ISME

www.nature.com/ismecomms

ARTICLE OPEN

Linking prokaryotic genome size variation to metabolic potential and environment

Alejandro Rodríguez-Gijón[1,2]✉, Moritz Buck[3], Anders F. Andersson[2,4], Dandan Izabel-Shen[1], Francisco J. A. Nascimento[1,5] and Sarahi L. Garcia[1,2]✉

While theories and models have appeared to explain genome size as a result of evolutionary processes, little work has shown that genome sizes carry ecological signatures. Our work delves into the ecological implications of microbial genome size variation in benthic and pelagic habitats across environmental gradients of the brackish Baltic Sea. While depth is significantly associated with genome size in benthic and pelagic brackish metagenomes, salinity is only correlated to genome size in benthic metagenomes. Overall, we confirm that prokaryotic genome sizes in Baltic sediments (3.47 Mbp) are significantly bigger than in the water column (2.96 Mbp). While benthic genomes have a higher number of functions than pelagic genomes, the smallest genomes coded for a higher number of module steps per Mbp for most of the functions irrespective of their environment. Some examples of this functions are amino acid metabolism and central carbohydrate metabolism. However, we observed that nitrogen metabolism was almost absent in pelagic genomes and was mostly present in benthic genomes. Finally, we also show that Bacteria inhabiting Baltic sediments and water column not only differ in taxonomy, but also in their metabolic potential, such as the Wood-Ljungdahl pathway or the presence of different hydrogenases. Our work shows how microbial genome size is linked to abiotic factors in the environment, metabolic potential and taxonomic identity of Bacteria and Archaea within aquatic ecosystems.

ISME Communications; https://doi.org/10.1038/s43705-023-00231-x

INTRODUCTION

Genomes in Bacteria and Archaea are information-rich [1], and known to range in size from 0.1 to 16 million base pairs (Mbp) [2]. They can vary over evolutionary time through genomic expansions and contractions via genetic drift, selection, homologous recombination, deletions and insertions [3–9]. Moreover, evolutionary studies have revealed extremely rapid and highly variable flux of genes [10] with evolutionary forces acting on individual genes [5]. With all these evolutionary forces acting on the genes, we can presume that gene content and, by consequence, genome size has an ecological meaning. Indeed, genome size has been linked to phylogenetic history [11, 12], lifestyle such as free-living, particle attached or host-associated [4, 13, 14], or environment such as marine, freshwater, different types of sediments or different hosts in host-associated microorganisms [2, 15, 16]. We aim to delve into the ecological implications of genome size in aquatic microorganisms, with emphasis on metabolic potential using a brackish environment as a model.

In the last decade, aquatic microorganisms have been extensively sampled and now have a large representation in genomic and metagenomic datasets [17]. Their genome size spans from 0.5 to 15 Mbp with an average of 3.1 Mbp [2]. Aquatic environments are heterogeneous and many different abiotic factors, such as salinity and depth, could be linked to microbial genome size variation. For example, pelagic microbes inhabiting deep marine environments are estimated to present bigger genome sizes than those in shallow marine waters [18, 19]. Within freshwater ecosystems, isolates from the family Methylophilaceae (class Gammaproteobacteria) show a smaller genome size for pelagic than for sediment dwellers [20]. Additionally, two studies have shown that marine Cyanobacteria have smaller genome sizes than freshwater [21, 22]. This literature already provides some insights on how genome size is linked to the environmental heterogeneity of freshwater and marine ecosystems. Yet, the studies are limited either to a specific marine station, or specific microbial lineages and it remains a question if these findings are applicable more widely. Moreover, genome size variation in the brackish realm remains debated: Actinobacteria in the brackish Baltic Sea show bigger genome sizes than in freshwaters [23], while picocyanobacteria show the opposite trend [22]. Additionally, further research must be done to elucidate the link between genome size and abiotic factors within aquatic environments, particularly brackish water bodies.

In the link between abiotic factors in the environment and genome size, gene content is selected accordingly. Metage-

[1]Department of Ecology, Environment and Plant Sciences, Stockholm University, Stockholm 106 91, Sweden. [2]Science for Life Laboratory, Stockholm, Sweden. [3]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden. [4]Department of Gene Technology, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Stockholm, Sweden. [5]Baltic Sea Centre, Stockholm University, Stockholm, Sweden.
✉email: alejandro.rgijon@gmail.com; sarahi.garcia@su.se

Received: 26 October 2022 Revised: 2 March 2023 Accepted: 14 March 2023
Published online: 27 March 2023

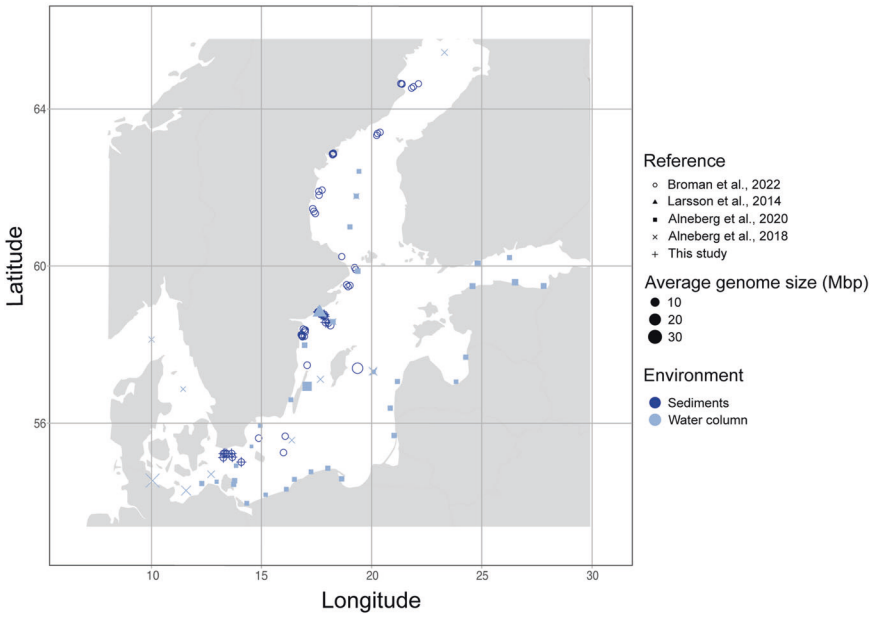

Fig. 1 Overview of the sampling locations and average genome size (AGS) of metagenomes (108 from sediments in dark blue, and 111 from the water column in light blue). This figure shows the geographic location of all metagenomes used in this study. For exact coordinates see Table S1. Shape type indicates the reference and shape size indicates the AGS of the given metagenome.

nomic studies show that a vast majority of the genes in Bacteria and Archaea are specific to particular environments, whereas very few genes are being shared between environments [24]. This remarks how relevant is the relationship between niche specificity and lineage specific functional traits [25]. These functional capabilities also differ between water column and sediments in Bacteria and Archaea in both marine [26, 27] and freshwater environments [28]. Since gene repertoires and genome size are related, they must be considered together with environmental gradients to better understand niche-specificity and the ecology of different prokaryotic lineages.

In this research article, we provide a comprehensive analysis to show the ecological implications of genome size variation of Bacteria and Archaea in pelagic and benthic communities in the Baltic Sea. Specifically, we investigate; (i) how genome size varies across abiotic factors and taxonomic lineages of Bacteria and Archaea from sediments and water column, (ii) what relationship can be found between genome size and the number of metabolic capabilities in Bacteria in the Baltic Sea, and (iii) which taxa and metabolic pathways contrast between pelagic and benthic Bacteria. To achieve this, we selected 111 pelagic and 59 benthic metagenomic samples that were previously published, and we provide one new and unpublished benthic dataset with 49 metagenomic samples. We use these 219 metagenomes to study genome size distribution across sediments and water column in the Baltic Sea (Fig. 1 and Table S1). For this, we use two different approaches to study genome size: the estimated average genome size (AGS) per metagenomic sample, and the estimated genome size of bacterial and archaeal metagenome-assembled genomes (MAGs). Our results show that Bacteria and Archaea with larger genomes in the sediments present lower coding density than the smaller genomes in the water column. We also find that microorganisms inhabiting the Baltic benthos have more metabolically-versatile genomes than pelagic prokaryotes, which mean they code for a wider range of metabolic capabilities. Finally, we also find that functions involved in the nitrogen metabolism are disproportionately more detected in benthic bacteria in the Baltic Sea.

RESULTS AND DISCUSSION

While depth is significantly associated with genome size in benthic and pelagic metagenomes, salinity is only correlated to genome size in benthic metagenomes

First, we calculated the average genome size (AGS) of metagenomes across the latitudinal gradient of the Baltic Sea comparing benthic and pelagic metagenomes. We observed that sediment-dwelling microbial communities present significantly larger AGS (mean = 6.01 Mbp, $n = 108$) than pelagic communities (mean AGS = 5.40 Mbp, $n = 69$) (Wilcoxon test, $p < 0.01$) (Fig. 2A). We then evaluated the relationships between AGS of metagenomes and each of the environmental variables (depth, salinity, temperature, and oxygen concentration) independently and their interactions (ANOVA type II). The AGS in the pelagic metagenomes is significantly associated only with depth and shows a negative correlation (Supplementary Material 1 and Fig. S1). Previous marine analysis have found the opposite effect, bigger genome sizes in deeper areas [19]. However, our pelagic analysis only covers 5 meters of depth and explains 19% of the genome size variation in pelagic metagenomes. AGS in sediment metagenomes is significantly associated with both salinity and depth and most of the interactions of these variables (Supplementary Material 1). Both depth and salinity have a weak positive correlation with AGS in the sediment samples (Fig. S1). To our knowledge, this is the first time where a positive correlation between genome size and salinity is reported for brackish sediment microorganisms.

Although the ANOVA did not show a significant effect of water oxygen concentration on genome size, we further investigate the effect of bottom water O_2 concentration in the AGS of metagenomes from sediments. We separated these metagenomes into those from oxygen concentration 0 to 2 mg/L (mostly metagenomes from the dead zone) and those metagenomes with oxygen concentration 2–12.45 mg/L. We observe that benthic metagenomes from lower oxygen concentration (mean AGS = 7.08 Mbp, $n = 14$) present significantly bigger genome sizes than those in sediments with higher oxygen concentration (mean AGS = 5.85 Mbp, $n = 94$) (Wilcoxon test, $p < 0.01$) (Fig. 2B). Complementarily, previous results show that the dead zone bacteria tend to be more metabolically similar to each other when

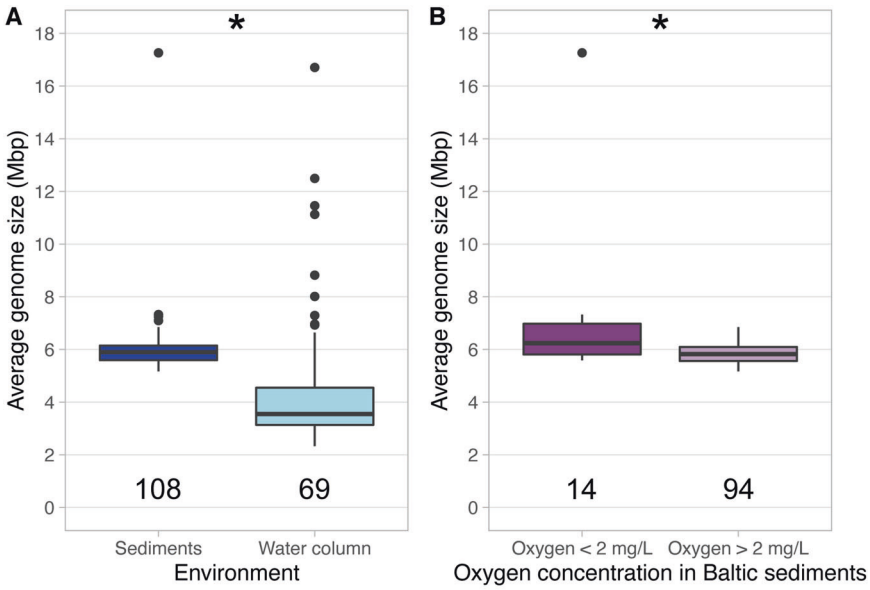

Fig. 2 Boxplots showing the AGS distribution of Baltic metagenomes. A Indicates the AGS distribution in both water column and sediments. **B** Indicates the AGS distribution in metagenomes from sediments across the oxygen gradient (two groups, from 0 to 2 and from 2 to 12.45 mg/L). Stars in both panels indicate significant differences $p < 0.05$ (Wilcoxon non-parametric test).

compared to bacteria from oxic sediments [29]. Our observation confirms a recent study that observed obligate anaerobes from diverse environments present bigger genome sizes than microaerophilic microorganisms [30]. Moreover, previous studies show a positive correlation between genome size and nutrient concentration [19, 31, 32]. Dead zones in the Baltic Sea are characterized by anthropogenic eutrophication [33], which would also promote bigger genome sizes. Altogether, these results indicate that high nutrient concentration and low oxygen concentrations in Baltic Sea dead zones may select for prokaryotes with bigger genome sizes.

While average benthic estimated genome size is bigger than pelagic, the biggest genomes in the Baltic Sea were found in the water column

To compare genome size between pelagic and benthic bacteria and archaea, we used metagenome-assembled genomes (MAGs; >75% completeness and <5% contamination) from our metagenome datasets. Our dataset compiles 216 MAGs from the sediments that dereplicated into 56 representative genomospecies (95% average nucleotide identity). Additionally, 1920 pelagic MAGs were dereplicated into 340 representative genomospecies. We observe that 12 phyla were detected in both sediments and water column, while 16 phyla were specific to either habitat. Seven phyla were found specific to the sediment (14 representative MAGs) and nine phyla were specific to the water column (29 representative MAGs) (Table 1). Interestingly, only one genomospecies representative was binned from both sediments and water column. This genome representative belongs to genus *Mycobacterium* (phylum Actinobacteriota, mOTU_124/pelagic and mOTU_027/sediments in Table S2), a genus that is not commonly found on brackish surface waters [34]. However, this genus was found to be abundant in sediments in some regions of the Baltic Sea, especially in anoxic areas close to Landsort [35]. These differences on taxonomical composition in microbial communities between water column and sediments altogether with latitudinal changes in microbial biodiversity in the Baltic Sea [34] show how heterogeneous is the microbial composition of brackish environments.

Additionally, the eight largest representative genomes of the dataset were observed in the water column and belonged to phylum Planctomycetota (family Planctomycetaceae) (7.95–9.69 Mbp). It has been previously observed that Planctomycetota is the phylum containing the aquatic MAG with the biggest known estimated genome size (14.93 Mbp) [2]. These large genomes contain large collections of genes that could be linked to the very complex cell structures and chromosomes observed in this phylum [36]. Still, further research is necessary to understand if extant genome size in phylum Plactomycetota is the result of ecological adaptation to abiotic gradients. On the other side of the genome size spectrum, the representative MAG with the smallest estimated genome size belongs to class Alphaproteobacteria (family Pelagibacteraceae, 1.08 Mbp; Table 1). Bacteria from this family have been widely reported to be streamlined, abundant and ubiquitous across all salinity gradients [37, 38].

Altogether, representative MAGs from sediments presented an average estimated genome size of 3.47 Mbp, which was significantly higher than for water column MAGs (2.96 Mbp) ($p < 0.01$) (Fig. 3A). This was true also at the phyla, class, and order level (Fig. 3C, D). Bacteria from sediments presented bigger estimated genome size on average (3.67 Mbp), followed by pelagic Bacteria (2.98 Mbp), pelagic Archaea (1.97 Mbp) and sediment Archaea (1.43 Mbp) (Fig. 3B). This is supported by previous results, as Bacteria show bigger genome sizes than Archaea regardless of the environment [2]. Moreover, the average estimated genome size of Baltic sediments (3.47 Mbp) is more similar to that of terrestrial microbial genomes (3.7 Mbp) [2]. Additionally, genomes in the Baltic sediments have also lower coding density than pelagic (Figure S2). Our results corroborate previous findings that streamlining is common in pelagic marine environments [7, 37, 39, 40] and pelagic brackish environments [23].

In our study, the average estimated AGS for the sediments (6.01 Mbp) and water column (4.44 Mbp) is larger than the average estimated genome size of the MAGs assembled and binned from the sediments (3.47 Mbp) and water column (2.96 Mbp), respectively (Figs. 2A and 3A). We calculated the AGS to estimate the average genome size of the whole microbial community. We used MicrobeCensus as a robust and accurate tool to calculate AGS [41]. However, this AGS of metagenomes overestimates the genome size because of the viral and eukaryotic reads that might be present in the sample [42, 43]. On the other hand, there are

Table 1. Summary of all 56 sediment and 340 water column representative MAGs (95% average nucleotide identity) with >75% completeness.

Phyla	Environment	*n*	Smallest estimated genome size (Mbp)	Largest estimated genome size (Mbp)	Average GC (%)	Average coding density (%)
Actinobacteriota	Sediments	6	2.84	3.7	67.55	91.63
	Water column	59	1.25	4.53	54.23	93.74
Bacteroidota	Sediments	2	3	4.61	41.08	89.69
	Water column	86	1.38	4.71	40.7	92.97
Chloroflexota	Sediments	3	2.38	5.19	63.57	91.14
	Water column	5	1.25	5.91	57.42	90.26
Desulfobacterota	Sediments	8	2.71	4.66	49.32	84.55
	Water column	1	–	2.73	54.4	87.06
Gemmatimonadota	Sediments	1	–	2.85	66.66	93.99
	Water column	1	–	4.32	62.47	91.21
Myxococcota	Sediments	3	4.95	6.74	63.54	90.84
	Water column	1	–	7.66	66.33	90.86
Planctomycetota	Sediments	1	–	5.09	66.06	90.33
	Water column	29	3.18	9.69	59.01	88.23
Proteobacteria (Alfa)	Sediments	1	–	3.43	59.14	91.26
	Water column	46	1.08	5.92	51.88	91.98
Proteobacteria (Gamma)	Sediments	10	1.93	4.8	59.56	90.96
	Water column	54	1.18	4.19	50.48	92.09
Verrucomicrobiota	Sediments	3	2.97	3.69	56.16	89.4
	Water column	25	1.98	6.82	57.36	90.41
Thermoplasmatota	Sediments	1	–	1.51	54.16	91.35
	Water column	2	2.35	2.68	41.63	93.32
Thermoproteota	Sediments	3	1.42	1.58	35.32	89.34
	Water column	2	1.42	1.48	31.86	89.96
Acidobacteriota	Sediments	5	4.13	5.64	65.98	91.6
Desulfobacterota D	Sediments	1	–	2.33	45.15	88.3
Desulfobacterota E	Sediments	2	2.35	2.74	64.01	91.94
Nitrospirota	Sediments	3	2.86	4	54.01	86.72
Omnitrophota	Sediments	1	–	1.47	41.17	91.99
Zixibacteria	Sediments	1	–	3.68	41.12	90.73
Micrarchaeota	Sediments	1	–	1.14	48.22	93.23
Bdellovibrionota	Water column	1	–	3.73	49.16	93.45
Bdellovibrionota C	Water column	1	–	3.33	47.41	85.46
Campylobacterota	Water column	2	2.6	3.3	37.02	93.25
Cyanobacteria	Water column	16	1.98	6.03	51.8	88.45
Firmicutes	Water column	3	1.09	1.16	38.25	93.69
Krumholzibacteriota	Water column	1	–	3.34	63.78	92.99
Marinisomatota	Water column	2	2.62	3.11	39.47	91.47
Nitrospinota	Water column	2	3.18	3.35	47.22	87.87
Nanoarchaeota	Water column	1	–	1.93	29.77	92.86

Table includes phyla, environment (either water column or sediments), number of representative genomes (*n*), smallest and largest estimated genome sizes (Mbp) observed for each phylum, average GC content (%) and average coding density (%). When only one MAG is indicated, estimated genome size of that MAG is expressed in the sixth column.

two biases with looking at the average estimated genome size of MAGs. One, assembly and binning biases make a MAG an average 3.7% smaller than genomes from isolates in the same genomospecies [2]. Second, a bias of overlooking all those bacteria and archaea that are hard to assemble and bin due to high genomic intrapopulation diversity [44]. For these reasons, in our study, we use two methods that have different biases to answer the same question; how is the genome size of microorganisms distributed in the Baltic Sea. Irrespective of the method used, the average estimated genome size in sediments is larger than in the water column.

The available metadata for benthic and pelagic carbon concentration was not comparable due to methods and metrics used for analysis. This made it not possible to look for a clear ecological link between bigger genome size in benthic zones and nutrient concentration (Table S1). Luckily a previous study has

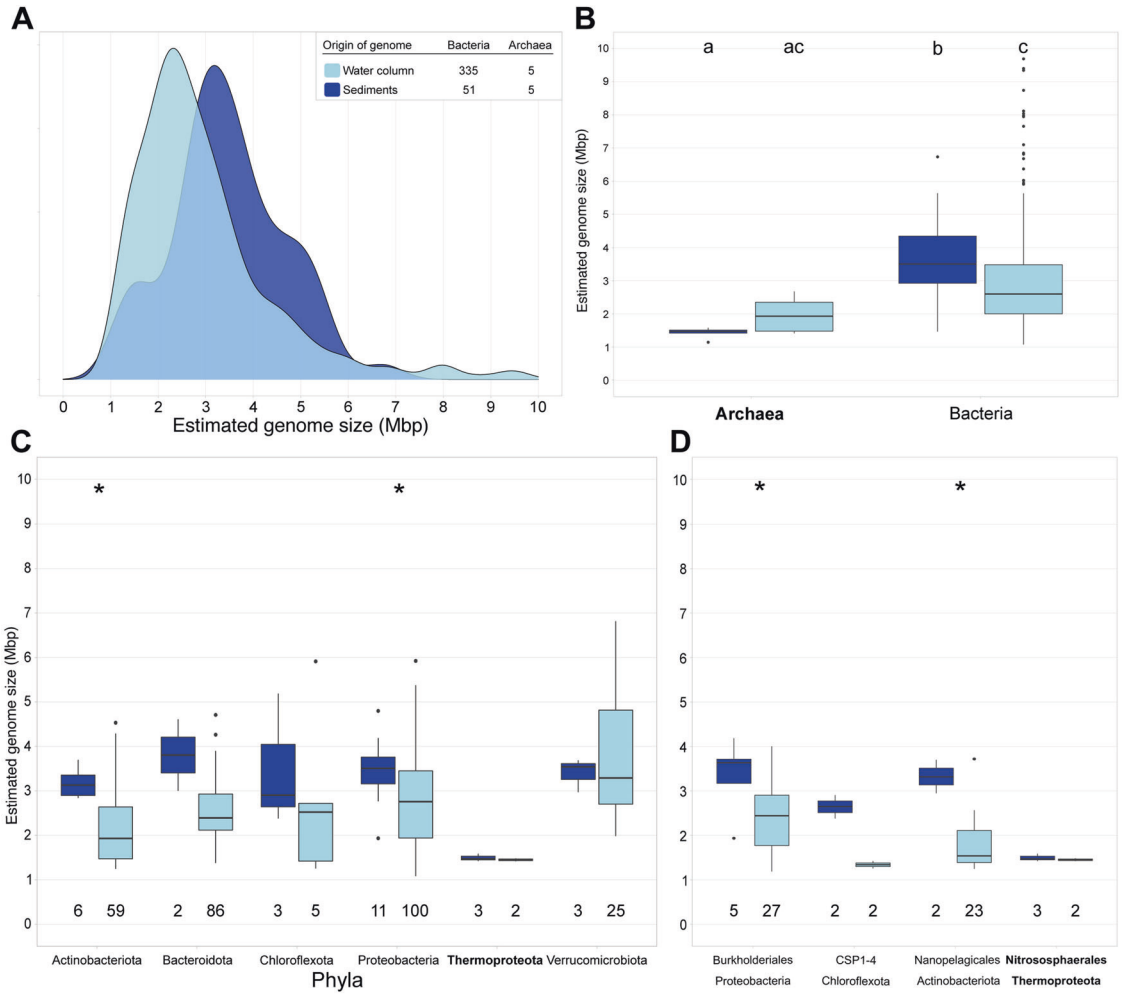

Fig. 3 Overview of the estimated genome size in bacteria and archaea obtained from Baltic Sea sediments (dark blue) and water column (light blue) using only the 397 representative MAGs (95% average nucleotide identity) with >75% completeness. A Shows the genome size distribution of archaea and bacteria obtained from Baltic water column and sediments for a total of 397 representative genomes. **B** Shows the estimated genome size per domain and environment. Different letters indicate significant differences $p < 0.05$ (Kruskal-Wallis non-parametric test; multiple testing corrected with Benjamini-Hochberg). **C** Shows the estimated genome size per phylum. We selected only phyla with at least 2 MAGs in each environment. **D** Shows the estimated genome size per order. We selected only orders with at least 2 MAGs in each environment. Numbers below the boxes indicate the number of MAGs per environment. Stars in Panel C and D indicate significant differences $p < 0.05$ (Wilcoxon non-parametric test).

compiled information on organic carbon stocks in the Baltic Sea [45]: by collecting information from many different studies and years, they show that in average the top 10 cm of sediments contain between 2 and 4 times more organic carbon per area than the water column in the Baltic Sea. More organic carbon available for bacteria and archaea would also pose a lower pressure on the genome to streamline. From the results of our study, we hypothesize that one of the reasons pelagic microbial genomes are smaller than benthic microbial genomes is the difference in organic carbon availability. Altogether, a positive correlation between genome size and nutrient concentration has been shown before [19, 31, 32].

Brackish pelagic microorganisms tend to show smaller genome sizes than marine and freshwater

In further analysis of genome size variation across salinity gradients, we compared the average estimated genome size of pelagic Baltic Sea MAGs to previously published genome size estimations [2]. This comparison includes all taxonomic groups found in three large MAG-datasets [17, 46, 47] (completeness >75%) that includes 4051 freshwater representative MAGs, 2118 marine representative MAGs and 340 pelagic representative MAGs from the Baltic Sea. We found that the average estimated genome size of the brackish pelagic MAGs (2.96 Mbp) is lower than in marine MAGs (3.10 Mbp) (Kruskal-Wallis test, $p < 0.01$). Furthermore, we observed the largest average estimated genome size in freshwater MAGs (3.48 Mbp) (Kruskal-Wallis test, $p < 0.01$) (Fig. 4A). These observed differences in genome sizes across different pelagic environments together with the previously observed differential functions [48] suggests that genome size has a potential signature across aquatic environments.

We then divided all MAGs into phyla and focus on the most common from aquatic environments: phyla Actinobacteriota, Bacteroidota, Cyanobacteria and Proteobacteria (classes Alpha and Gammaproteobacteria) (Fig. 4B–F). We test if the genome size differences are consistent across phyla. We observe that only phylum Bacteroidota follows the same trend as the full dataset in the average estimated genome size (Fig. 4C). On the other hand, MAGs from phylum Actinobacteriota present the biggest genome sizes for marine environments, while no difference on average genome size is observed between freshwater and brackish (Fig. 4B). This result updates previous observations on aquatic

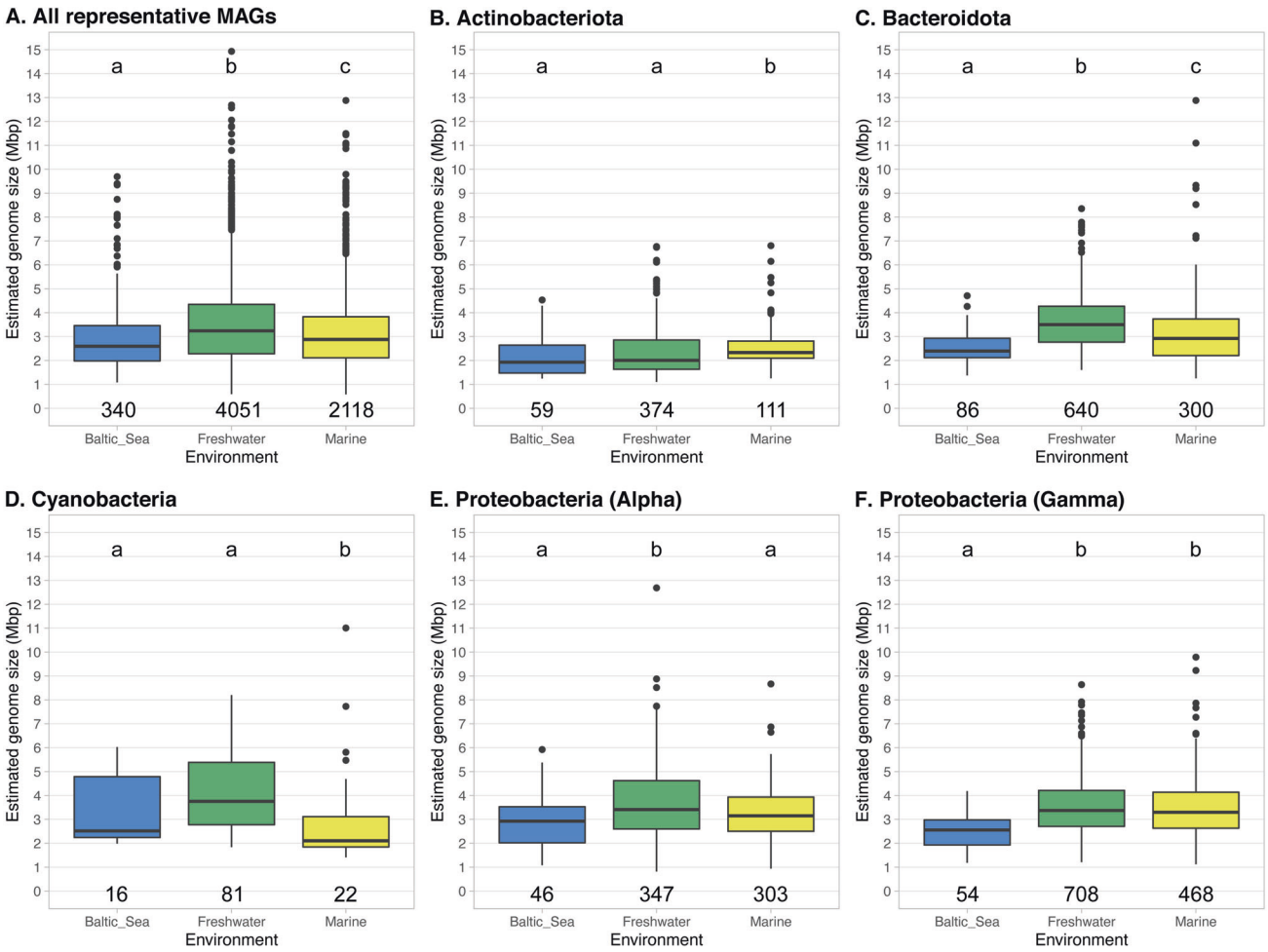

Fig. 4 Overview of the estimated genome size of pelagic Bacteria and Archaea obtained from Baltic Sea (blue), freshwater (green) and marine (yellow) using only representative MAGs (calculates using 95% average nucleotide identity) with >75% completeness. We compare all representative MAGs (**A**), only phylum Actinobacteriota (**B**), phylum Bacteroidota (**C**), phylum Cyanobacteria (**D**), class Alphaproteobacteria (**E**) and class Gammaproteobacteria (**F**). Different letters indicate significant differences $p < 0.05$ (Kruskal-Wallis non-parametric test; multiple testing corrected with Benjamini-Hochberg). Numbers below the boxes indicate the number of MAGs per environment.

Actinobacteriota genome size variation [23]. Opposite to Actinobacteriota, MAGs from phylum Cyanobacteria show the smallest average genome size for marine environments, while we do not observe statistical differences between brackish and freshwater MAGs (Fig. 4D). Similar trends were observed for isolates and MAGs of picocyanobacteria [22, 49]. However, it is important to remark that DNA extraction, assembly, binning and/or quality check of aquatic cyanobacterial MAGs is still a big challenge that needs to be addressed (Supplementary Material 2) [50, 51]. All in all, these results hint at the complicated ecological role of genome size in pelagic bacterial groups, where environment [2, 15, 16], lifestyle [4, 13, 14] and taxonomy [11, 12] are intertwined.

Smaller bacterial genomes in the Baltic Sea tend to lack certain functional categories

We selected the bacterial MAGs with >90% completeness for metabolic annotation and analyze which functional categories correlate with genome size in brackish sediments and water column (Fig. 5A–R and Supplementary Material 3). Functional categories include different but related metabolic pathways (KEGG modules), and each module comprises multiple specific metabolic reactions (module steps) [52]. In our results, we observe patterns of negative correlation between estimated genome size and number of module steps per Mbp in most of the functional categories analyzed (Fig. 5). For example, given the core functions amino acid metabolism, aminoacyl tRNAs and central carbohydrate metabolism, we observe a higher number of module steps per Mbp at lower genome sizes in both environments both in pelagic and benthic MAGs (Fig. 5A, D, J). However, genome size does not explain the number of module steps per Mbp in all functional categories and metabolisms, especially in the case of non-essential functions such as drug resistance and transport systems (Supplementary Material 3F and N). These results suggest that streamlining of genomes select for specific functions and not the whole genome. This could be explained by the Black Queen Hypothesis [53]; when the fitness cost of a function is higher than its benefit, microbes might lose it and, instead, obtain benefit from leaky metabolites from neighboring cells, establishing interdependencies. The loss of those specific functions in the Baltic Sea might have consequential long-lasting metabolic partnerships within the community.

When considering the total number of module steps, we observe that genomes of benthic bacteria code for a higher number of module steps than pelagic bacteria in amino acid metabolism, aminoacyl tRNAs, carbon fixation, nitrogen and sulfur metabolism (Fig. 5B, E, H, N, Q) (Wilcoxon test, $p < 0.01$). This could be the result of a confounding effect of genome size: microorganisms from sediments have bigger genome sizes than

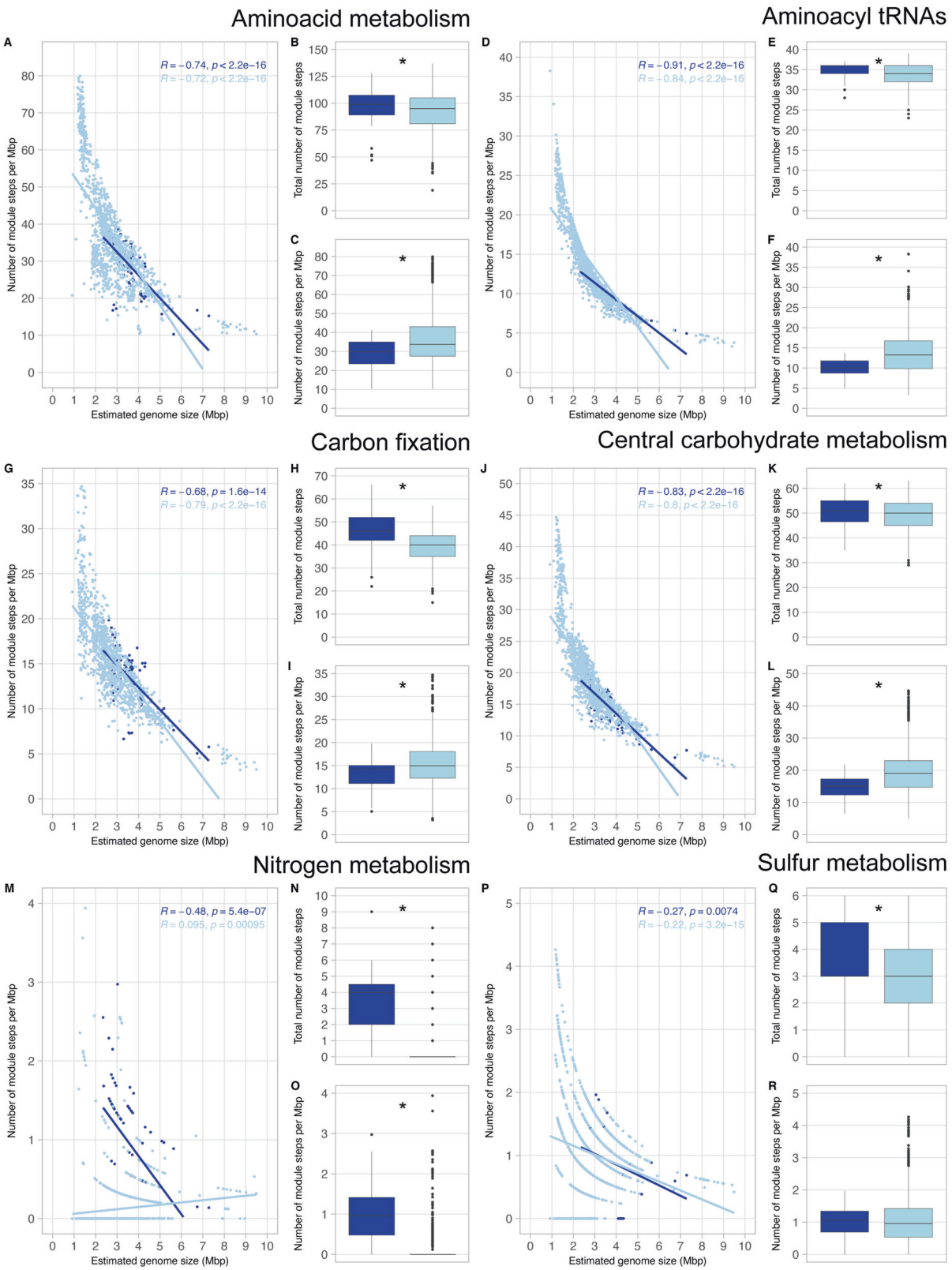

Fig. 5 Overview of the presence of module steps for six metabolic categories. The analyzed metabolic capabilities include amino acid metabolism (**A**–**C**), aminoacyl tRNAs (**D**–**F**), carbon fixation (**G**–**I**), central carbohydrate metabolism (**J**–**L**), nitrogen metabolism (**M**–**O**) and sulfur metabolism (**P**–**R**). We used all bacterial MAGs with very-high quality (>90% completeness and <5% contamination), from both environments (99 MAGs from sediments and 1215 MAGs from water column). Stars in boxplots indicate significant differences $p < 0.05$ (Wilcoxon non-parametric test). **B**, **E**, **H**, **K**, **N** and **Q** indicate total number of module steps. **C**, **F**, **I**, **L**, **O** and **R** indicate the number of module steps per Mbp.

those in the water column (Fig. 2) and therefore, code for a higher number of functions. Hence, the total number of module steps per Mbp in all six categories was analyzed. If streamlining affects all functional categories similarly, coding density trends would be similar for all functional categories. Indeed, we observe that pelagic bacteria present a greater number of module steps per Mbp than sediment bacteria in amino acid metabolism (Fig. 5C), aminoacyl tRNAs (Fig. 5F), carbon fixation (Fig. 5I) and central carbohydrate metabolism (Fig. 5L). However, nitrogen metabolism shows the opposite trend: sediment bacteria have a greater number of module steps per Mbp than water column bacteria (Fig. 5O). Literature report that Baltic sediments contain about 95% of the total pool of nitrogen while the water column only 5%, hence the water column only carries a small part of the overall nitrogen cycle [54]. As mentioned above for carbon, a higher availability of resources, including nitrogen-derived compounds, would also imply a lower evolutionary pressure for streamlined genome sizes in Baltic sediments. These results indicate that benthic bacteria potentially play a bigger role than pelagic bacteria in nitrogen cycling of autotropic systems like the Baltic Sea [55, 56].

Although most of the MAGs presented at least 1 module step related to drug resistance (96.97% sediment MAGs and 98.81% pelagic MAGs), bacteria from sediments presented a higher number of antibiotic resistance module steps than pelagic bacteria. This applies both to total number of module steps and the number of module steps per Mbp (Supplementary Material 3). Our results confirm that aquatic sediments are reservoirs of antibiotic resistance genes [57, 58]. Just as sediments harbor more than double the organic carbon than the water column, this allows microbial genomes to have a bigger size and code for a higher number of genes. This allows microbes to upkeep non-essential functions and allow metabolic reservoir in the sediments.

Baltic sediments and water column harbor bacteria with different metabolic capabilities

From all >90% completeness MAGs, we selected only bacterial phyla with five or more high-quality MAGs to observe how metabolism differs between different taxa in Baltic sediments and water column (Fig. 6).

We observed the presence of the genes *cdhH* | *cdhE* | *cooS* from the Wood-Ljungdhal pathway exclusively in marine sediments (Fig. 6), particularly in phyla Desulfobacterota, Desulfobacterota E and Verrucomicrobiota (class Kiritimatiellae). This is a carbon fixation pathway predominant in acetogenic bacteria found in anoxic conditions [59]. Complementarily, we also find putative fermentation genes for acetogenesis (*acdA* | *ack* | *pta*) to be widespread across taxa in Baltic sediments. This would explain the potential success of acetogenic metabolism in brackish sediments [60]. We also find that the *acs* gene for acetate fermentation into acetyl-CoA is widely distributed in both sediments and water column. These results support the common distribution of acetogens and the Wood-Ljungdhal pathway in Baltic glacial sediments [61].

No FeFe hydrogenases were detected, but different NiFe hydrogenases were spotted to differ between environments: NiFe groups 3abd were detected mainly in pelagic bacteria, while NiFe group 1 in sediments (Fig. 6). NiFe group 1 hydrogenases could be playing a vital role in nitrate (NO_3^-), sulfate (SO_4^{2-}) and iron (Fe^{3+}) reduction: these molecules can act as acceptors of electrons coupled to H_2 oxidation in anoxic conditions [62]. Moreover, putative genes coding for the reduction of the above-mentioned molecules were also detected on our sediment dataset (*napAB* | *narGH* for nitrate reduction, *aprA* | *sat* for sulfate reduction, and iron reduction series genes). These results suggest a key role of sediment bacteria in sulfur, nitrogen, and iron cycling in the Baltic Sea. For example, sulfate reducers such as Desulfobacterota found in our benthic MAGs collection, most likely contribute to the release of Fe-bound phosphorus from sediments to the water column [63].

Conclusions

In this research article, we provide a comprehensive analysis to investigate the ecological implications of microbial genome in the Baltic Sea. We show that genome size in Bacteria and Archaea is linked to the environment (Figs. 2, 3, 4 and S1 and Supplementary Material 1), taxonomic identity (Table 1 and Fig. 3) and metabolic potential (Fig. 5 and Supplementary Material 3). We also provide some insights on how distinct pelagic and benthic microbial communities in the Baltic Sea are: not only microbial MAGs retrieved from these two environments are different in taxonomy (Table 1), but also in genome size (Fig. 3) and metabolism (Figs. 5 and 6). This highlights water bodies are highly heterogeneous biomes, with highly distinct microbial communities between micro-niches. With the continuous progression of aquatic microbial ecology and the development of new isolation, omics and bioinformatic techniques, future research should provide a more complete and unbiased view of genome sizes distribution in nature and its ecological implication in microbial populations.

MATERIAL AND METHODS

Baltic sea metagenomes collection

For this study we compiled new and public Baltic Sea metagenomes from the water column and the sediments. The final dataset consisted of 219 metagenome samples from a wide range of locations in the Baltic Sea that include 5 independent datasets (Table S1 and Fig. 1). We compiled 108 sediment metagenomes, of which 59 were collected in 2019 and recently published [29], and we collected 49 metagenomes from 2016 to 2018. The pelagic dataset consists of 118 pelagic samples collected from 2011 to 2015 published in three different studies [46, 64, 65]. All five datasets have abiotic metadata of depth (m), salinity (PSU), temperature (C) and oxygen concentration (mg/L) (Table S1).

Environmental sampling

The top 2 cm of sediment was collected at soft bottom clay-muddy habitats from 59 stations from north to south in the Baltic Sea in 2019, following the sampling described in Broman et al., (2022) (Table S1 for coordinates). Briefly, one sediment core was collected per station using a Kajak gravity corer (surface area: 50 cm^2, one core per station) and the top 0–2 cm layer was sliced into a 215 ml polypropylene container (207.0215PP, Noax Lab, Sweden). The sediment was homogenized and stored at −20 °C on the boat, kept on an iced cooler without thawing for ~2 h during transportation to the university, and finally stored again at −20 °C until DNA extraction. Bottom water (~20 cm above the sediment surface) was collected at each station with a Niskin bottle. This was followed by on deck measurements of bottom water temperature, salinity, and dissolved O2 using a portable multimeter (HQ40D, Hach).

DNA extraction and sequencing

The sediment samples were thawed, homogenized, and a subsample of 0.25 g was used for DNA extraction using the DNeasy PowerSoil kit (Qiagen) according to the manufacturer's protocol. The quantity and quality of eluted DNA were measured using NanoDrop One spectrophotometer and Qubit 2 (both by ThermoFisher Scientific) to ensure that samples meet the minimum requirements for sequencing. The samples were then sequenced at the Science for Life Laboratories facility on one NovaSeq 6000 S4 lanes using a 2 × 150 bp setup. Sequencing yielded on average 53.0 million reads per sample.

MAGs collection

Assembling and binning of the 108 sediment metagenomes resulted in 2248 bins. To obtain bins from metagenomes, we followed the 0053_metasssnake2 pipeline (https://github.com/moritzbuck/0053_metasssnake2) (v0.0.2). In this pipeline we used Sourmash [66] to create signatures, Megahit [67] to obtain single-sample assemblies and Metabat2 [68] for the binning of the assemblies. We used default parameters throughout the whole pipeline. Quality of the bins was assessed using CheckM (v1.1.3) [69]: we used the typical workflow (*lineage_wf*) with default parameters, and selected only bins

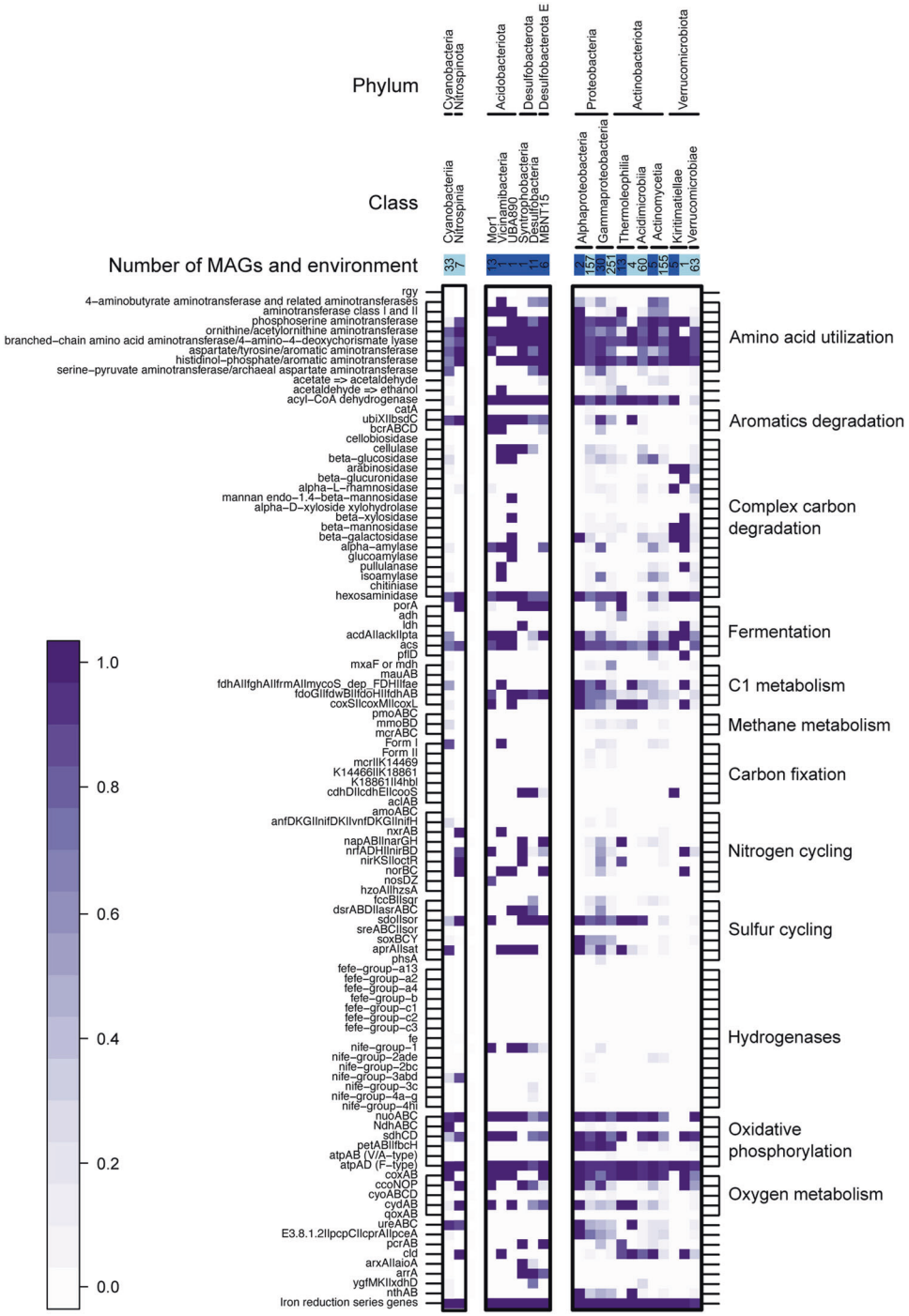

Fig. 6 Metabolic potential of all high-quality MAGs (>90% completeness and <5% contamination). We selected all phyla with at least 5 MAGs, and then divided by class. Boxes on the top of the figure indicate environment (pelagic in light blue and sediments in dark blue), and inner numbers indicate the number of MAGs per category. In the heatmap, white squares indicate absence of a given gene in all MAGs, and the darkest purple indicates presence in all of them (gradient scale on the bottom-left part of the figure for reference).

with quality of completeness >75% and contamination <5%. From all the bins, only 216 passed our quality threshold and we named those MAGs (metagenome-assembled genomes). All MAGs were taxonomically classified using GTDB-tk (v1.5.0) [70] according to the GTDB classification (data version r202) [71]. The quality of the MAGs belonging to phyla Actinobacteriota and Patescibacteria were assessed separately using a custom set of marker genes. Preliminary quality check of Actinobacteriota genomes in a publicly available freshwater dataset [47] show that default parameters underestimate the quality of the MAGs that are classified as Actinobacteriota compared to using a custom marker gene set (Supplementary Material 2). In the case of phylum Patescibacteria, we used a custom set of maker genes provided by CheckM [69, 72].

Complementarily, we used 1920 pelagic MAGs that were published [46] and passed the >75% completeness and <5% contamination threshold. All high-quality MAGs from the sediments and water column were dereplicated using fastANI (95% ANI threshold as estimator of genetic unit) and mOTUlizer (v0.3.2) [73, 74]. From the 216 with >75% completeness MAGs, 56 were chosen as representatives. From the 1920 pelagic MAGs, 340 were chosen as representatives. All genomic information for pelagic and benthic MAGs is included in Table S2.

Genome size analysis

We studied genome size in two different levels: entire microbial community and bacterial/archaeal MAGs. To study differences in genome size between Baltic sediments and water column at the community level, we first calculated the average genome size (AGS) of the metagenomes using MicrobeCensus (v1.1.0) [75]. MicrobeCensus estimates the AGS of a microbial community from metagenomes by aligning reads to a set of single-copy genes that are widely distributed across taxa to calculate their abundance, with highly accurate estimations [41]. We excluded one of the pelagic datasets [65] due to the presence of spike-in DNA. We used default parameters, but we set the number of reads sampled to 10 million (*-n 10 000 000*). To estimate the genome size of the microbial MAGs, we divided the MAGs assembly size by the completeness (provided by CheckM, ranging from 0 to 1). To study genome size variation between our pelagic brackish dataset and other major aquatic environments (freshwater and marine), we compiled the metagenomic information from all pelagic MAGs from marine and freshwater environments (>75% completeness and <5% contamination) from [2] (Table S3).

Metabolic annotation

To analyze the metabolic potential of sedimentary and pelagic bacteria, we selected all MAGs with completeness >90% and contamination <5%. In total, we obtained 99 MAGs from sediments and 1241 MAGs from the water column. The metabolic potential of sediment and pelagic MAGs was reconstructed using Prodigal annotations (v2.6.3) [76]. We used the resulting protein translation files to predict biogeochemical and metabolic functional traits using METABOLIC (v4.0) [52]. We used the METABOLIC-G script, using default settings.

Statistical analysis

We performed Wilcoxon non-parametric test to analyze if there were significant differences between pairs of boxplots (Figs. 2, 3 and 5 and Supplementary Material 3). Asterisks in boxplots indicate significant differences $p < 0.05$. In Figs. 3B and 4 we performed Kruskal-Wallis test corrected with Benjamini-Hochberg, to test statistical differences between groups. Different letters are the result of this non-parametric test; $p < 0.05$. We performed a ANOVA type II analysis to test the effect of abiotic factors and their interactions on the AGS, using the function *aov* from the R package *stats* v3.6.3 [77] (Supplementary Material 1). We obtained the correlation coefficients on scatterplots (Fig. 5 and S1) to test the fit of our data to linear regressions using the function *stat_cor* from the R package *ggpubr* v0.4.0 [78].

DATA AVAILABILITY

All metagenome datasets are available in public repositories under NCBI project accession number SRP077551 and ENA accession numbers PRJEB34883, PRJEB22997 and PRJEB41834. Specific accession numbers of all metagenomes are available in Table S1. Pelagic MAGs can be found on project PRJEB34883 and benthic MAGs on project PRJNA891615. Assembly and binning of the dataset provided in this paper used scripts available at https://github.com/moritzbuck/0053_metasssnake2. Supplemental material can also be accessed 10.17044/scilifelab.21378294.

REFERENCES

1. Kirchberger PC, Schmidt ML, Ochman H. The ingenuity of bacterial genomes. Annu Rev Microbiol. 2020;74:815–34.
2. Rodríguez-Gijón A, Nuy JK, Mehrshad M, Buck M, Schulz F, Woyke T, et al. A genomic perspective across Earth's microbiomes reveals that genome size in Archaea and Bacteria is linked to ecosystem type and trophic strategy. Front Microbiol. 2022;12:761869.
3. Lynch M. Streamlining and simplification of microbial genome architecture. Annu Rev Microbiol. 2006;60:327–49.
4. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. ISME J. 2014;8:1553–65.
5. Kuo CH, Moran NA, Ochman H. The consequences of genetic drift for bacterial genome complexity. Genome Res. 2009;19:1450–4.
6. Wolf YI, Koonin EV. Genome reduction as the dominant mode of evolution. BioEssays. 2013;35:829–37.
7. Batut B, Knibbe C, Marais G, Daubin V. Reductive genome evolution at both ends of the bacterial population size spectrum. Nat Rev Microbiol. 2014;12:841–50.
8. Bobay LM, Ochman H. Factors driving effective population size and pan-genome evolution in bacteria. BMC Evol Biol. 2018;18:153.
9. Konstantinidis KT, Tiedje JM. Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci. 2004;101:3160–5.
10. Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. BMC Biology. 2014;12:66.
11. Biller SJ, Berube PM, Lindell D, Chisholm SW. Prochlorococcus: the structure and function of collective diversity. Nat Rev Microbiol. 2015;13:13–27.
12. Martinez-Gutierrez CA, Aylward FO. Genome size distributions in bacteria and archaea are strongly linked to evolutionary history at broad phylogenetic scales. PLoS Genet. 2022;18:e1010220.
13. Moran NA, Mira A. The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. Genome Biol. 2001;2:research0054.1.
14. van Ham RCHJ, Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, et al. Reductive genome evolution in Buchnera aphidicola. Proc Natl Acad Sci USA. 2003;100:581–6.
15. Maistrenko OM, Mende DR, Luetge M, Hildebrand F, Schmidt TSB, Li SS, et al. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. ISME J. 2020;14:1247–59.
16. Simonsen AK. Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria. ISME J. 2022;16:423–34.
17. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. Nat Biotechnol. 2021;39:499–509.
18. Konstantinidis KT, Braff J, Karl DM, DeLong EF. Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. Appl Environ Microbiol. 2009;75:5345–55.
19. Mende DR, Bryant JA, Aylward FO, Eppley JM, Nielsen T, Karl DM, et al. Environmental drivers of a microbial genomic transition zone in the ocean's interior. Nat Microbiol. 2017;2:1367–73.
20. Salcher MM, Schaefle D, Kaspar M, Neuenschwander SM, Ghai R. Evolution in action: habitat transition from sediment to the pelagial leads to genome streamlining in Methylophilaceae. ISME J. 2019;13:2764–77.
21. Chen MY, Teng WK, Zhao L, Hu CX, Zhou YK, Han BP, et al. Comparative genomics reveals insights into cyanobacterial evolution and habitat adaptation. ISME J. 2021;15:211–27.
22. Cabello-Yeves PJ, Callieri C, Picazo A, Schallenberg L, Huber P, Roda-Garcia JJ, et al. Elucidating the picocyanobacteria salinity divide through ecogenomics of new freshwater isolates. BMC Biol. 2022;20:175.
23. Hugerth LW, Larsson J, Alneberg J, Lindh MV, Legrand C, Pinhassi J, et al. Metagenome-assembled genomes uncover a global brackish microbiome. Genome Biol. 2015;16:279.
24. Coelho LP, Alves R, del Río ÁR, Myers PN, Cantalapiedra CP, Giner-Lamia J, et al. Towards the biogeography of prokaryotic genes. Nature. 2022;601:252–6.
25. Zhou Z, Tran PQ, Kieft K, Anantharaman K. Genome diversification in globally distributed novel marine Proteobacteria is linked to environmental adaptation. ISME J. 2020;14:2060–77.
26. Acinas SG, Sánchez P, Salazar G, Cornejo-Castillo FM, Sebastián M, Logares R, et al. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. Commun Biol. 2021;4:604.
27. Louca S, Parfrey LW, Doebeli M. Decoupling function and taxonomy in the global ocean microbiome. Science. 2016;353:1272–7.
28. Li M, Wei G, Shi W, Sun Z, Li H, Wang X, et al. Distinct distribution patterns of ammonia-oxidizing archaea and bacteria in sediment and water column of the Yellow River estuary. Sci Rep. 2018;8:1584.
29. Broman E, Izabel-Shen D, Rodríguez-Gijón A, Bonaglia S, Garcia SL, Nascimento FJA. Microbial functional genes are driven by gradients in sediment stoichiometry, oxygen, and salinity across the Baltic benthic ecosystem. Microbiome. 2022;10:126.
30. Nielsen DA, Fierer N, Geoghegan JL, Gillings MR, Gumerov V, Madin JS, et al. Aerobic bacteria and archaea tend to have larger and more versatile genomes. Oikos. 2021;130:501–11.
31. Allen LZ, Allen EE, Badger JH, McCrow JP, Paulsen IT, Elbourne LD, et al. Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic. ISME J. 2012;6:1403–14.
32. Aylward FO, Santoro AE. Heterotrophic thaumarchaea with small genomes are widespread in the Dark Ocean. mSystems. 2020;5:e00415–20. Jun 16/msystems/5/3/msys.00415-20.atom
33. Conley DJ, Björck S, Bonsdorff E, Carstensen J, Destouni G, Gustafsson BG, et al. Hypoxia-related processes in the baltic sea. Environ Sci Technol. 2009;43:3412–20.
34. Herlemann DP, Labrenz M, Jürgens K, Bertilsson S, Waniek JJ, Andersson AF. Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. ISME J. 2011;5:1571–9.
35. Thureborn P, Lundin D, Plathan J, Poole AM, Sjöberg BM, Sjöling S. A metagenomics transect into the deepest point of the baltic sea reveals clear stratification of microbial functional capacities. Gilbert JA, editor. PLoS ONE. 2013;8:e74983.

36. Seeger C, Dyrhage K, Mahajan M, Odelgard A, Lind SB, Andersson SGE. The subcellular proteome of a planctomycetes bacterium shows that newly evolved proteins have distinct fractionation patterns. Front Microbiol. 2021;12:643045.
37. Giovannoni SJ. SAR11 bacteria: the most abundant plankton in the oceans. Annu Rev Mar Sci. 2017;9:231–55.
38. Lanclos VC, Rasmussen AN, Kojima CY, Cheng C, Henson MW, Faircloth BC, et al. Ecophysiology and genomics of the brackish water adapted SAR11 subclade IIIa. ISME J. 2023;17:620–9.
39. Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, Gonzalez JM, et al. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. Proc Natl Acad Sci. 2013;110:11463–8.
40. Giovannoni SJ. Genome streamlining in a cosmopolitan oceanic bacterium. Science. 2005;309:1242–5.
41. Pereira-Flores E, Glöckner FO, Fernandez-Guerra A. Fast and accurate average genome size and 16S rRNA gene average copy number computation in metagenomic data. BMC Bioinformatics. 2019;20:453.
42. Kristensen DM, Mushegian AR, Dolja VV, Koonin EV. New dimensions of the virus world discovered through metagenomics. Trends Microbiol. 2010;18:11–9.
43. Lind AL, Pollard KS. Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing. Microbiome. 2021;9:58.
44. Ghurye JS. Metagenomic assembly: overview, challenges and applications. Yale J Biol Med. 2016;89:353–62.
45. Scheffold MIE, Hense I. Quantifying contemporary organic carbon stocks of the baltic sea ecosystem. Front Mar Sci. 2020;7:571956.
46. Alneberg J, Bennke C, Beier S, Bunse C, Quince C, Ininbergs K, et al. Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. Commun Biol. 2020;3:119.
47. Buck M, Garcia SL, Fernandez L, Martin G, Martinez-Rodriguez GA, Saarenheimo J, et al. Comprehensive dataset of shotgun metagenomes from oxygen stratified freshwater lakes and ponds. Sci Data. 2021;8:131.
48. Jurdzinski KT, Mehrshad M, Delgado LF, Deng Z, Bertilsson S, Andersson AF Large-scale phylogenomics of aquatic bacteria reveal molecular mechanisms for adaptation to salinity. Microbiology; 2022 [cited 2023 Jan 17].
49. Sánchez-Baracaldo P, Bianchini G, Di Cesare A, Callieri C, Chrismas NAM. Insights into the evolution of picocyanobacteria and phycoerythrin genes (mpeBA and cpeBA). Front Microbiol. 2019;10:45.
50. Alvarenga DO, Fiore MF, Varani AM. A metagenomic approach to cyanobacterial genomics. Front Microbiol. 2017;8:809.
51. Kim Tiam S, Comte K, Dalle C, Duval C, Pancrace C, Gugger M, et al. Development of a new extraction method based on high-intensity ultra-sonication to study RNA regulation of the filamentous cyanobacteria Planktothrix. PLoS ONE. 2019;14:e0222029.
52. Zhou Z, Tran PQ, Breister AM, Liu Y, Kieft K, Cowley ES, et al. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. Microbiome. 2022;10:33.
53. Morris JJ, Lenski RE, Zinser ER. The black queen hypothesis: evolution of dependencies through adaptive gene loss. mBio. 2012;3:e00036–12.
54. Lønborg C, Markager S. Nitrogen in the Baltic Sea: Long-term trends, a budget and decadal time lags in responses to declining inputs. Estuarine, Coastal Shelf Sci. 2021;261:107529.
55. Albert S, Bonaglia S, Stjärnkvist N, Winder M, Thamdrup B, Nascimento FJA. Influence of settling organic matter quantity and quality on benthic nitrogen cycling. Limnol Oceanogr. 2021;66:1882–95.
56. Griffiths JR, Kadin M, Nascimento FJA, Tamelander T, Törnroos A, Bonaglia S, et al. The importance of benthic-pelagic coupling for marine ecosystem functioning in a changing world. Glob Change Biol. 2017;23:2179–96.
57. Guo XP, Zhao S, Chen YR, Yang J, Hou LJ, Liu M, et al. Antibiotic resistance genes in sediments of the Yangtze Estuary: From 2007 to 2019. Science of The Total Environment. 2020;744:140713.
58. Marti E, Variatza E, Balcazar JL. The role of aquatic ecosystems as reservoirs of antibiotic resistance. Trends Microbiol. 2014;22:36–41.
59. Esposito A, Tamburini S, Triboli L, Ambrosino L, Chiusano ML, Jousson O. Insights into the genome structure of four acetogenic bacteria with specific reference to the Wood–Ljungdahl pathway. MicrobiologyOpen. 2019;8:e938. Dec [cited 2022 Aug 26]
60. Lever MA. Acetogenesis in the energy-starved deep biosphere – a paradox? Front Microbio. 2012;2:284. Jan [cited 2022 Aug 26]
61. Marshall IPG, Karst SM, Nielsen PH, Jørgensen BB. Metagenomes from deep Baltic Sea sediments reveal how past and present environmental conditions determine microbial community composition. Marine Genomics. 2018;37:58–68.
62. Peters JW, Schut GJ, Boyd ES, Mulder DW, Shepard EM, Broderick JB, et al. [FeFe]- and [NiFe]-hydrogenase diversity, mechanism, and maturation. Biochimica et Biophysica Acta (BBA)—Mol Cell Res. 2015;1853:1350–69.
63. Sinkko H Sediment bacterial communities in nutrient cycling and in the history of the Baltic Sea [Doctoral dissertation]. [Helsinki, Finland]: University of Helsinki; 2013.
64. Larsson J, Celepli N, Ininbergs K, Dupont CL, Yooseph S, Bergman B, et al. Picocyanobacteria containing a novel pigment gene cluster dominate the brackish water Baltic Sea. ISME J. 2014;8:1892–903.
65. Alneberg J, Sundh J, Bennke C, Beier S, Lundin D, Hugerth LW, et al. BARM and BalticMicrobeDB, a reference metagenome and interface to meta-omic data for the Baltic Sea. Sci Data. 2018;5:180146.
66. Titus Brown C, Irber L. sourmash: a library for MinHash sketching of DNA. JOSS. 2016;1:27.
67. Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, et al. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods. 2016;102:3–11.
68. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7:e7359.
69. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
70. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2020;36:1925–7.
71. Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. 2020;38:1079–86.
72. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature. 2015;523:208–11.
73. Buck M, Mehrshad M, Bertilsson S. mOTUpan: a robust Bayesian approach to leverage metagenome-assembled genomes for core-genome estimation. NAR Genomics Bioinformatics. 2022;4:lqac060.
74. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:5114.
75. Nayfach S, Pollard KS. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. Genome Biol. 2015;16:51.
76. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
77. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2020. Available from: https://www.R-project.org/.
78. Kassambara, A ggpubr: 'ggplot2' Based Publication Ready Plots. 2020. Available from: https://CRAN.R-project.org/package=ggpubr

ACKNOWLEDGEMENTS

We thank Elias Broman for his help on the acquisition of the benthic dataset of metagenomes and metadata and Ola Svensson, Caroline Raymond and Jonas Gunnarsson for assistance during sampling. We also thank Juanita Gutiérrez-Valencia for her suggestions in statistical analysis.

This work was supported by SciLifeLab. The authors acknowledge support from the National Genomics Infrastructure in Stockholm funded by Science for Life Laboratory, the Knut and Alice Wallenberg Foundation and the Swedish Research Council, and SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure. Computational work and data handling were enabled by resources in the projects SNIC 2020/5-159, 2021/5-133, 2022/5-137, 2020-6-60 and 2022/6-77 provided by the Swedish National Infrastructure for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

AUTHOR CONTRIBUTIONS

AR-G and SLG conceptualized and designed research project. AR-G, MB and SLG refined the project idea. AR-G, AFA and FJAN participated in data collection. AR-G and DIS did the molecular work. AR-G and MB performed bioinformatic analysis to obtain MAGs from sediment raw sequences. AR-G and SLG performed data analysis. AR-G and SLG drafted the first manuscript. AR-G, DIS and SLG did literature searches. All authors contributed to the writing and editing of the manuscript.

FUNDING

Open access funding provided by Stockholm University.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43705-023-00231-x.

Correspondence and requests for materials should be addressed to Alejandro Rodríguez-Gijón or Sarahi L. Garcia.