenomic bins (European Nucleotide Archive PRJEB41575). Accession numbers for the metagenomic and metatranscriptomic samples used in the fragment recruitment analyses can be found in Supplementary Table 13. Publicly available datasets used in this study include the following: CheckM v.1.0.11 (https://github.com/Ecogenomics/CheckM/releases/tag/v1.1.0), GTDB release 89 (https://data.gtdb.ecogenomic.org/releases/release89/), SILVA 132 (https://www.arb-silva.de/documentation/release-132/), KEGG release 89.1 (https://www.genome.jp/kegg/docs/relnote.html) and Pfam database release 31.0 (http://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam31.0/). Source data are provided with this paper.

Received: 30 July 2020; Accepted: 13 September 2021;
Published online: 15 November 2021

References

1. IPCC. *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate* (in the press).
2. Cavicchioli, R. et al. Scientists' warning to humanity: microorganisms and climate change. *Nat. Rev. Microbiol.* **17**, 569–586 (2019).
3. Meltofte, H. (ed.) *Arctic Biodiversity Assessment: Status and Trends in Arctic Biodiversity* (CAFF International Secretariat, 2013).
4. Wassmann, P. & Reigstad, M. Future Arctic Ocean seasonal ice zones and implications for pelagic-benthic coupling. *Oceanography* **24**, 220–231 (2011).
5. Bunse, C. & Pinhassi, J. Marine bacterioplankton seasonal succession dynamics. *Trends Microbiol.* **25**, 494–505 (2017).
6. Olli, K. et al. Seasonal variation in vertical flux of biogenic matter in the marginal ice zone and the central Barents Sea. *J. Mar. Syst.* **38**, 189–204 (2002).
7. Riedel, A., Michel, C., Gosselin, M. & LeBlanc, B. Winter–spring dynamics in sea-ice carbon cycling in the coastal Arctic Ocean. *J. Mar. Syst.* **74**, 918–932 (2008).
8. Joli, N., Monier, A., Logares, R. & Lovejoy, C. Seasonal patterns in Arctic prasinophytes and inferred ecology of *Bathycoccus* unveiled in an Arctic winter metagenome. *ISME J.* **11**, 1372–1385 (2017).
9. Alonso-Sáez, L., Sánchez, O., Gasol, J. M., Balagué, V. & Pedrós-Alio, C. Winter-to-summer changes in the composition and single-cell activity of near-surface Arctic prokaryotes. *Environ. Microbiol.* **10**, 2444–2454 (2008).
10. Alonso-Sáez, L. et al. Role for urea in nitrification by polar marine Archaea. *Proc. Natl Acad. Sci. USA* **109**, 17989–17994 (2012).
11. Boetius, A., Anesio, A. M., Deming, J. W., Mikucki, J. A. & Rapp, J. Z. Microbial ecology of the cryosphere: sea ice and glacial habitats. *Nat. Rev. Microbiol.* **13**, 677–690 (2015).

12. Circumpolar Biodiversity Monitoring Program, Conservation of Arctic Flora and Fauna. *State of the Arctic Marine Biodiversity Report* (Conservation of Arctic Flora and Fauna International Secretariat, 2017).
13. Kirchman, D. L., Cottrell, M. T. & Lovejoy, C. The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. *Environ. Microbiol.* **12**, 1132–1143 (2010).
14. Galand, P. E., Casamayor, E. O., Kirchman, D. L., Potvin, M. & Lovejoy, C. Unique archaeal assemblages in the Arctic Ocean unveiled by massively parallel tag sequencing. *ISME J.* **3**, 860–869 (2009).
15. Pedrós-Alió, C., Potvin, M. & Lovejoy, C. Diversity of planktonic microorganisms in the Arctic Ocean. *Prog. Oceanogr.* **139**, 233–243 (2015).
16. Amaral-Zettler, L. et al. in *Life in the World's Oceans: Diversity, Distribution, and Abundance* (ed. McIntyre, A. D.) 221–245 (Blackwell Publishing Ltd, 2010).
17. Christman, G. D., Cottrell, M. T., Popp, B. N., Gier, E. & Kirchman, D. L. Abundance, diversity, and activity of ammonia-oxidizing prokaryotes in the coastal Arctic Ocean in summer and winter. *Appl. Environ. Microbiol.* **77**, 2026–2034 (2011).
18. Alonso-Sáez, L., Galand, P. E., Casamayor, E. O., Pedrós-Alió, C. & Bertilsson, S. High bicarbonate assimilation in the dark by Arctic bacteria. *ISME J.* **4**, 1581–1590 (2010).
19. Galand, P. E., Lovejoy, C., Pouliot, J., Garneau, M.-È. & Vincent, W. F. Microbial community diversity and heterotrophic production in a coastal Arctic ecosystem: a stamukhi lake and its source waters. *Limnol. Oceanogr.* **53**, 813–823 (2008).
20. Nguyen, D. et al. Winter diversity and expression of proteorhodopsin genes in a polar ocean. *ISME J.* **9**, 1835–1845 (2015).
21. Cifuentes-Anticevic, J. et al. Proteorhodopsin phototrophy in Antarctic coastal waters. *mSphere* **6**, e00525–21 (2021).
22. Ghiglione, J.-F. et al. Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proc. Natl Acad. Sci. USA* **109**, 17633–17638 (2012).
23. Salazar, G. et al. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell* **179**, 1068–1083. e21 (2019).
24. 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).
25. Cao, S. et al. Structure and function of the Arctic and Antarctic marine microbiota as revealed by metagenomics. *Microbiome* **8**, 47 (2020).
26. Sunagawa, S. et al. Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **18**, 428–445 (2020).
27. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
28. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
29. Delmont, T. O. et al. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat. Microbiol.* **3**, 804–813 (2018).
30. Ibarbalz, F. M. et al. Global trends in marine plankton diversity across kingdoms of life. *Cell* **179**, 1084–1097.e21 (2019).
31. Aagaard, K., Swift, J. H. & Carmack, E. C. Thermohaline circulation in the Arctic Mediterranean Seas. *J. Geophys. Res. Oceans* **90**, 4833–4846 (1985).
32. Dupont, C. L. et al. Genomes and gene expression across light and productivity gradients in eastern subtropical Pacific microbial communities. *ISME J.* **9**, 1076–1092 (2015).
33. Franzosa, E. A. et al. Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl Acad. Sci. USA* **111**, E2329–E2338 (2014).
34. Jones, S. E. & Lennon, J. T. Dormancy contributes to the maintenance of microbial diversity. *Proc. Natl Acad. Sci. USA* **107**, 5881–5886 (2010).
35. Mestre, M. & Höfer, J. The microbial conveyor belt: connecting the globe through dispersion and dormancy. *Trends Microbiol.* **29**, 482–492 (2021).
36. Ciufo, S. et al. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int. J. Syst. Evol. Microbiol.* **68**, 2386–2392 (2018).
37. Chaumeil, P-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
38. Nelson, W. C., Tully, B. J. & Mobberley, J. M. Biases in genome reconstruction from metagenomic data. *PeerJ* **8**, e10119 (2020).
39. Alneberg, J. et al. Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. *Commun. Biol.* **3**, 119 (2020).
40. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* **5**, 170203 (2018).
41. Christensen, M. & Nilsson, A. E. Arctic sea ice and the communication of climate change. *Pop. Commun.* **15**, 249–268 (2017).
42. Jaffe, A. L., Castelle, C. J., Dupont, C. L. & Banfield, J. F. Lateral gene transfer shapes the distribution of RuBisCO among candidate phyla radiation bacteria and DPANN Archaea. *Mol. Biol. Evol.* **36**, 435–446 (2019).
43. Kono, T. et al. A RuBisCO-mediated carbon metabolic pathway in methanogenic archaea. *Nat. Commun.* **8**, 14007 (2017).
44. Sato, T., Atomi, H. & Imanaka, T. Archaeal type III RuBisCOs function in a pathway for AMP metabolism. *Science* **315**, 1003–1006 (2007).
45. Tabita, F. R., Satagopan, S., Hanson, T. E., Kreel, N. E. & Scott, S. S. Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships. *J. Exp. Bot.* **59**, 1515–1524 (2008).
46. Yelton, A. P. et al. Global genetic capacity for mixotrophy in marine picocyanobacteria. *ISME J.* **10**, 2946–2957 (2016).
47. Cordero, P. R. F. et al. Atmospheric carbon monoxide oxidation is a widespread mechanism supporting microbial survival. *ISME J.* **13**, 2868–2881 (2019).
48. King, G. M. & Weber, C. F. Distribution, diversity and ecology of aerobic CO-oxidizing bacteria. *Nat. Rev. Microbiol.* **5**, 107–118 (2007).
49. Sunagawa, S. et al. Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
50. Sul, W. J., Oliver, T. A., Ducklow, H. W., Amaral-Zettler, L. A. & Sogin, M. L. Marine bacteria exhibit a bipolar distribution. *Proc. Natl Acad. Sci. USA* **110**, 2342–2347 (2013).
51. Roller, B. R. K., Stoddard, S. F. & Schmidt, T. M. Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. *Nat. Microbiol.* **1**, 16160 (2016).
52. Levins, R. *Evolution in Changing Environments: Some Theoretical Explorations* (Princeton Univ. Press, 1968).
53. Colwell, R. K. & Futuyma, D. J. On the measurement of niche breadth and overlap. *Ecology* **52**, 567–576 (1971).
54. Massana, R. & Logares, R. Eukaryotic versus prokaryotic marine picoplankton ecology. *Environ. Microbiol.* **15**, 1254–1261 (2013).
55. Székely, A. J., Berga, M. & Langenheder, S. Mechanisms determining the fate of dispersed bacterial communities in new environments. *ISME J.* **7**, 61–71 (2013).
56. Brooks, J. P. et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol.* **15**, 66 (2015).
57. Logares, R. et al. Biogeography of bacterial communities exposed to progressive long-term environmental change. *ISME J.* **7**, 937–948 (2013).
58. Ruiz-González, C. et al. Higher contribution of globally rare bacterial taxa reflects environmental transitions across the surface ocean. *Mol. Ecol.* **28**, 1930–1945 (2019).
59. Staley, J. T. & Gosink, J. J. Poles apart: biodiversity and biogeography of sea ice bacteria. *Annu. Rev. Microbiol.* **53**, 189–215 (1999).
60. Chaffron, S. et al. Environmental vulnerability of the global ocean epipelagic plankton community interactome. *Sci. Adv.* **7**, eabg1921 (2021).
61. Estrada, E. Characterization of topological keystone species: local, global and "meso-scale" centralities in food webs. *Ecol. Complex.* **4**, 48–57 (2007).
62. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
63. Tully, B. J., Sachdeva, R., Graham, E. D. & Heidelberg, J. F. 290 metagenome-assembled genomes from the Mediterranean Sea: a resource for marine microbiology. *PeerJ* **2017**, e3558 (2017).
64. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. *Commun. Biol.* **4**, 604 (2021).
65. Pesant, S. et al. Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2**, 150023 (2015).
66. Alberti, A. et al. Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci. Data* **4**, 170093 (2017).
67. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
68. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
69. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
70. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
71. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
72. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
73. Huang, X. & Madan, A. CAP3: a DNA sequence assembly program. *Genome Res.* **9**, 868–877 (1999).
74. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

75. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
76. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
77. Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487–2489 (2013).
78. Jain, C., Rodriguez-R, L. M., Phillipy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
79. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
80. Vieira-Silva, S. & Rocha, E. P. C. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* **6**, e1000808 (2010).
81. Pertea, G. & Pertea, M. GFF utilities: GffRead and GffCompare. *F1000Res.* **9**, ISCB Comm J-304 (2020).
82. Aylward, F. O. & Santoro, A. E. Heterotrophic Thaumarchaeota with ultrasmall genomes are widespread in the ocean. *mSystems* **5**, e00415–20 (2020).
83. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
84. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2––approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
85. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
86. Louca, S., Doebeli, M. & Parfrey, L. W. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome* **6**, 41 (2018).

Acknowledgements

Tara Oceans (which includes both the *Tara* Oceans and *Tara* Oceans Polar Circle expeditions) would not exist without the leadership of the *Tara* Ocean Foundation and the continuous support of 23 institutes (http://oceans.taraexpeditions.org). We thank SHOOK Studio for assistance with designing the figures. This work acknowledges the 'Severo Ochoa Centre of Excellence' accreditation (CEX2019-000928-S). We thank the commitment of the following sponsors and research funding agencies: the Spanish Ministry of Economy and Competitiveness (project MAGGY, grant no. CTM2017-87736-R and Polar EcoGen PID2020-116489RB-I00), Horizon 2020-Research and Innovation Framework Programme (Atlantic ECOsystems assessment, forecasting & sustainability, grant no. H2020-BG-2019-2), Centre National de la Recherche Scientifique (in particular Groupement de Recherche GDR3280 and the Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/*Tara* Oceans-GOSEE), European Molecular Biology Laboratory, Genoscope/Commissariat à l'Énergie Atomique et aux Énergies Alternatives, the French Ministry of Research and the French Government's 'Investissements d'Avenir' programmes OCEANOMICS (project no. ANR-11-BTBR-0008), FRANCE GENOMIQUE (project no. ANR-10-INBS-09-08), MEMO LIFE (project no. ANR-10-LABX-54), Paris Sciences et Lettres University (project no. ANR-11-IDEX-0001-02), Eidgenössische Technische Hochschule Zürich and Helmut Horten Foundation, the Swiss National Foundation (project no. 205321_184955), MEXT/JSPS/KAKENHI (project nos. 16H06429, 16K21723, 16H06437 and 18H02279). We also thank the support and commitment of agnès b. and E. Bourgois, the Prince Albert II de Monaco Foundation, the Veolia Foundation, Region Bretagne, Lorient Agglomération, Serge Ferrari, World Courier and King Abdullah University of Science and Technology. The global sampling effort was enabled by countless scientists and crews who sampled aboard the *Tara* from 2009 to 2013. We thank Mercator/Coriolis and ACRI-ST for providing daily satellite data during the expedition. We also thank the countries who graciously granted sampling permissions. All data reported herein are fully and freely available from the date of publication, with no restrictions; all of the analyses, publications and ownership of data are free from legal entanglement or restriction by the various nations whose waters the *Tara* Oceans expeditions sampled in. This article is contribution number 122 of *Tara* Oceans.

Author contributions

M.R.-L. developed the methodology, analysed the data, designed the data visualizations and wrote the manuscript. P.S. developed the methodology and analysed the data. C.R.-G., G.S., L.P., F.I., B.C. and M.S. analysed the data. L.Z. and S.C. analysed the data and contributed to the interpretation of the findings. C.P.-A., L.K.-B. and C.B. contributed to the interpretation of the findings and provided critical reading of the manuscript. K.L. and P.W. coordinated all sequencing efforts. D.E. and S.S. provided critical reading of the manuscript. E.K. is the director of the *Tara* Oceans expedition. C.B., L.K.-B. and M.B. directed the *Tara* Oceans Polar Circle expedition. The *Tara* Oceans coordinators conceptualized the research, organized the sampling efforts and revised the manuscript. S.G.A. created the study design, developed the methodology, analysed the data and helped to write the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41564-021-00979-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-021-00979-9.

Correspondence and requests for materials should be addressed to Silvia G. Acinas.

Peer review information *Nature Microbiology* thanks Eric Collins, David Pearce and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

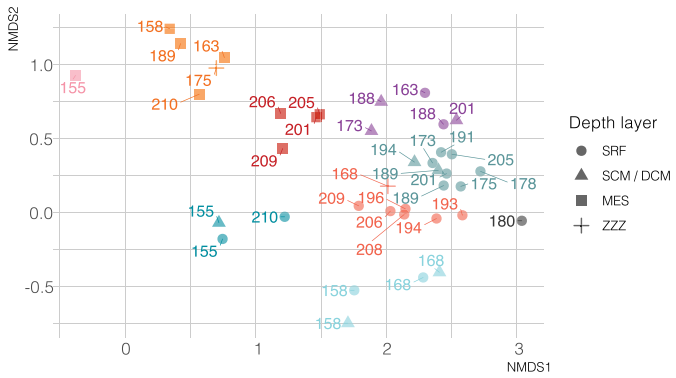
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Tara Oceans Coordinators

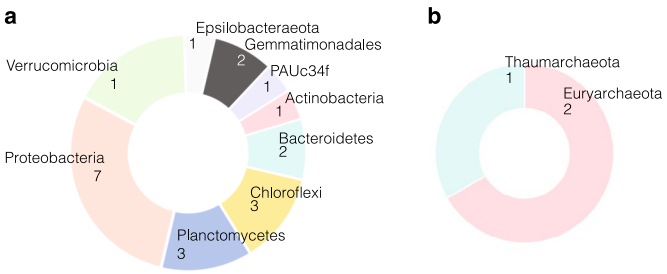
Silvia G. Acinas[1], Marcel Babin[12], Peer Bork[13,14,15], Emmanuel Boss[11], Chris Bowler[5,8], Guy Cochrane[16], Colomban de Vargas[8,17], Gabriel Gorsky[8,18], Nigel Grimsley[8,19,20], Lionel Guidi[8,18], Pascal Hingamp[8,21], Daniele Iudicone[22], Olivier Jaillon[4,8], Stefanie Kandels[9,13], Lee Karp-Boss[11], Eric Karsenti[5,8,9], Fabrice Not[8,23], Hiroyuki Ogata[24], Stéphane Pesant[25,26], Nicole Poulton[27], Jeroen Raes[28,29,30], Christian Sardet[8,18], Sabrina Speich[8,31,32], Lars Setmmann[8,18], Matthew B. Sullivan[33], Shinichi Sunagawa[2] and Patrick Wincker[8,10]

[12]Takuvik International Research Laboratory (IRL 3376), Université Laval-Centre National de la Recherche Scientifique, Université Laval, Quebec City, Quebec, Canada. [13]Structural and Computational Biology, European Molecular Biology Laboratory, Heidelberg, Germany. [14]Max Delbrück Center for Molecular Medicine, Berlin, Germany. [15]Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany. [16]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK. [17]Sorbonne Université, Centre National de la Recherche Scientifique, Station Biologique de Roscoff, AD2M ECOMAP, Roscoff, France. [18]Sorbonne Université, Centre National de la Recherche Scientifique, Laboratoire d'Océanographie de Villefanche, Villefranche-sur-mer, France. [19]Centre National de la Recherche Scientifique Unité Mixte de Recherche 7232, Biologie Intégrative des Organismes Marins, Banyuls-sur-Mer, France. [20]Sorbonne Universités Paris 06, Observatoire Océanologique de Banyuls Université Pierre et Marie Curie, Banyuls-sur-Mer, France. [21]Aix-Marseille Université, Université de Toulon, Centre National de la Recherche Scientifique, Institut de Recherche pour le Développement, Mediterranean Institute of Oceanography UM 110, Marseille, France. [22]Stazione Zoologica Anton Dohrn, Naples, Italy. [23]Sorbonne Université, Centre National de la Recherche Scientifique - UMR7144 - Ecology of Marine Plankton Group, Station Biologique de Roscoff, Place Georges Teissier, Roscoff, France. [24]Institute for Chemical Research, Kyoto University, Kyoto, Japan. [25]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [26]MARUM, Center for Marine Environmental Sciences, University of Bremen,

Bremen, Germany. [27]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA. [28]Department of Microbiology and Immunology, Rega Institute, Katholieke Universiteit Leuven, Leuven, Belgium. [29]Center for the Biology of Disease, Vlaams Instituut voor Biotechnologie Katholieke Universiteit Leuven, Leuven, Belgium. [30]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Brussels, Belgium. [31]Department of Geosciences, Laboratoire de Météorologie Dynamique, Ecole Normale Supérieure, Paris, France. [32]Ocean Physics Laboratory, University of Western Brittany, Brest, France. [33]Departments of Microbiology and Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH, USA.

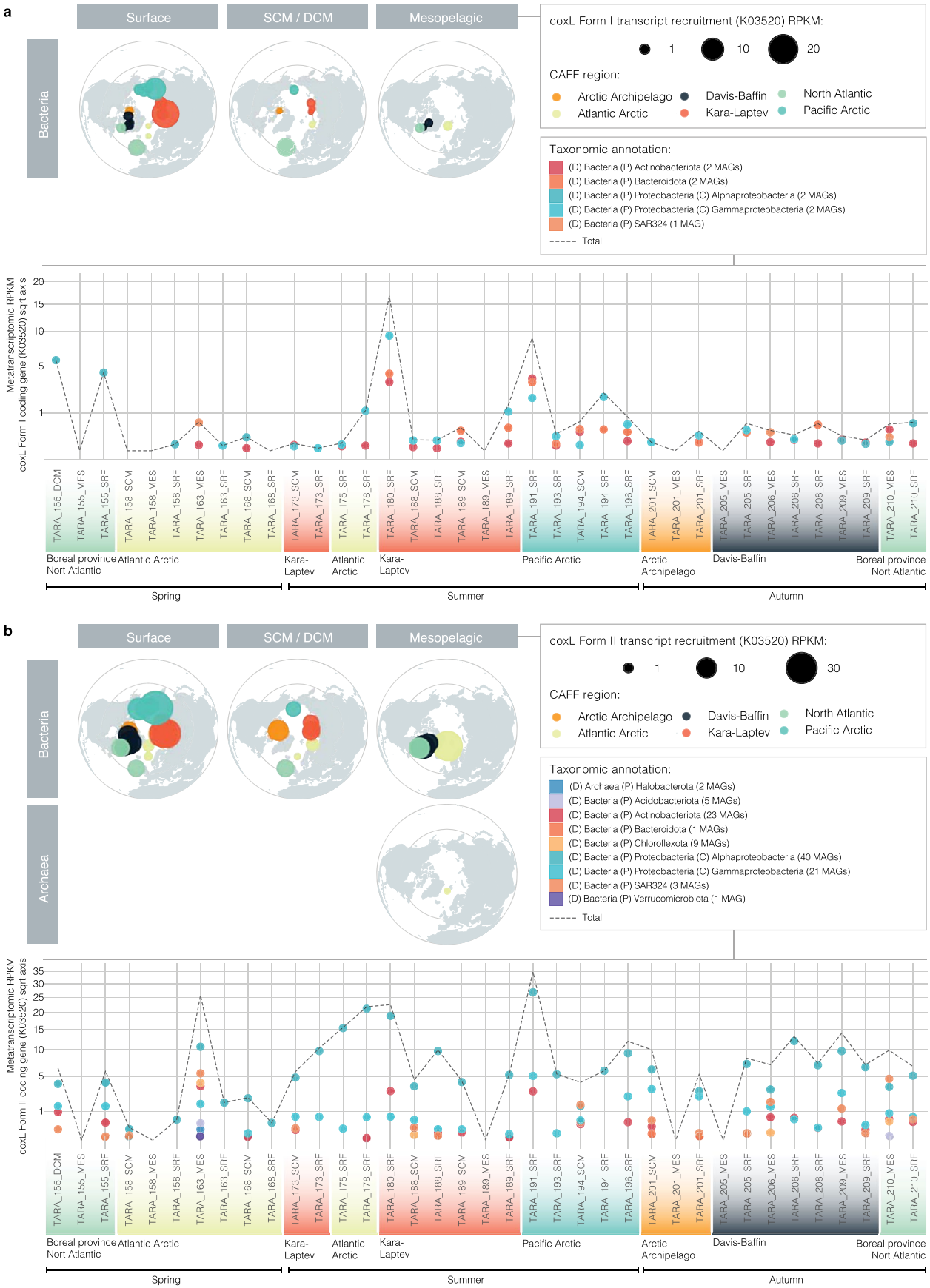


Extended Data Fig. 1 | NMDS of the 16S miTAG community composition of the 41 *Tara* Oceans Polar Circle metagenomes. Colors delimit the 9 groups of samples used for co-assembly in order to build Arctic bins. Shape indicates the ocean layer from which each metagenomic sample was collected.



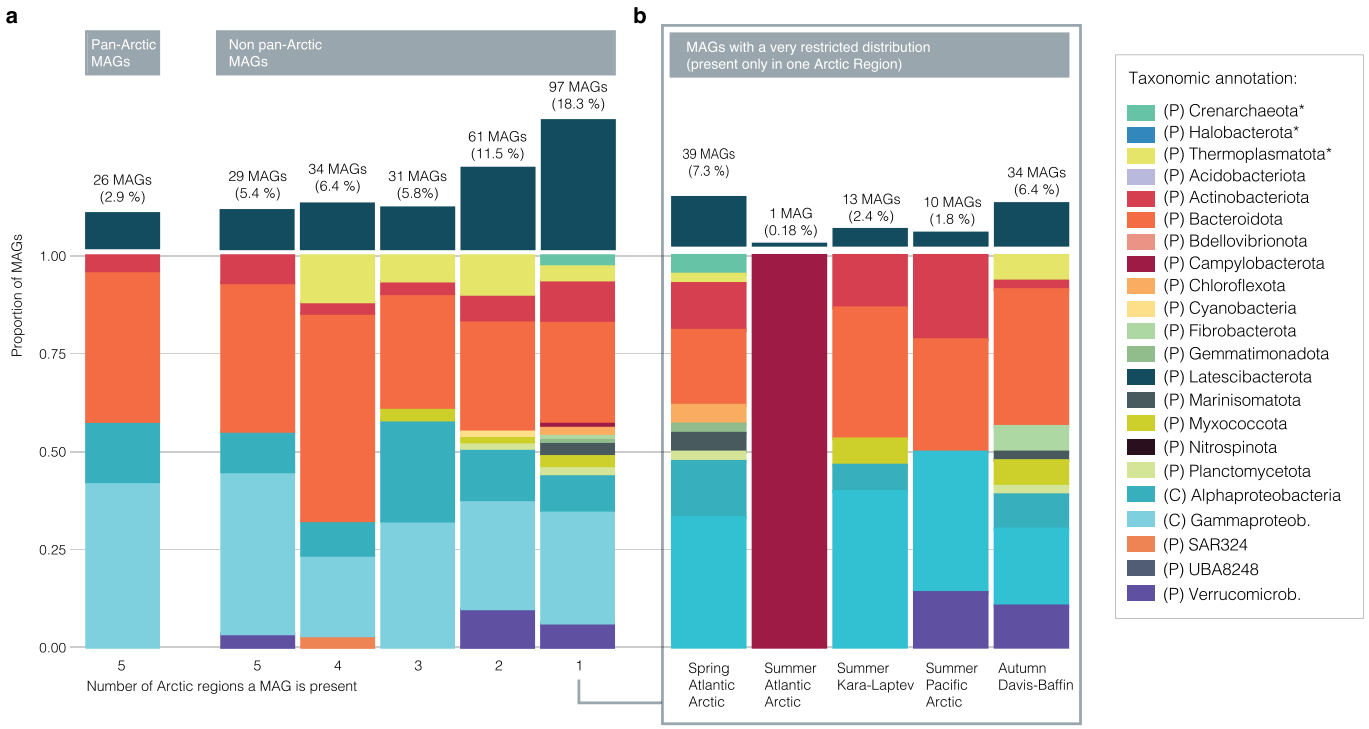
Extended Data Fig. 2 | Taxonomic classification of the 27 partial ribosomal genes encoded in the 530 MQ and HQ Arctic MAGs. a, Number of Arctic MAGs assigned to each phylum in the Bacteria domain. **b**, Number of Arctic MAGs assigned to each phylum in the Archaea domain.

Extended Data Fig. 3 | Rank-abundance curve of Arctic MAGs in Arctic metagenomes. MAGs are sorted in X axis by their accumulated RPKGs in the 37 Arctic metagenomes (including the sub-Arctic North Atlantic) used in this study. MAGs are colored by phyla and the those recruiting at least 200 RPKGs are labelled with extended taxonomic annotation. Taxonomic annotation reaches the furthest level of classification for each MAG and the number in parenthesis is the MAG's identification code.

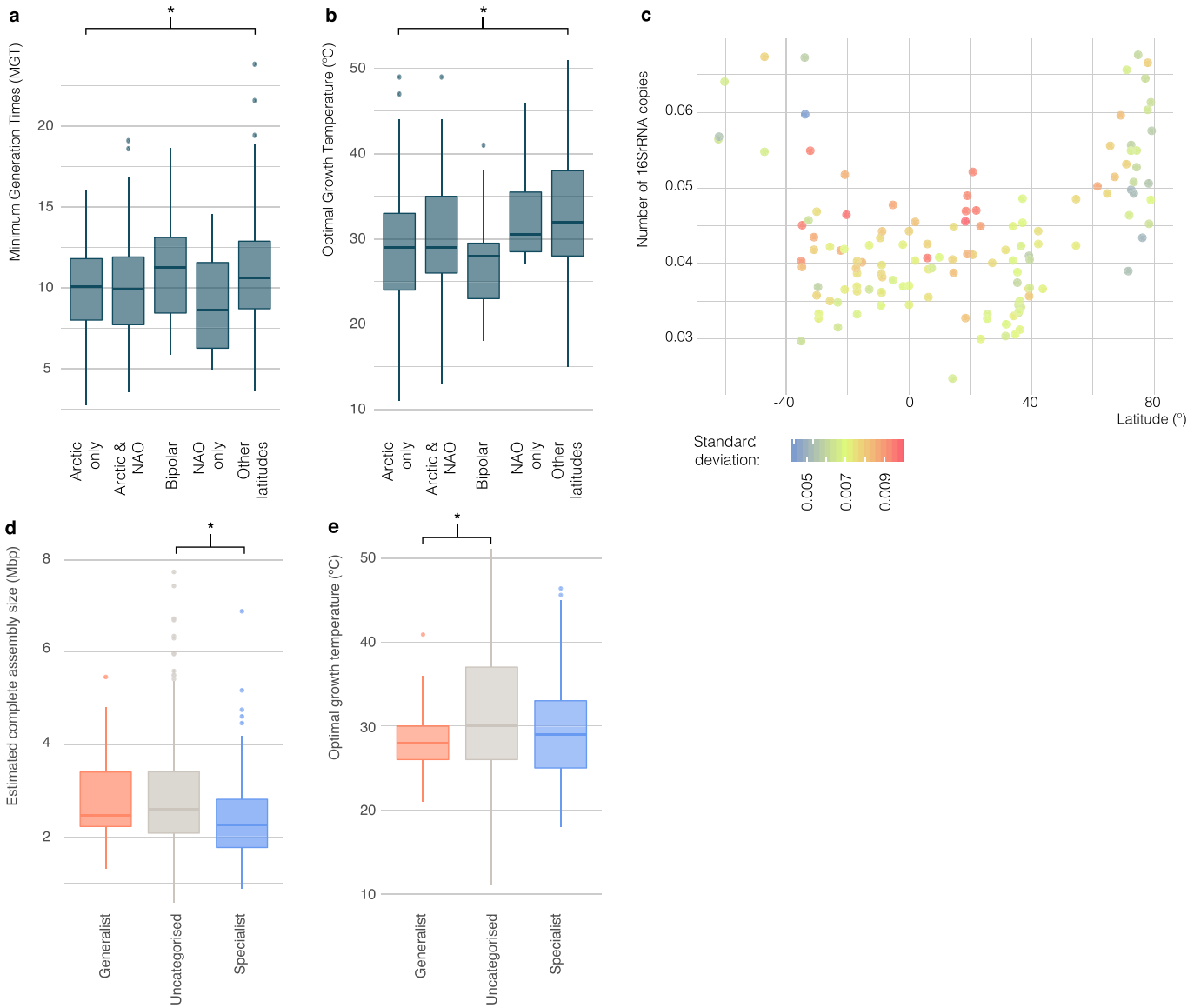


Extended Data Fig. 4 | See next page for caption.

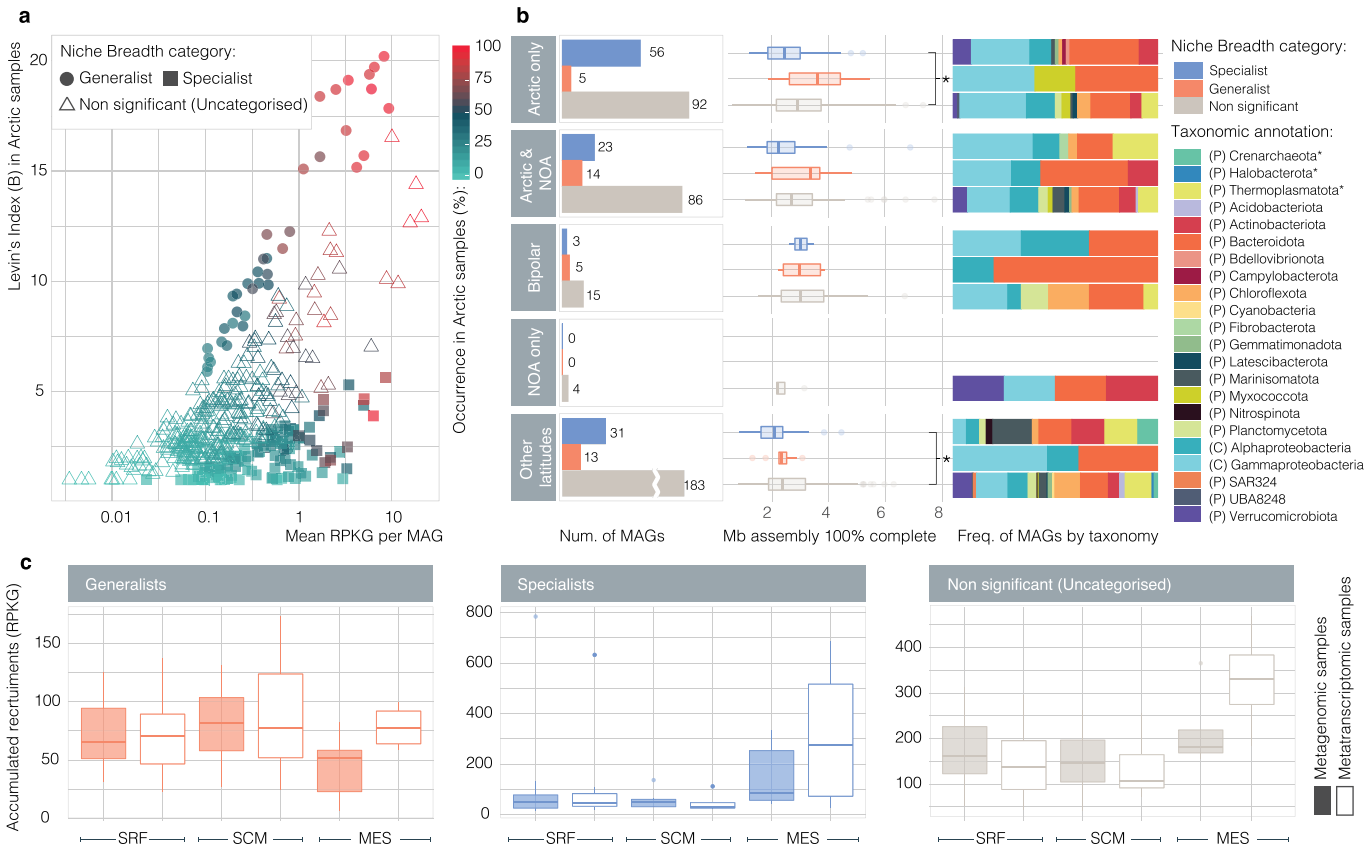
Extended Data Fig. 4 | Transcript abundance (RPKM; reads per gene kilobase per million of sequenced reads) of the marker gene *coxL* (Carbon monoxide dehydrogenase large chain) K03520 from the aerobic carbon-monoxide dehydrogenase. a, Polar maps with the accumulated metatranscriptomic RPKMs of 9 Arctic MAGs expressing the CO fixing *coxL* Form I, color-coded by CAFF region. The size of the dot is proportional to the accumulated metatranscriptomic RPKMs. In the dot plot below, RPKMs of 9 Arctic MAGs expressing CO fixing *coxL* Form I colored based on taxonomic annotation at the phylum level. Accumulated RPKMs per sample is depicted with a dashed black line. **b**, Polar maps with the accumulated metatranscriptomic RPKMs of 105 Arctic MAGs expressing *coxL* Form II, color-coded by CAFF region. The size of the dot is proportional to the accumulated metatranscriptomic RPKGs. Absent maps mean that no recruitment was found for that specific metabolism/domain/layer. In the dot plot below, RPKMs of 9 Arctic MAGs expressing *coxL* Form II colored based on taxonomic annotation at the phylum level. Accumulated RPKMs per sample is depicted with a dashed black line.



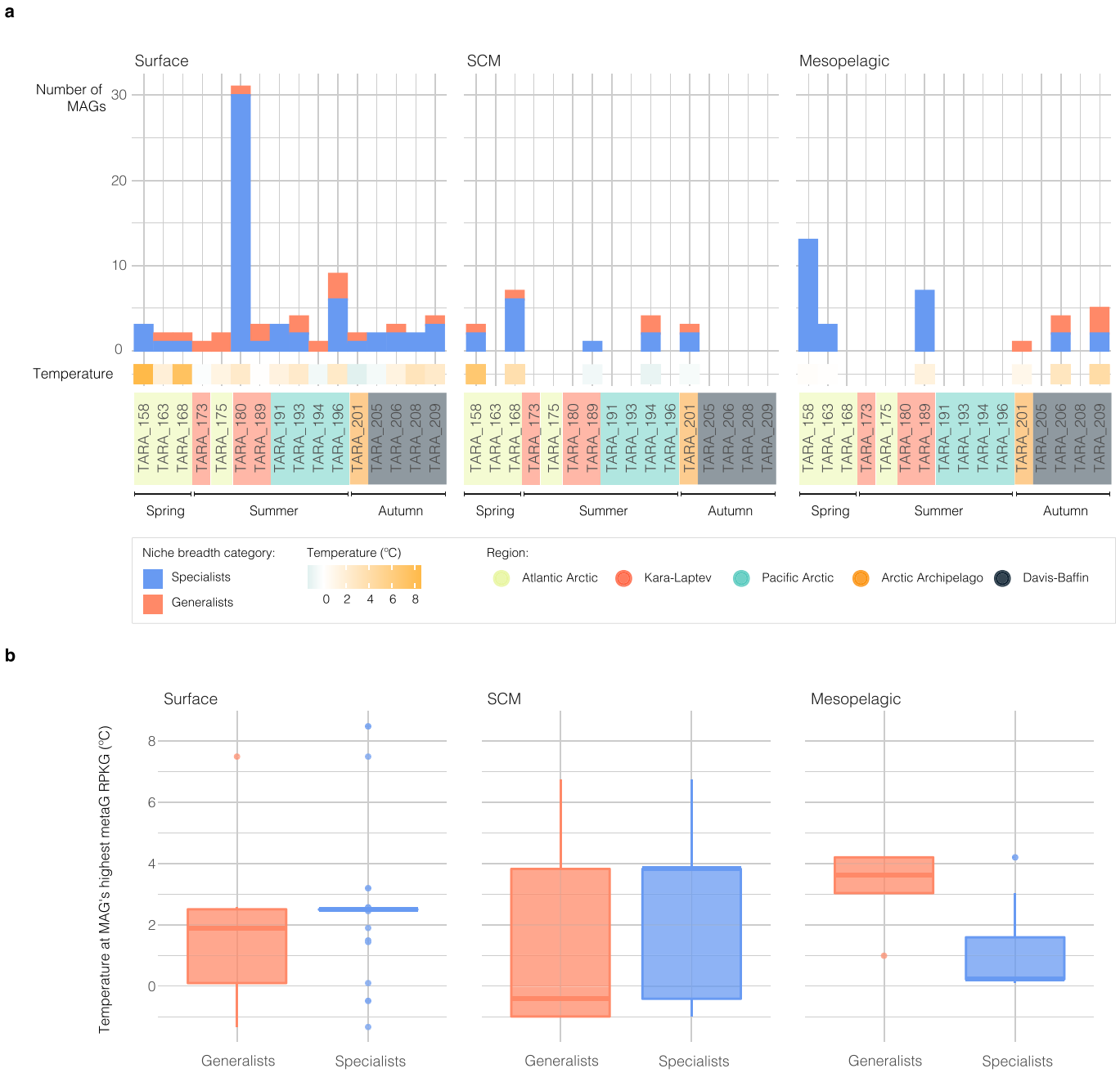
Extended Data Fig. 5 | Pan-Arctic profiles of *Tara* Arctic Ocean MAGs catalogue. a, Quantification and taxonomic classification of MAGs based on their pan-Arctic profiles. Stacked barplots representative of the number of pan-Arctic MAGs (left) and non pan-Arctic MAGs (right), colored by phylum. MAGs have been classified based on the number of Arctic regions they are present, represented in the X axis. Absolute number of MAGs per X axis category can be seen in the top barplots. More details for pan-Arctic categorization can be found in the Methods sections. **b**, Quantification and taxonomic classification of MAGs with a limited distribution, found only in one Arctic Region. Stacked barplots representative of the number of non pan-Arctic MAGs present only in one Arctic region, colored by phylum. The different Arctic regions and their season of sampling are found in axis X. Absolute number of MAGs per X axis category can be seen in the top barplots.



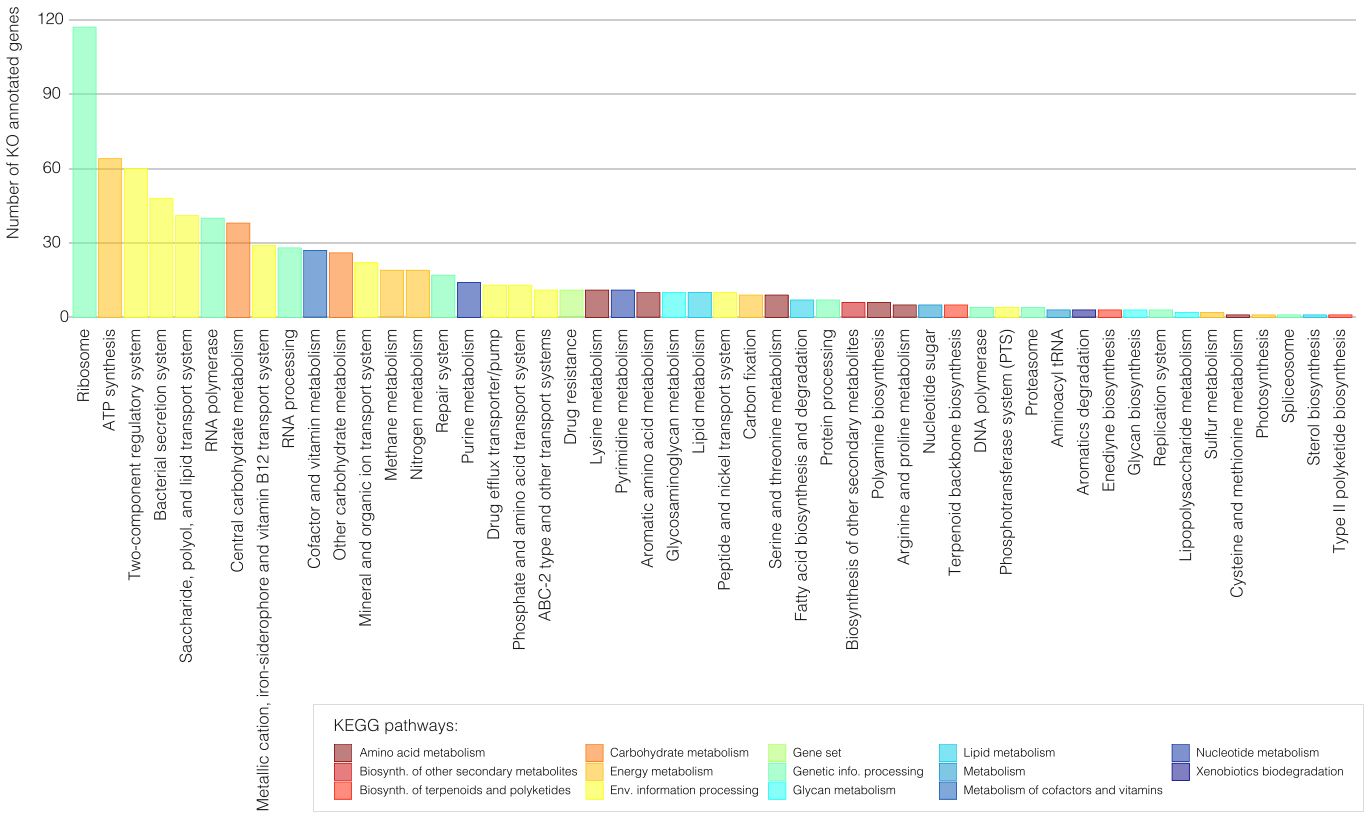
Extended Data Fig. 6 | Differences in estimated minimum growth times and optimal growth temperatures across biogeographic categories, mean 16S rRNA gene copy number across latitude and differences in estimated complete genome sizes and optimal growth temperatures between different niche breadth categories. a, Distribution of Minimum Generation Times estimated for each individual MAG, grouped by their biogeographical categorization (n = 153 MAGs classified as Arctic only, 123 MAGs classified as Arctic & NAO, 23 MAGs classified as bipolar, 4 MAGs classified as NAO only and 227 MAGs classified as Other latitudes). Data are shown as box plots (Tukey style): the lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles), the horizontal line indicates the median and the whiskers indicate the lowest and highest points within 1.5× the interquartile ranges of the lower (first) or upper (third) quartile, respectively. Data beyond the end of the whiiskers are outlying points and are plotted individually. Statistical support was calculated using the two-sided Dunnett-Tukey-Kramer Pairwise Multiple Comparison Test Adjusted for Unequal Variances and Unequal Sample Sizes (DTK) and CI 95. DTK test shows significant differences between the 'Arctic only' and the 'Other latitudes' MAGs. **b**, Distribution of Optimal Growth Temperatures estimated for each individual MAG, grouped by their biogeographical categorization (n = 153 MAGs classified as Arctic only, 123 MAGs classified as Arctic & NAO, 23 MAGs classified as bipolar, 4 MAGs classified as NAO only and 227 MAGs classified as Other latitudes). Boxplots describe the data as in **a**. DTK was performed as in **a** and shows significant differences (p-value < 0.05) between the 'Arctic only' and the 'Other latitudes' MAGs. **c**, Dots correspond to *Tara* Oceans samples from surface and subsurface chlorophyll maxima and are place across latitude depending on their estimated number of ribosomal copies (derived from miTAGs, see Methods). **d**, Distribution of estimated complete assembly size of MAGs based on their niche breadth category. Statistical support was calculated using the two-sided Dunnett-Tukey-Kramer Pairwise Multiple Comparison Test Adjusted for Unequal Variances and Unequal Sample Sizes (DTK) and CI 95. DTK test shows significant differences in estimated complete assembly size (p-value < 0.05) between MAGs classified as habitat specialists and those uncategorized (n = 38 MAGs classified as generalists, 111 MAGs classified as specialists, 381 uncategorised MAGs). Data are shown as box plots (Tukey style): the lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles), the horizontal line indicates the median and the whiskers indicate the lowest and highest points within 1.5× the interquartile ranges of the lower (first) or upper (third) quartile, respectively. Data beyond the end of the whiskers are outlying points and are plotted individually. **e)** Distribution of optimal growth temperatures of MAGs based on their niche breadth category (n = 38 MAGs classified as generalists, 111 MAGs classified as specialists, 381 uncategorised MAGs). DTK was performed as in **a** and shows (p-value < 0.05) between the generalists and the uncategorised MAGs.



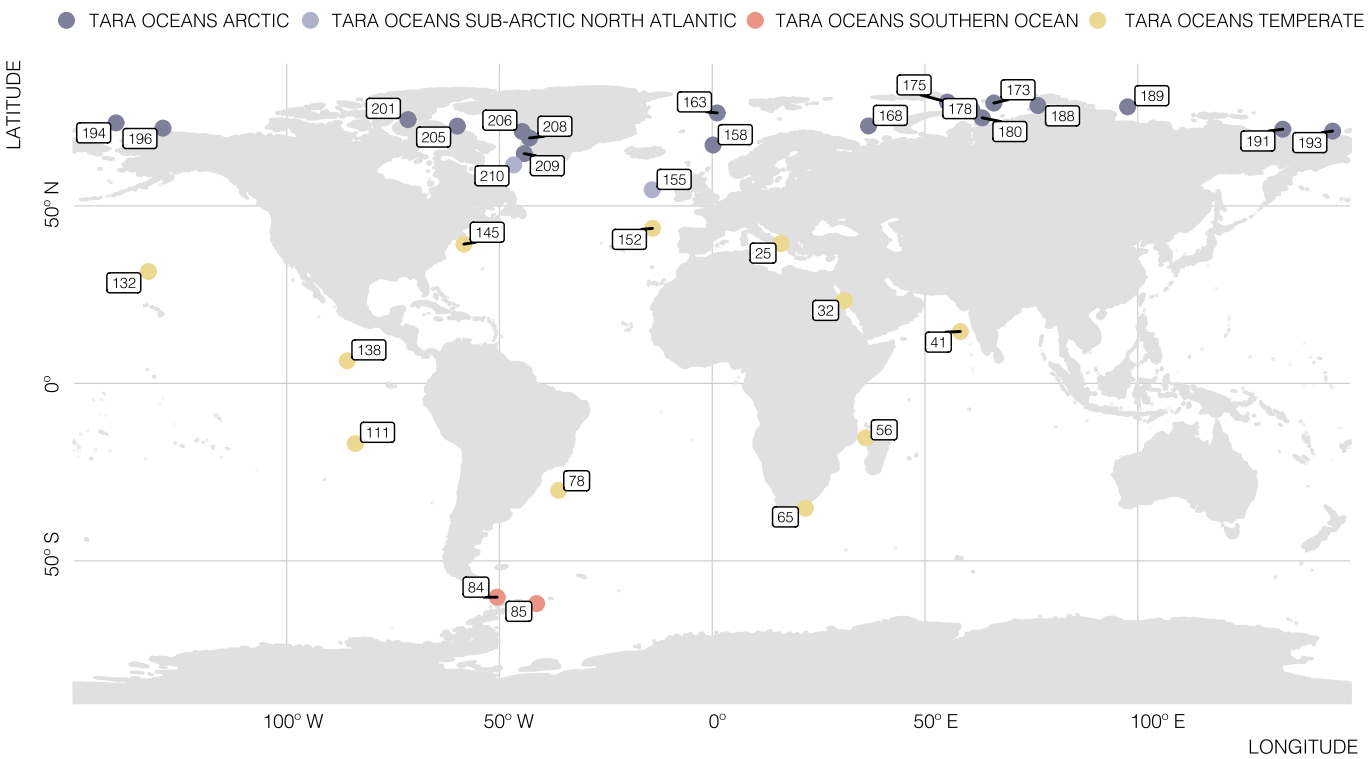
Extended Data Fig. 7 | Disentangling generalists and specialists within the 530 Arctic MAGs. a, Distribution of Arctic MAGs based on their mean read recruitments in Arctic metagenomic samples (RPKG, X axis) and their Levin's Index (i.e, niche breadth, Y axis). The color gradient depicts the occurrence (that is, % of samples where a given MAG is present) in the Arctic metagenomic dataset and shape indicates their niche breadth category (generalists, specialists and uncategorised). **b**, Number of habitat generalists (orange), specialists (blue) and uncategorised MAGs (grey) in each biogeographic category shown in bar plots (n = 530 MAGs examined over 32 Arctic metagenomes). The adjacent boxplots show the distribution of assembly sizes within each subcategory (upscaled to 100% of genome completeness) and statistically significant differences have been highlighted with an asterisk (DTK test, p-value < 0.05). Box plots are presented horizontally and in Tukey style: the lower (left) and upper (right) hinges correspond to the first and third quartiles (the 25th and 75th percentiles), the vertical line indicates the median and the whiskers indicate the lowest and highest points within 1.5× the interquartile ranges of the lower (first) or upper (third) quartile, respectively. Data beyond the end of the whiskers are outlying points and are plotted individually. Statistical support was calculated using the two-sided Dunnett-Tukey-Kramer Pairwise Multiple Comparison Test Adjusted for Unequal Variances and Unequal Sample Sizes (DTK) and CI 95. Adjacent stacked barplots indicate their taxonomic composition at the phylum level. Asterisks in the taxonomic annotation legend indicate phyla from domain Archaea, lack of asterisk indicates domain Bacteria. **c**, Abundances of generalists (n = 38 MAGs; orange), specialists (n = 111 MAGs; blue) and uncategorised (n = 381 MAGs; grey) MAGs in Arctic metagenomes (n = 32 samples, 18 SRF, 7 SCM, 7 MES; filled boxplots) and metatranscriptomes (n = 29 samples, 18 SRF, 7 SCM, 4 MES; empty boxplots) across the three ocean layers. Boxplots describe the data as in **b**. There are no significant differences between the groups.



Extended Data Fig. 8 | Seawater temperature in samples where maximum metagenomic recruitment occurred per MAG, per niche breadth category. a, Histogram of the number of MAGs in the station where their maximum metagenomic RPKG occurred, colored by niche breadth category and horizontally separated by layer. Bottom heatmap above X axis represents the temperature of each sample. **b**, Distribution of temperatures in those samples where maximum metagenomic RPKG per MAG occurred, by niche category and layer (n = 38 MAGs classified as generalists and 111 MAGs classified as specialists tested against 32 metagenomes, including 18 SRF, 7 SCM and 7 MES). Data are shown as box plots (Tukey style): the lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles), the horizontal line indicates the median and the whiskers indicate the lowest and highest points within 1.5× the interquartile ranges of the lower (first) or upper (third) quartile, respectively. Data beyond the end of the whiiskers are outlying points and are plotted individually.



Extended Data Fig. 9 | Genes found in specialists but not in generalists. Quantification of genes annotated against KEGG database that were found in specialist MAGs but not in generalist, colored by pathway.



Extended Data Fig. 10 | Map with the reference stations with metagenomic and metatranscriptomic samples used in the study. Samples are colored based on the expedition. Supplementary Table 4 contains more details about environmental metadata of these stations.

nature research

Corresponding author(s): Silvia G. Acinas

Last updated by author(s): Aug 23, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Metagenomic reads were co-assembled with megahit v1.1.2. Contigs were de-replicated with cd-hit-est v4.6.8-2017-0621. Binning used bowtie2 v2.3.2 and metabat2 v2.12.1. Bin quality used checkM v1.0.11 and curation was made with Geneious v10.2.4. Taxonommic annotation was done using GTDBTk v0.3.2 and SINA v1.2.11 against SILVA132. Functional annotation used prokka v1.13, KEGG orthology database (release 89.1) and aligner diamond v0.9.22, and PFAM database release 31.0 and hmmer v3.1b2. ANI was calculated with fastANI v1.2 and AAI used compareM v0.0.23. Read recruitment used Nucleotide-Nucleotide BLAST v2.7.1+.
Data analysis	Data analysis was done in R v3.4.0 using Rstudio v1.0.143. Horizontal genomic coverage was assessed with package GenomicRanges v1.34.0. Niche breadth analysis used package EcolUtils v0.1, that depends on packages vegan v2.5-6 and spaa v0.2.2. Figures were done using ggplot2 v3.3.2 and post-processing was done in Adobe Illustrator CC 2018. Dunnett-Tukey-Kramer Pairwise Multiple Comparison Test was done with package DTK v3.5, MANOVA test used function adonis from vegan v2.5-6 and NMDS used function metaMDS from vegan v2.5-6 using 100 iterations. Minimum Generation Times prediction used Growthpred v1.08 using a snakemake pipeline available at https://gitlab.univ-nantes.fr/combi-ls2n/growthsnake. Phylogenies used Clustal Omega v1.2.3 for alignment and FastTree v2.1.11 for phylogenetic reconstruction. Phylogenetic tree editing was done in iTol (https://itol.embl.de).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Accession numbers for the data used and generated in this study can be found in Supplementary Table 12, which includes the Arctic MAGs Catalogue and their functional annotation (EBI Biostudies ID: S-BSST451) and the co-assembly of metagenomic samples used to generate the metagenomic bins (ENA PRJEB41575). Accession numbers for the metagenomic and metatranscriptomic samples used in the Fragment Recruitment Analyses can be found in Supplementary Table 13. Source data has also been provided for all main figures and extended data figures. Public available datasets used in this study include the checkM v1.0.11 database (https://github.com/Ecogenomics/CheckM/releases/tag/v1.1.0), the GTDB database release 89 (https://data.gtdb.ecogenomic.org/releases/release89/), the SILVA 132 database (https://www.arb-silva.de/documentation/release-132/), KEGG release 89.1 (https://www.genome.jp/kegg/docs/relnote.html) and PFAM database release 31.0 (http://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam31.0/).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Co-assembly used the 41 metagenomes collected during the Tara Oceans Polar Circle, including Arctic and Sub-Arctic North Atlantic metagenomes. These metagenomes are representative of all Arctic Regions surrounding the Arctic ice-cap and represent a balance between accurate representation of the Arctic Ocean, the limitations of navigating such waters and the amount of water to be collected by the scheduled time with the team that could be hosted in the Tara vessel. Further analyses focus on Arctic regions were done with the 37 metagenomes with absolute latitude values above 64º. The 27 metagenomes from temperate latitudes are representative all the oceanic regions sampled in the Tara Oceans expeditions and have enough sequencing depth to be compared to the Arctic dataset. Another requisite of these samples was that they had a corresponding metatranscriptome with enough sequencing depth. The Tara Arctic MAGs catalogue includes 530 genomes selected according to common quality standards suggested in previous literature (Bowers et al., 2017).
Data exclusions	There was no data exclusion in this analysis.
Replication	The unique nature of the seawater sampling (unique cruises) and filtration (multitude of liters collected through sampling protocols that last hours) does not allow replication.
Randomization	Randomization was applied in the classification of Niche Breadth Analysis in 1000 permutations of Levin's Index calculations to avoid sampling or sequencing biases that would affect the MAG's read recruitments. Permutations were also applied in Permutational MANOVA test applied to the ordination of samples based on their MAG composition.
Blinding	Blinding was not necessary for the development of this study as it is mostly descriptive and treats environmental sequencing samples that cannot be influenced by human manipulation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging