

Supplementary Methods

Metagenomic assembly and co-assembly

Details of seawater sampling, DNA extraction, shotgun sequencing, and generation of single sample metagenome assemblies were previously described in¹. In the current study, we generated metagenome co-assemblies of collection of samples from the Canada Basin and Amundsen Gulf (**Supplementary Data 1**) to recover a greater diversity of MAGs. All metagenome co-assemblies were generated using MEGAHIT v.1.2.7² with k-mer sizes of 27, 37, 47, 57, 67, 77, and 87.

Binning, dereplication, and taxonomic classification of MAGs

The input metagenome reads for each co-assembly were mapped to the co-assemblies with BWA v.0.7.17 using the mem option³. The mapping results were processed using jgi_summarize_bam_contig_depths. Contig binning was performed for each of the assemblies using MetaBAT2 v.2.12.14 with default settings, resulting in 4,824 genome bins. Genome quality was evaluated using CheckM v.1.0.11⁵ with the lineage_wf workflow. The 924 bins with >50% completeness and <10% contamination and strain heterogeneity were considered at least medium-quality MAGs. Dereplicated MAGs with >50% completeness and <10% contamination and strain heterogeneity were taxonomically classified with the GTDB-tk v.1.3.0 using the classify_wf workflow⁶. The medium-quality MAGs were dereplicated with dRep using 95% ANI cut-off to remove species level redundancy⁷, resulting in 663 representative MAGs.

MAG abundances and distributions across the global ocean

To determine the coverage of MAGs in the Canada Basin metagenomes (**Supplementary Data 1**) and 129 TARA global oceans metagenomes⁸ and 20 Southern Ocean metagenomes⁹, trimmed metagenome reads were initially recruited to a concatenation of the MAGs at a minimum sequence identity of 98% using BBMap v.35. The mapping files were parsed to extract the mapping information for each individual MAG using a custom Unix script which implemented SAMtools¹⁰. Mapping files for each MAG were converted from SAM to sorted BAM format files

using SAMtools¹⁰. Horizontal coverage was calculated following the method in¹¹ in which the average sequencing depth is truncated to the central 80% of the mapped positions (TAD₈₀) normalized by the number of genome equivalents. The number of genome equivalents in each metagenome was estimated using MicrobeCensus v.1.1.0¹². A principal coordinate analysis was performed to assess the community composition of medium-quality MAG assemblages based on the Bray-Curtis dissimilarities in TAD₈₀ abundances using the function pcoa() in the R package vegan¹³.

Functional annotation of MAGs

Gene prediction and annotation was performed using Prokka v.1.12¹⁴, which implemented Prodigal v.2.6.3 for protein coding genes¹⁵, Barrnap for ribosomal RNA genes, and ARAGORN v.1.2¹⁶ for transfer and transfer-messenger RNA genes. Functional annotation of proteins was performed using Prokka v.1.12¹⁴ with default settings. Functional annotation of proteins was also performed using KofamKOALA¹⁷ with default settings and a threshold score greater than 0.7 or higher and an e-value below 1×10^{10} or lower to identify metabolic genes within the MAG catalogue.

Supplementary Results

The Western Arctic Ocean MAG catalogue

We generated a MAG catalogue from an eight station transect of the western Arctic Ocean that included the Amundsen Gulf and a latitudinal gradient of the Canada Basin (73-81°N) (**Figure S1a**). We performed a metagenomic survey of the oligotrophic Canada Basin samples (Table S1). All stations were ice-free during the sampling period (September 2017), except for the most northern station (CBN3). The survey targeted distinct water layers, including the surface mixed layer (SML; 5 m and 20 m), the subsurface chlorophyll maximum (SCM; 55-95 m), Pacific-origin winter water (PWW; 90-250 m) and Atlantic-origin water (AW; 360-1000 m) (**Figure S1b**). To maximize the MAG diversity recovered, we performed contig binning on each of the 31 single metagenome assemblies, as well as on three metagenome coassemblies generated from 1) all 24 Canada Basin samples (CB coassembly), 2) 11 Canada Basin samples from the SML and SCM (Upper water coassembly), and 3) 7 Amundsen Gulf samples (AG coassembly). In total,

we recovered 924 MAGs exceeding MIMAG medium quality through an additional <10% strain heterogeneity criterion. MAGs were dereplicated across assemblies/ coassemblies (average nucleotide identity $\geq 95\%$), resulting in a genomospecies-level set of 663 MAGs. Of the genomospecies, 84% (559) were represented by MAG originating from metagenome coassemblies and 16% (104) originated from single metagenome assemblies (**Supplementary Data 2**), demonstrating the value of combining coassembly and single assemblies in generation of MAG datasets.

To map MAG biogeography across the water layers of the Canada Basin, we calculated the central 80% truncated average depths of coverage (TAD_{80}) from fragment recruitment. Based on TAD_{80} , between 62 and 471 MAGs were detected across samples, with a pattern of increased observed richness with depth (**Figure S1c**). Correspondingly, fragment recruitment to the MAG dataset was lowest in the surface waters (average 8.34%) and increased with depth with the highest mean read recruitment from the PWW samples (26.34%), indicating less genomic representation from the surface waters (**Figure S1d**). A principal coordinate analysis demonstrated distinct MAG assemblages were strongly structured by the stratified waters masses of the Western Arctic Ocean (**Figure S1e**).

We next investigated the distribution of the Arctic Ocean MAG catalogue across other regions of the global ocean. The Arctic Ocean MAG catalogue recruited on average 0.05% (0.04% and 0.08%) of reads from surface and chlorophyll maximum layer metagenomes from other regions of the ocean (**Figure S1d**). Recruitment was greater for mesopelagic metagenomes from lower latitudes (0.14 to 6.19 %) and was greatest for a single mesopelagic metagenome from the Southern Ocean (9.74 %). This restricted distribution across global oceans demonstrate a strong Arctic Ocean endemism of the MAGs.

The MAG dataset represented a wide phylogenetic range, including 619 bacterial (**Figure S2a**) and 44 archaeal taxa (**Figure S2b**) of which the majority (93%) were novel at the species level (**Figure S2c**).

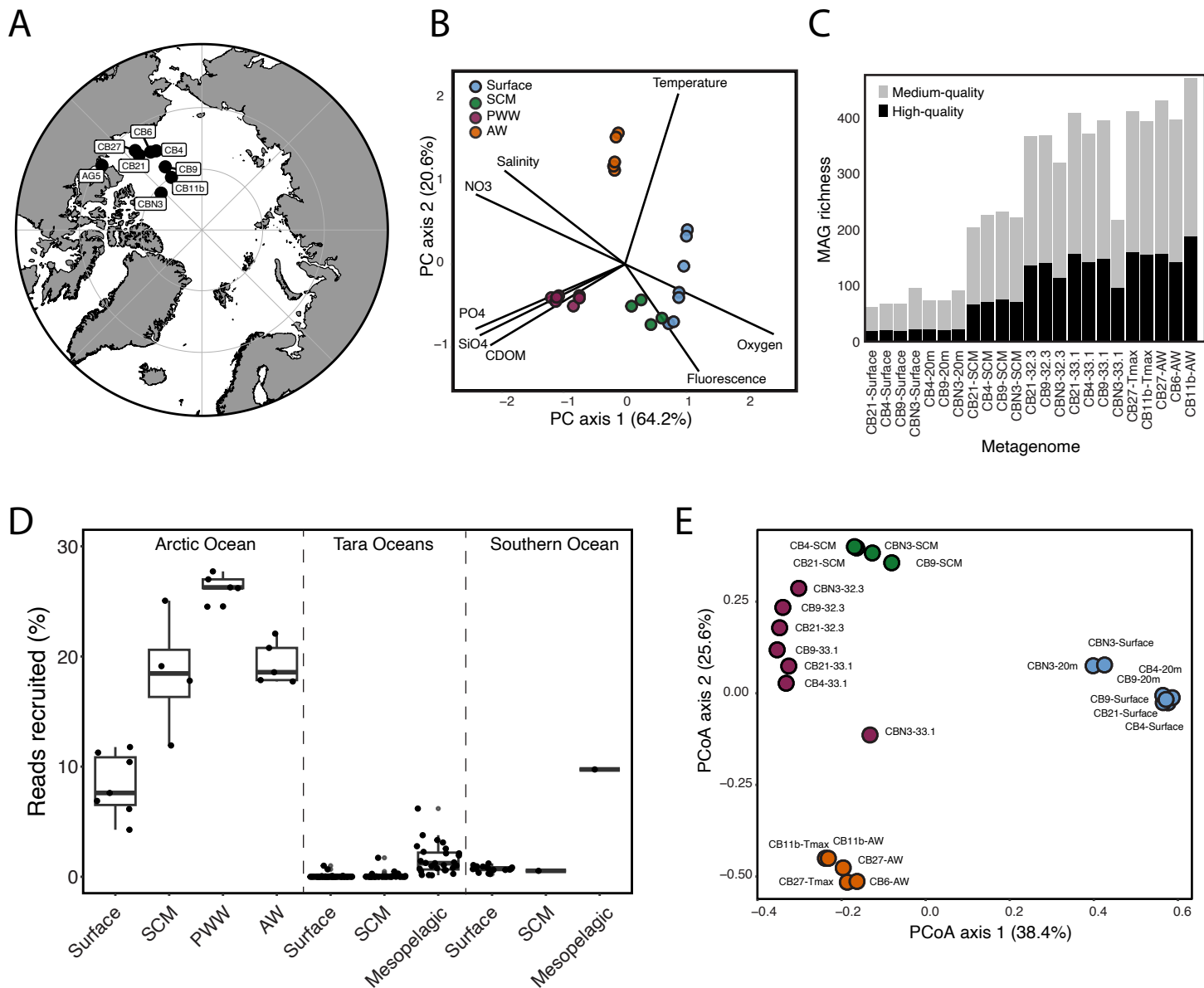
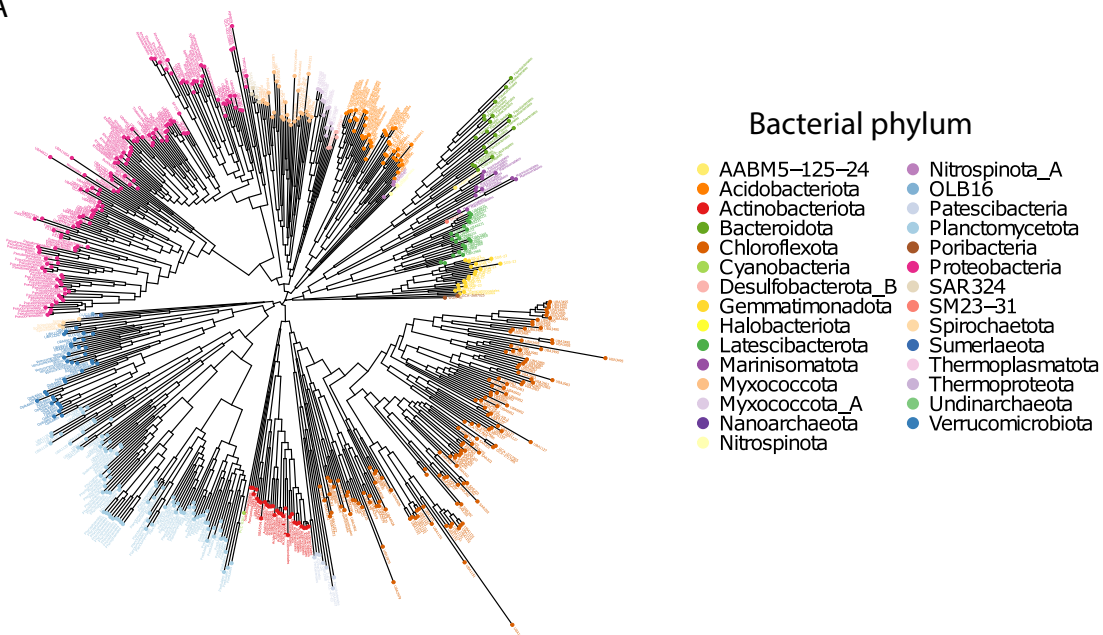
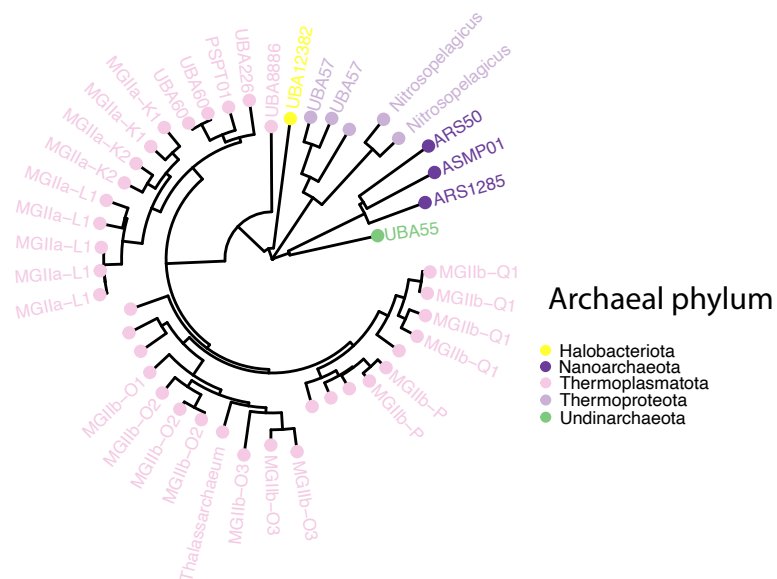


Figure S1. Vertical structuring of the microbial community in the Canada Basin, Arctic Ocean. **(A)** Map of sampling stations in the Canada Basin, Arctic Ocean. The map was generated using ggOceanMaps in R. **(B)** Principal component analysis of the environmental characteristics of the Canada Basin metagenome sampling locations. **(C)** The number of MAGs recovered from metagenome assemblies. **(D)** Abundance of the MAG collection in the Arctic Ocean and global oceans based on fragment recruitment **(E)** Principal coordinate analysis of the taxonomic variation among Arctic Ocean MAG assemblages based on TAD₈₀ community fraction of the MAGs.

A



B



C

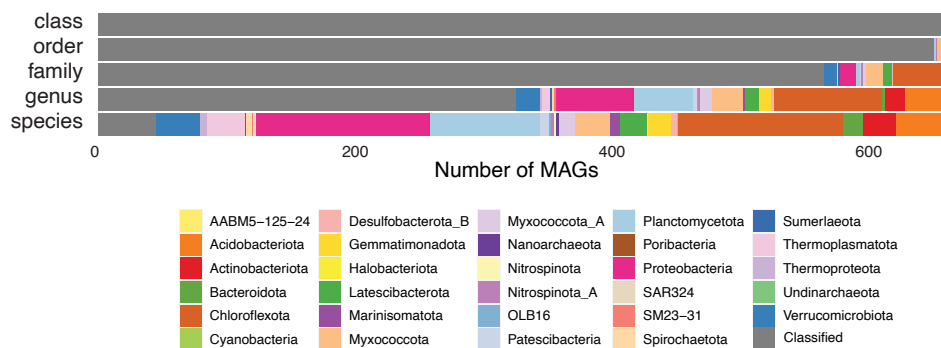


Figure S2. A MAG catalogue from the western Arctic Ocean. Phylogenetic diversity of (A) bacterial and (B) archaeal MAGs in the Western Arctic Ocean MAG catalogue. For the archaeal tree, taxa IDs were not assigned for some MAGs at this taxonomic level. (C) Phylogenetic novelty of the MAGs. The bars represent the number of MAGs within that taxonomic group that are novel at each of the five taxonomic levels.

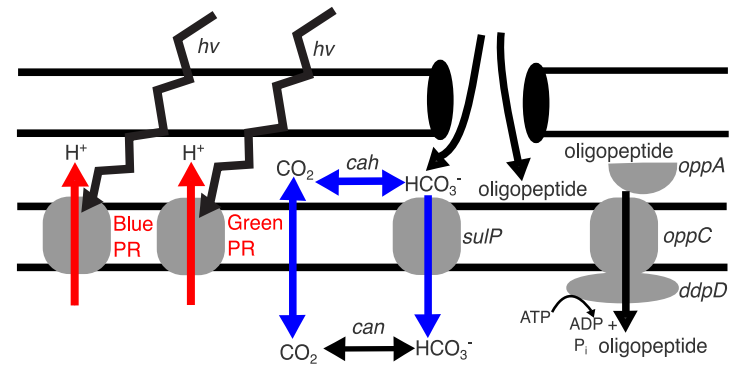
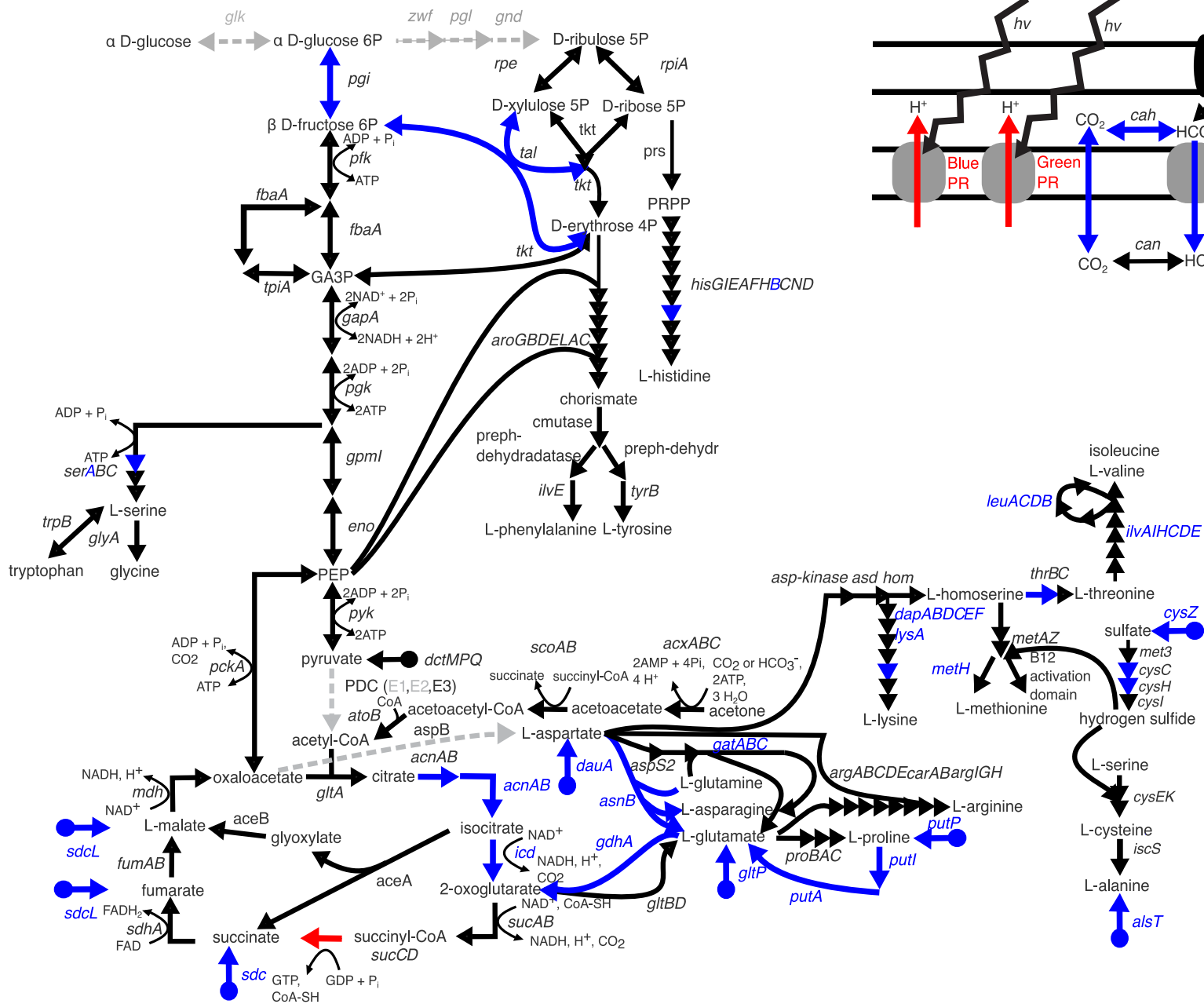


Figure S3. Metabolic reconstruction of the central carbon metabolism of *P. arcticus* and *P. hydrocarbonoclasticus*. Black arrows indicate genes of pathways conserved between *P. arcticus* and *P. hydrocarbonoclasticus*. Red arrows indicate genes of pathways possessed only by *P. arcticus*. Blue arrows indicate genes of pathways possessed only by *P. hydrocarbonoclasticus*. Grey dashed arrows indicate genes of pathways that are absent in *P. arcticus* and *P. hydrocarbonoclasticus*.

Supplementary References

1. Grevesse, T., Guéguen, C., Onana, V. E. & Walsh, D. A. Degradation pathways for organic matter of terrestrial origin are widespread and expressed in Arctic Ocean microbiomes. *Microbiome* **10**, 237 (2022).
2. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
3. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
4. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
5. 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).
6. 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).
7. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
8. Karsenti, E. *et al.* A Holistic Approach to Marine Eco-Systems Biology. *PLoS Biol.* **9**, e1001177 (2011).
9. Brown, M. V. *et al.* Global biogeography of SAR11 marine bacteria. *Mol. Syst. Biol.* **8**, MSB201228 (2012).
10. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
11. Rodriguez-R, L. M., Tsementzi, D., Luo, C. & Konstantinidis, K. T. Iterative subtractive binning of freshwater chronoserries metagenomes identifies over 400 novel species and their ecologic preferences. *Environ. Microbiol.* **22**, 3394–3412 (2020).
12. Nayfach, S. & Pollard, K. S. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol.* **16**, 51 (2015).
13. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).

- 34 14. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069
35 (2014).
- 36 15. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
37 identification. *BMC Bioinform.* **11**, 119–119 (2010).
- 38 16. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes
39 in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).
- 40 17. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and
41 adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

42