Genome size variation between pelagic and benthic communities across prokaryotic taxonomy and environmental gradients in the Baltic Sea

RUNNING TITLE: Genome size in benthic and pelagic Baltic Sea

AUTHORS

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

$ Correspondent authors

alejandro.rgijon@gmail.com

sarahi.garcia@su.se

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

Keywords: microbial ecology, genome size, bacteria, Baltic Sea, streamlining theory

ABSTRACT (250 words)

Microbial genome size can be used as a predictor to explain the ecology and metabolism of Bacteria and Archaea across major biomes. Despite their ecological significance, the contribution of microbial genome size to differences in metabolic potential of benthic and pelagic prokaryotes are poorly studied. Here, we investigated how taxonomy and microbial genome size varies between benthic and pelagic habitats across environmental gradients of the brackish Baltic Sea. We also explored the relationships between that variation, the environmental heterogeneity, and microbial functions in these habitats. By analyzing Baltic metagenomes and MAGs, we observe that pelagic brackish Bacteria and Archaea present smaller genome sizes on average than pelagic marine and freshwater prokaryotes. Moreover, we found that prokaryotic genome sizes in Baltic sediments (3.47 Mbp) are significantly bigger than in the water column (2.96 Mbp). These differences in genome size persisted in Bacteria from the phyla level to the order level. For pelagic prokaryotes, the smallest genomes coded for a higher number of module steps per Mbp than bigger genomes for most of the functions, such as 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 presence of different hydrogenases.

IMPORTANCE (150 words)

In Bacteria and Archaea, genome size is the result of strong selection pressures shaping the ecology and metabolism of microbial lineages. Thanks to the development of metagenomic analysis of environmental samples in the last three decades, we can now survey the size of their genomes, quantify their functional capabilities, and infer the role they play in the environment. Here, by using four publicly available datasets together with one new unpublished dataset, we provide new insights on how genome size, metabolism and taxonomy are linked and how they differ between sediments and water column in the Baltic Sea. Moreover, we investigate differences in genome size of pelagic microorganisms across ecosystems, from the brackish Baltic Sea to marine and freshwater Bacteria and Archaea. Our results provide a comprehensive and contrasting view between the taxonomic and genetic

diversity of sediments and water column and its implications on nutrient cycling in the Baltic Sea.

INTRODUCTION

Genome size in Bacteria and Archaea can be used as a predictor of microbial metabolic complexity and niche occupation across Earth's biomes (Lynch, 2006; Maistrenko et al., 2020; Rodríguez-Gijón et al., 2022; Swan et al., 2013). Over the last two decades, different explanations have been postulated to find the mechanisms responsible for the variations in prokaryotic genome length. Reported drivers of genome size are phylogenetic history (Biller et al., 2015; Martinez-Gutierrez and Aylward, 2022), lifestyle such as free-living, particle attached or host-associated (Giovannoni et al., 2014; Moran and Mira, 2001; van Ham et al., 2003), or environment such as marine, freshwater, different types of sediments or different hosts in host-associated microorganisms (Rodríguez-Gijón et al., 2022; Simonsen, 2022). While many factors have been found to partially explain genome size, ultimately, this implies that multiple evolutionary forces and their interactions shape genome size in prokaryotes. Untangling how and why different lineages have dissimilar genome sizes across different ecosystems can be key to understand the ecology and evolutionary history of Bacteria and Archaea.

Aquatic microorganisms have been extensively sampled and have a large representation in genomic and metagenomic datasets (Nayfach et al., 2021). Their genome size spans from ~0.5 to ~15, averaging 3.1 million base pairs (Mbp) (Rodríguez-Gijón et al., 2022). Aquatic environments are heterogeneous and many different abiotic factors, such as salinity and depth, could be responsible for the microbial genome size variation. Within marine environments, it has been postulated that pelagic microbes inhabiting deep marine environments present bigger genome sizes than those in shallow marine waters (Konstantinidis et al., 2009; Mende et al., 2017). Within freshwater ecosystems, we can observe that isolates from the family Methylophilaceae (class Gammaproteobacteria) show a decrease in genome size in the transition from sediments to pelagic forms (Salcher et al., 2019). Additionally, several studies have shown that marine microbes have smaller genome sizes than freshwater microorganisms (Cabello-Yeves et al., 2022; Chen et al., 2021; Rodríguez-Gijón et al., 2022). Yet, genome size variation in the brackish realm remains debated: Actinobacteria in the brackish Baltic Sea show bigger genome sizes than in

freshwaters (Hugerth et al., 2015), while picocyanobacteria show the opposite trend (Cabello-Yeves et al., 2022). Further research must be done to understand how genome sizes vary between and within different aquatic environments, particularly brackish water bodies including pelagic and benthic microorganisms.

In pelagic Bacteria, many abundant and ubiquitous lineages present reduced genomes (Hugerth et al., 2015). The Streamlining Theory has been proposed to explain this tendency towards simplification in aquatic environments (Giovannoni et al., 2014). This evolutionary theory indicates that abundant and ubiquitous aquatic microorganisms exhibit small cells and reduced genomes, as observed for *Pelagibacter ubique* (class Alphaproteobacteria) (Giovannoni, 2005; Grote et al., 2012) or *Prochlorococcus* spp. (phylum Cyanobacteria) (Biller et al., 2015). In many cases, reduced genomes imply the absence of key metabolic capabilities, such as the electron transport chains in Patescibacteria (Beam et al., 2020) or the catalase-peroxidase in *Prochlorococcus* (Scanlan et al., 2009). The absence of essential metabolic capabilities is explained by the Black Queen Hypothesis (Morris et al., 2012): when the fitness cost of a function is higher than its benefit, microbes might stop performing it and, instead, obtain benefit from leaky metabolites from neighboring cells, establishing strong interdependencies. As a whole, genetic streamlining has significant consequences on genome size, metabolism, adaptability, and lifestyle in microorganisms.

Environmental conditions strongly select which metabolic functions are present or absent on microbial communities and, consequently, the length of the genome. Metagenomic 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 (Coelho et al., 2022). This remarks how relevant is the relationship between niche specificity and lineage specific functional traits to understand microbial ecology and evolution (Zhou et al., 2020). These functional capabilities also differ between water column and sediments in Bacteria and Archaea in both marine (Acinas et al., 2021; Louca et al., 2016) and freshwater environments (Li et al., 2018). As a whole, gene repertoires and genome size are interrelated, and researchers must consider them together with environmental gradients to better understand niche-specificity and the ecology of different prokaryotic lineages.

In this research article, our main aim is to link the genome size of Bacteria and Archaea with their ecology and functional potential across the environmental gradients in the

Baltic Sea in both pelagic and benthic habitats. Specifically, we investigate; i) how genome size varies among brackish communities 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 three pelagic and one benthic publicly available metagenomic datasets and we provide one new benthic dataset to study genome size distribution across sediments and water column in the Baltic Sea (**Figure 1**). 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 inhabiting brackish sediments present bigger genome sizes than those 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 some functions, such as nitrogen metabolism, are mainly found in benthic Bacteria but are almost absent in pelagic bacteria the Baltic Sea.

RESULTS AND DISCUSSION

Benthic metagenomes and MAGs show bigger estimated average genome sizes than pelagic in the Baltic Sea. First, we calculated the average genome size (AGS) of metagenomes across the latitudinal gradient of the Baltic Sea using two sediment and two pelagic datasets. We observed that sediment-dwelling microbial communities present significantly larger average genome sizes (mean AGS = 6.01 Mbp, n = 108) than pelagic communities (mean AGS = 5.40 Mbp, n = 69) (t-test, $p < 0.01$) (**Figure 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 **Figure S1**). Previous marine analysis have found the opposite effect, bigger genome sizes in deeper areas (Mende et al., 2017). 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 average genome size in the sediment samples (**Figure 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 Death Zone) and those metagenomes with oxygen concentration 2 to 12.45 mg/L. We observe that benthic metagenomes from lower oxygen concentration (mean AGS = 7.08 Mbp, n = 14) present significantly higher genome sizes than those in sediments with oxygen concentration higher than 2 mg/L (mean AGS = 5.85 Mbp, n = 94) (t-test, $p < 0.01$) (**Figure 2B**). Complementarily, previous results show that the Dead Zone bacteria tend to be more metabolically similar to each other when compared to bacteria from oxic sediments (Broman et al., 2022). Our observation confirms a recent study that observed microaerophilic microorganisms from diverse environments present smaller genome sizes than obligate anaerobic (Nielsen et al., 2021). Moreover, previous studies show a positive correlation between genome size and nutrient concentration (Allen et al., 2012; Aylward and Santoro, 2020; Mende et al., 2017). Dead Zones in the Baltic Sea are characterized by anthropogenic eutrophication (Conley et al., 2009), which would also promote bigger genome sizes. Altogether, these results indicate that high nutrient concentration and low oxygen concentrations in Baltic Sea Dead Zones promote bigger genome sizes with similar metabolic capabilities.

To compare genome size between pelagic and benthic Bacteria and Archaea, we obtained 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 MAGs (95% average nucleotide identity). Additionally, 1920 pelagic MAGs were dereplicated into 340 representative MAGs. 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 species-level representative MAG was binned from both sediments and water column, belonging to genus *Mycobacterium* (phylum

Actinobacteriota). These differences on taxonomical composition in microbial communities between water column and sediments altogether with latitudinal changes in microbial biodiversity in the Baltic Sea (Herlemann et al., 2011) show how heterogeneous brackish environments are.

Additionally, the eight largest representative genomes of the dataset belonged to phylum Planctomycetota (family Planctomycetaceae) (7.95 – 9.69 Mbp) belonging to the water column and not to sediments. It has been previously observed that Planctomycetota is the phylum containing the aquatic MAG with the biggest known estimated genome size (14.93 Mbp) (Rodríguez-Gijón et al., 2022). Further research to understand if extant genome size in phylum Plactomycetota is the result of taxonomical divergence or ecological adaptation to abiotic gradients would enlighten the evolution in this group. 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 has been widely reported to be streamlined, abundant and ubiquitous (Giovannoni, 2017).

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) (t-test, $p < 0.01$) (**Figure 3A**). Bacteria from sediments presented the biggest 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) (**Figure 3B**). This is supported by previous results, as Bacteria show bigger genome sizes than Archaea regardless of the environment (Rodríguez-Gijón et al., 2022). Moreover, the average estimated genome size of Baltic sediments (3.47 Mbp) is more similar to that of terrestrial microbial genomes (3.7 Mbp) (Rodríguez-Gijón et al., 2022). We also observe that pelagic microorganisms tend to be more streamlined than those present in sediments, at the phyla (**Figure 3C**), class (data not shown) and order level (**Figure 3D**). Our results corroborate previous findings that streamlining is common in pelagic marine environments (Batut et al., 2014; Giovannoni, 2017, 2005) and pelagic brackish environments (Hugerth et al., 2015).

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

(**Figure 2A** and **3A**). We calculated the AGS for the metagenomic samples to estimate the average genome size of the whole microbial community. We used MicrobeCensus as a robust and accurate tool to calculate AGS (Pereira-Flores et al., 2019). However, this AGS of metagenomes is biased by the viral and eukaryotic reads that might be present in the sample (Kristensen et al., 2010; Lind and Pollard, 2021). On the other hand, there is a bias with looking at the average estimated genome size of MAGs. This bias stems from all those bacteria and archaea that are hard to assemble and bin due to high genomic intrapopulation diversity (Ghurye, 2016). 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.

In our available homogeneous metadata for all datasets, we could not find a clear answer of why benthic genome size is bigger than pelagic genome size. While there is accompanying information on DOC for pelagic metagenomes and organic carbon for benthic metagenomes, they use methods and metrics that are non-comparable. Luckily a previous study has compiled information on organic carbon stocks in the Baltic Sea (Scheffold and Hense, 2020): 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 available. After all, a positive correlation between genome size and nutrient concentration has been shown before (Allen et al., 2012; Aylward and Santoro, 2020; Mende et al., 2017).

Brackish pelagic microorganisms tend to show smaller genome sizes than marine and freshwater. We compare the average estimated genome size of pelagic Baltic Sea microorganisms to previously published genome size estimations (Rodríguez-Gijón et al., 2022). This comparison includes all taxonomic groups found in three large datasets that includes 4051 freshwater representative MAGs, 2118 marine representative MAGs and 341 pelagic representative MAGs from the Baltic Sea (completeness >75%) (Alneberg et al., 2020; Buck et al., 2021; Nayfach et al., 2021).We found that the average estimated genome

size of the brackish pelagic MAGs (2.96 Mbp) is lower than marine MAGs (3.10 Mbp), and these two smaller than freshwater MAGs (3.48 Mbp) (t-test, $p < 0.01$) (**Figure 4A**).

We then subset all MAGs into phyla and focus on the most common from aquatic environments: phyla Actinobacteriota, Bacteroidota, Cyanobacteria and Proteobacteria (classes Alpha and Gammaproteobacteria) (**Figure 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 (**Figure 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 (**Figure 4B**). This result updates previous observations on aquatic Actinobacteriota genome size variation (Hugerth et al., 2015). 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 (**Figure 4D**). Similar trends were observed for isolates and MAGs of picocyanobacteria (Cabello-Yeves et al., 2022; Sánchez-Baracaldo et al., 2019). 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 3**) (Alvarenga et al., 2017; Kim Tiam et al., 2019). All in all, these results support that both environment (Rodríguez-Gijón et al., 2022) and taxonomy (Martinez-Gutierrez and Aylward, 2022) impact on existent pelagic genome size distributions.

Streamlining may affect certain functional categories in Bacteria. 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 (**Figure 5 A-R** and **Supplementary Material 2**). Functional categories included different but related metabolic pathways (KEGG modules), and each module comprises multiple specific metabolic reactions (module steps) (Zhou et al., 2022). In our results, we observe two different patterns: negative correlation between estimated genome size and number of module steps per Mbp, or the almost absence of the module steps (**Figure 5**). For 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 (**Figures 5A, D** & **J**). Genome size does not explain the

number of module steps per Mbp in all functional categories and metabolisms, such as drug resistance and transport systems (**Supplementary Material 2F** & **N**). These results suggest that streamlining of genomes select for specific functions and not the whole genome. This can be explained by the Black Queen Hypothesis (Morris et al., 2012), where only functions that are more costly than beneficial are lost. The loss of those specific functions has consequential long-lasting metabolic partnerships within the community (Mas et al., 2016).

When considering the total number of module steps, we observe that sediments code for a higher number of module steps than pelagic Bacteria in amino acid metabolism, aminoacyl tRNAs, carbon fixation, nitrogen and sulfur metabolism (**Figures 5B, E, H, N** & **Q**) (t-test, $p < 0.01$). This could be the result of a cofounding effect of genome size: microorganisms from sediments have bigger genome sizes than those in the water column (**Figure 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 can observe that pelagic Bacteria present a greater number of module steps per Mbp than sediment Bacteria in amino acid metabolism (**Figure 5C**), aminoacyl tRNAs (**Figure 5F**), carbon fixation (**Figure 5I**) and central carbohydrate metabolism (**Figure 5L**). However, nitrogen metabolism shows the opposite trend: sediment Bacteria have a greater number of module steps per Mbp than water column Bacteria (**Figure 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 (Lønborg and Markager, 2021). This indicates that benthic Bacteria potentially play a bigger role than pelagic bacteria in nitrogen cycling of autotropic systems like the Baltic Sea (Albert et al., 2021; Griffiths et al., 2017).

Moreover, although the majority 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 2**). Our results confirm that aquatic sediments are reservoirs of antibiotic resistance genes (Guo et al., 2020; Marti et al., 2014). Just as sediments harbor more than double the organic carbon than the water column, this allows

337 microbial genomes to have a bigger size and code for a higher number of genes. This allows
338 microbes to upkeep non-essential functions and allow metabolic reservoir in the sediments.
339
340 **Baltic sediments and water column harbor various Bacteria with different**
341 **metabolic capabilities.** From all >90% completeness MAGs, we selected only bacterial
342 phyla with five or more high-quality MAGs to observe how metabolism differs between
343 different taxa in Baltic sediments and water column (**Figure 6**).
344
345 We observed the presence of the genes *cdhH* | *cdhE* | *cooS* from the Wood-Ljungdhal
346 pathway exclusively in marine sediments (**Figure 6**), particularly in phyla Desulfobacterota,
347 Desulfobacterota E and Verrucomicrobiota (class Kiritimatiellae). This is a carbon fixation
348 pathway predominant in acetogenic bacteria found in anoxic conditions (Esposito et al.,
349 2019). Complementarily, we also find putative fermentation genes for acetogenesis (*acdA* |
350 *ack* | *pta*) to be widespread across taxa in Baltic sediments. This would explain the potential
351 success of acetogenic metabolism in brackish sediments (Lever, 2012). We also find that the
352 *acs* gene for acetate fermentation into acetyl-CoA is widely distributed in both sediments and
353 water column. These results support the common distribution of acetogens and the Wood-
354 Ljungdhal pathway in Baltic glacial sediments (Marshall et al., 2018).
355
356 No FeFe hydrogenases were detected, but different NiFe hydrogenases were spotted
357 to differ between environments: NiFe groups 3abd were detected mainly in pelagic bacteria,
358 while NiFe group 1 in sediments (**Figure 6**). NiFe group 1 hydrogenases could be playing a
359 vital role in nitrate (NO_3^-), sulfate (SO_4^{2-}) and iron (Fe^{3+}) reduction: these molecules can act
360 as acceptors of electrons coupled to H_2 oxidation in anoxic conditions (Peters et al., 2015).
361 Moreover, putative genes coding for the reduction of the above-mentioned molecules were
362 also detected on our sediment dataset (*napAB* | *narGH* for nitrate reduction, *aprA* | *sat* for
363 sulfate reduction, and iron reduction series genes). These results suggest a key role of
364 sediment Bacteria in sulfur, nitrogen, and iron cycling in the Baltic Sea. For example,
365 sulphate reducers such as Desulfobacterota found in our benthic MAGs collection, most
366 likely contribute to the release of Fe-bound phosphorus from sediments to the water column
367 (Sinkko, 2013).
368
369 **Conclusions**. Our analyses delve into genome size distributions across the sediments and
370 water column in the Baltic Sea, and its relationships with metabolic potential, taxonomic

identity and abiotic variables. Our key findings show that these three ecological and evolutionary variables are linked to genome size and, while each of them may partially explain the genome size patterns, it is the interplay between the three of them that drives genome size variation in Bacteria in the Baltic Sea. We also provide the first comprehensive overview of microbial MAGs comparing genome size variation across the three major pelagic ecosystems, where brackish Bacteria have smaller estimated genome sizes than marine and freshwater. Within the Baltic Sea, benthic Bacteria have the biggest estimated genome size. As a future perspective, with the continuous progression of aquatic microbial ecology and the development of new isolation, omics and bioinformatic techniques, we will be able to grasp a more complete and unbiased view of genome sizes distribution in nature and, therefore, microbial evolution and ecology.

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 (**Supplementary Table 1** and **Figure 1**). We compiled 108 sediment metagenomes, of which 59 were collected in 2019 and recently published (Broman et al., 2022), 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 (Alneberg et al., 2020, 2018; Larsson et al., 2014). All five datasets have abiotic metadata of depth (m), salinity (PSU), temperature (C) and oxygen concentration (mg/L) (**Supplementary Table 1**).

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) (**Supplementary Table 1** 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 (Titus Brown and Irber, 2016) to create signatures, Megahit (Li et al., 2016) to obtain single-sample assemblies and Metabat2 (Kang et al., 2019) 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) (Parks et al., 2015): we used the typical workflow (*lineage_wf*) with default parameters, and selected only bins 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) (Chaumeil et al., 2019) according to the GTDB classification (Parks et al., 2020). 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 (Buck et al., 2021) 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 3**). We obtained the custom marker gene set using Ac1 circular genomes (Kang et al., 2017; Neuenschwander et al., 2018). In the case of phylum Patescibacteria, we used a custom set of maker genes provided by CheckM (Brown et al., 2015; Parks et al., 2015).

Complementarily, we used 1920 pelagic MAGs that were published (Alneberg et al., 2020) 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) (Buck et al., 2022; Jain et al., 2018). From the 216 >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 **Supplementary Table 2**.

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) (Nayfach and Pollard, 2015). 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 (Pereira-Flores et al., 2019). We excluded one of the pelagic datasets (Alneberg et al., 2018) 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 (Rodríguez-Gijón et al., 2022) (**Supplementary Table 3**).

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 1 241 MAGs from the water column. The metabolic potential of sediment and pelagic MAGs was reconstructed using Prodigal annotations (v2.6.3) (Hyatt et al., 2010). We used the resulting protein translation files to predict biogeochemical and metabolic functional traits using METABOLIC (v4.0) (Zhou et al., 2022). 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 (**Figures 2, 3** and **5**). Stars in boxplots indicate significant differences $p < 0.05$. In **Figures 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 (R Core Team, 2020) (**Figure S1**). We obtained the correlation coefficients on scatterplots (**Figure S1** and **Figure 5**) to test the fit of our data to linear regressions using the function *stat_cor* from the R package *ggpubr* v0.4.0 (Kassambara, A., 2020).

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 **Supplementary Table 1**. 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.

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.

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.

CONFLICTS OF INTEREST

The authors declare that there are no competing interests.

REFERENCES

Acinas, S.G., Sánchez, P., Salazar, G., Cornejo-Castillo, F.M., Sebastián, M., Logares, R., Royo-Llonch, M., Paoli, L., Sunagawa, S., Hingamp, P., Ogata, H., Lima-Mendez, G., Roux, S., González, J.M., Arrieta, J.M., Alam, I.S., Kamau, A., Bowler, C., Raes, J., Pesant, S., Bork, P., Agustí, S., Gojobori, T., Vaqué, D., Sullivan, M.B., Pedrós-Alió, C., Massana, R., Duarte, C.M., Gasol, J.M., 2021. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. Commun Biol 4, 604. https://doi.org/10.1038/s42003-021-02112-2

Albert, S., Bonaglia, S., Stjärnkvist, N., Winder, M., Thamdrup, B., Nascimento, F.J.A., 2021. Influence of settling organic matter quantity and quality on benthic nitrogen cycling. Limnol Oceanogr 66, 1882–1895. https://doi.org/10.1002/lno.11730

Allen, L.Z., Allen, E.E., Badger, J.H., McCrow, J.P., Paulsen, I.T., Elbourne, L.D., Thiagarajan, M., Rusch, D.B., Nealson, K.H., Williamson, S.J., Venter, J.C., Allen, A.E., 2012. Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic. ISME J 6, 1403–1414. https://doi.org/10.1038/ismej.2011.201

Alneberg, J., Bennke, C., Beier, S., Bunse, C., Quince, C., Ininbergs, K., Riemann, L., Ekman, M., Jürgens, K., Labrenz, M., Pinhassi, J., Andersson, A.F., 2020. Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. Commun Biol 3, 119. https://doi.org/10.1038/s42003-020-0856-x

Alneberg, J., Sundh, J., Bennke, C., Beier, S., Lundin, D., Hugerth, L.W., Pinhassi, J., Kisand, V., Riemann, L., Jürgens, K., Labrenz, M., Andersson, A.F., 2018. BARM and BalticMicrobeDB, a reference metagenome and interface to meta-omic data for the Baltic Sea. Sci Data 5, 180146. https://doi.org/10.1038/sdata.2018.146

Alvarenga, D.O., Fiore, M.F., Varani, A.M., 2017. A Metagenomic Approach to Cyanobacterial Genomics. Front. Microbiol. 8, 809. https://doi.org/10.3389/fmicb.2017.00809

Aylward, F.O., Santoro, A.E., 2020. Heterotrophic Thaumarchaea with Small Genomes Are Widespread in the Dark Ocean. mSystems 5, e00415-20, /msystems/5/3/msys.00415-20.atom. https://doi.org/10.1128/mSystems.00415-20

Batut, B., Knibbe, C., Marais, G., Daubin, V., 2014. Reductive genome evolution at both ends of the bacterial population size spectrum. Nat Rev Microbiol 12, 841–850. https://doi.org/10.1038/nrmicro3331

551 Beam, J.P., Becraft, E.D., Brown, J.M., Schulz, F., Jarett, J.K., Bezuidt, O., Poulton, N.J.,
552 Clark, K., Dunfield, P.F., Ravin, N.V., Spear, J.R., Hedlund, B.P., Kormas, K.A.,
553 Sievert, S.M., Elshahed, M.S., Barton, H.A., Stott, M.B., Eisen, J.A., Moser, D.P.,
554 Onstott, T.C., Woyke, T., Stepanauskas, R., 2020. Ancestral Absence of Electron
555 Transport Chains in Patescibacteria and DPANN. Front. Microbiol. 11, 1848.
556 https://doi.org/10.3389/fmicb.2020.01848
557 Biller, S.J., Berube, P.M., Lindell, D., Chisholm, S.W., 2015. Prochlorococcus: the structure
558 and function of collective diversity. Nat Rev Microbiol 13, 13–27.
559 https://doi.org/10.1038/nrmicro3378
560 Broman, E., Izabel-Shen, D., Rodríguez-Gijón, A., Bonaglia, S., Garcia, S.L., Nascimento,
561 F.J.A., 2022. Microbial functional genes are driven by gradients in sediment
562 stoichiometry, oxygen, and salinity across the Baltic benthic ecosystem. Microbiome
563 10, 126. https://doi.org/10.1186/s40168-022-01321-z
564 Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., Wilkins, M.J.,
565 Wrighton, K.C., Williams, K.H., Banfield, J.F., 2015. Unusual biology across a group
566 comprising more than 15% of domain Bacteria. Nature 523, 208–211.
567 https://doi.org/10.1038/nature14486
568 Buck, M., Garcia, S.L., Fernandez, L., Martin, G., Martinez-Rodriguez, G.A., Saarenheimo,
569 J., Zopfi, J., Bertilsson, S., Peura, S., 2021. Comprehensive dataset of shotgun
570 metagenomes from oxygen stratified freshwater lakes and ponds. Sci Data 8, 131.
571 https://doi.org/10.1038/s41597-021-00910-1
572 Buck, M., Mehrshad, M., Bertilsson, S., 2022. mOTUpan: a robust Bayesian approach to
573 leverage metagenome-assembled genomes for core-genome estimation. NAR
574 Genomics and Bioinformatics 4, lqac060. https://doi.org/10.1093/nargab/lqac060
575 Cabello-Yeves, P.J., Callieri, C., Picazo, A., Schallenberg, L., Huber, P., Roda-Garcia, J.J.,
576 Bartosiewicz, M., Belykh, O.I., Tikhonova, I.V., Torcello-Requena, A., De Prado,
577 P.M., Puxty, R.J., Millard, A.D., Camacho, A., Rodriguez-Valera, F., Scanlan, D.J.,
578 2022. Elucidating the picocyanobacteria salinity divide through ecogenomics of new
579 freshwater isolates. BMC Biol 20, 175. https://doi.org/10.1186/s12915-022-01379-z
580 Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., Parks, D.H., 2019. GTDB-Tk: a toolkit to
581 classify genomes with the Genome Taxonomy Database. Bioinformatics btz848.
582 https://doi.org/10.1093/bioinformatics/btz848
583 Chen, M.-Y., Teng, W.-K., Zhao, L., Hu, C.-X., Zhou, Y.-K., Han, B.-P., Song, L.-R., Shu, W.-
584 S., 2021. Comparative genomics reveals insights into cyanobacterial evolution and
585 habitat adaptation. ISME J 15, 211–227. https://doi.org/10.1038/s41396-020-00775-z
586 Coelho, L.P., Alves, R., del Río, Á.R., Myers, P.N., Cantalapiedra, C.P., Giner-Lamia, J.,
587 Schmidt, T.S., Mende, D.R., Orakov, A., Letunic, I., Hildebrand, F., Van Rossum, T.,
588 Forslund, S.K., Khedkar, S., Maistrenko, O.M., Pan, S., Jia, L., Ferretti, P.,
589 Sunagawa, S., Zhao, X.-M., Nielsen, H.B., Huerta-Cepas, J., Bork, P., 2022.
590 Towards the biogeography of prokaryotic genes. Nature 601, 252–256.
591 https://doi.org/10.1038/s41586-021-04233-4
592 Conley, D.J., Björck, S., Bonsdorff, E., Carstensen, J., Destouni, G., Gustafsson, B.G.,
593 Hietanen, S., Kortekaas, M., Kuosa, H., Markus Meier, H.E., Müller-Karulis, B.,
594 Nordberg, K., Norkko, A., Nürnberg, G., Pitkänen, H., Rabalais, N.N., Rosenberg, R.,
595 Savchuk, O.P., Slomp, C.P., Voss, M., Wulff, F., Zillén, L., 2009. Hypoxia-Related
596 Processes in the Baltic Sea. Environ. Sci. Technol. 43, 3412–3420.
597 https://doi.org/10.1021/es802762a
598 Esposito, A., Tamburini, S., Triboli, L., Ambrosino, L., Chiusano, M.L., Jousson, O., 2019.
599 Insights into the genome structure of four acetogenic bacteria with specific reference
600 to the Wood–Ljungdahl pathway. MicrobiologyOpen 8.
601 https://doi.org/10.1002/mbo3.938
602 Ghurye, J.S., 2016. Metagenomic Assembly: Overview, Challenges and Applications 10.
603 Giovannoni, S.J., 2017. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. Annu.
604 Rev. Mar. Sci. 9, 231–255. https://doi.org/10.1146/annurev-marine-010814-015934

605 Giovannoni, S.J., 2005. Genome Streamlining in a Cosmopolitan Oceanic Bacterium.
606 Science 309, 1242–1245. https://doi.org/10.1126/science.1114057
607 Giovannoni, S.J., Cameron Thrash, J., Temperton, B., 2014. Implications of streamlining
608 theory for microbial ecology. ISME J 8, 1553–1565.
609 https://doi.org/10.1038/ismej.2014.60
610 Griffiths, J.R., Kadin, M., Nascimento, F.J.A., Tamelander, T., Törnroos, A., Bonaglia, S.,
611 Bonsdorff, E., Brüchert, V., Gårdmark, A., Järnström, M., Kotta, J., Lindegren, M.,
612 Nordström, M.C., Norkko, A., Olsson, J., Weigel, B., Žydelis, R., Blenckner, T.,
613 Niiranen, S., Winder, M., 2017. The importance of benthic-pelagic coupling for
614 marine ecosystem functioning in a changing world. Glob Change Biol 23, 2179–2196.
615 https://doi.org/10.1111/gcb.13642
616 Grote, J., Thrash, J.C., Huggett, M.J., Landry, Z.C., Carini, P., Giovannoni, S.J., Rappé,
617 M.S., 2012. Streamlining and Core Genome Conservation among Highly Divergent
618 Members of the SAR11 Clade. mBio 3, e00252-12.
619 https://doi.org/10.1128/mBio.00252-12
620 Guo, X., Zhao, S., Chen, Y., Yang, J., Hou, L., Liu, M., Yang, Y., 2020. Antibiotic resistance
621 genes in sediments of the Yangtze Estuary: From 2007 to 2019. Science of The Total
622 Environment 744, 140713. https://doi.org/10.1016/j.scitotenv.2020.140713
623 Herlemann, D.P., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J.J., Andersson, A.F.,
624 2011. Transitions in bacterial communities along the 2000 km salinity gradient of the
625 Baltic Sea. ISME J 5, 1571–1579. https://doi.org/10.1038/ismej.2011.41
626 Hugerth, L.W., Larsson, J., Alneberg, J., Lindh, M.V., Legrand, C., Pinhassi, J., Andersson,
627 A.F., 2015. Metagenome-assembled genomes uncover a global brackish
628 microbiome. Genome Biol 16, 279. https://doi.org/10.1186/s13059-015-0834-7
629 Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010.
630 Prodigal: prokaryotic gene recognition and translation initiation site identification.
631 BMC Bioinformatics 11, 119. https://doi.org/10.1186/1471-2105-11-119
632 Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., Aluru, S., 2018. High
633 throughput ANI analysis of 90K prokaryotic genomes reveals clear species
634 boundaries. Nat Commun 9, 5114. https://doi.org/10.1038/s41467-018-07641-9
635 Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., Wang, Z., 2019. MetaBAT 2: an
636 adaptive binning algorithm for robust and efficient genome reconstruction from
637 metagenome assemblies. PeerJ 7, e7359. https://doi.org/10.7717/peerj.7359
638 Kang, I., Kim, S., Islam, Md.R., Cho, J.-C., 2017. The first complete genome sequences of
639 the acI lineage, the most abundant freshwater Actinobacteria, obtained by whole-
640 genome-amplification of dilution-to-extinction cultures. Sci Rep 7, 42252.
641 https://doi.org/10.1038/srep42252
642 Kassambara, A., 2020. ggpubr: “ggplot2” Based Publication Ready Plots.
643 Kim Tiam, S., Comte, K., Dalle, C., Duval, C., Pancrace, C., Gugger, M., Marie, B.,
644 Yéprémian, C., Bernard, C., 2019. Development of a new extraction method based
645 on high-intensity ultra-sonication to study RNA regulation of the filamentous
646 cyanobacteria Planktothrix. PLoS ONE 14, e0222029.
647 https://doi.org/10.1371/journal.pone.0222029
648 Konstantinidis, K.T., Braff, J., Karl, D.M., DeLong, E.F., 2009. Comparative Metagenomic
649 Analysis of a Microbial Community Residing at a Depth of 4,000 Meters at Station
650 ALOHA in the North Pacific Subtropical Gyre. Appl Environ Microbiol 75, 5345–5355.
651 https://doi.org/10.1128/AEM.00473-09
652 Kristensen, D.M., Mushegian, A.R., Dolja, V.V., Koonin, E.V., 2010. New dimensions of the
653 virus world discovered through metagenomics. Trends in Microbiology 18, 11–19.
654 https://doi.org/10.1016/j.tim.2009.11.003
655 Larsson, J., Celepli, N., Ininbergs, K., Dupont, C.L., Yooseph, S., Bergman, B., Ekman, M.,
656 2014. Picocyanobacteria containing a novel pigment gene cluster dominate the
657 brackish water Baltic Sea. ISME J 8, 1892–1903.
658 https://doi.org/10.1038/ismej.2014.35

659 Lever, M.A., 2012. Acetogenesis in the Energy-Starved Deep Biosphere – A Paradox?
660 Front. Microbio. 2. https://doi.org/10.3389/fmicb.2011.00284
661 Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., Lam, T.-
662 W., 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by
663 advanced methodologies and community practices. Methods 102, 3–11.
664 https://doi.org/10.1016/j.ymeth.2016.02.020
665 Li, M., Wei, G., Shi, W., Sun, Z., Li, H., Wang, X., Gao, Z., 2018. Distinct distribution patterns
666 of ammonia-oxidizing archaea and bacteria in sediment and water column of the
667 Yellow River estuary. Sci Rep 8, 1584. https://doi.org/10.1038/s41598-018-20044-6
668 Lind, A.L., Pollard, K.S., 2021. Accurate and sensitive detection of microbial eukaryotes from
669 whole metagenome shotgun sequencing. Microbiome 9, 58.
670 https://doi.org/10.1186/s40168-021-01015-y
671 Lønborg, C., Markager, S., 2021. Nitrogen in the Baltic Sea: Long-term trends, a budget and
672 decadal time lags in responses to declining inputs. Estuarine, Coastal and Shelf
673 Science 261, 107529. https://doi.org/10.1016/j.ecss.2021.107529
674 Louca, S., Parfrey, L.W., Doebeli, M., 2016. Decoupling function and taxonomy in the global
675 ocean microbiome. Science 353, 1272–1277.
676 https://doi.org/10.1126/science.aaf4507
677 Lynch, M., 2006. Streamlining and Simplification of Microbial Genome Architecture 25.
678 Maistrenko, O.M., Mende, D.R., Luetge, M., Hildebrand, F., Schmidt, T.S.B., Li, S.S.,
679 Rodrigues, J.F.M., von Mering, C., Pedro Coelho, L., Huerta-Cepas, J., Sunagawa,
680 S., Bork, P., 2020. Disentangling the impact of environmental and phylogenetic
681 constraints on prokaryotic within-species diversity. ISME J.
682 https://doi.org/10.1038/s41396-020-0600-z
683 Marshall, I.P.G., Karst, S.M., Nielsen, P.H., Jørgensen, B.B., 2018. Metagenomes from deep
684 Baltic Sea sediments reveal how past and present environmental conditions
685 determine microbial community composition. Marine Genomics 37, 58–68.
686 https://doi.org/10.1016/j.margen.2017.08.004
687 Marti, E., Variatza, E., Balcazar, J.L., 2014. The role of aquatic ecosystems as reservoirs of
688 antibiotic resistance. Trends in Microbiology 22, 36–41.
689 https://doi.org/10.1016/j.tim.2013.11.001
690 Martinez-Gutierrez, C.A., Aylward, F.O., 2022. Genome size distributions in bacteria and
691 archaea are strongly linked to evolutionary history at broad phylogenetic scales.
692 PLoS Genet 18, e1010220. https://doi.org/10.1371/journal.pgen.1010220
693 Mas, A., Jamshidi, S., Lagadeuc, Y., Eveillard, D., Vandenkoornhuyse, P., 2016. Beyond the
694 Black Queen Hypothesis. ISME J 10, 2085–2091.
695 https://doi.org/10.1038/ismej.2016.22
696 Mende, D.R., Bryant, J.A., Aylward, F.O., Eppley, J.M., Nielsen, T., Karl, D.M., DeLong,
697 E.F., 2017. Environmental drivers of a microbial genomic transition zone in the
698 ocean's interior. Nat Microbiol 2, 1367–1373. https://doi.org/10.1038/s41564-017-
699 0008-3
700 Moran, N.A., Mira, A., 2001. The process of genome shrinkage in the obligate symbiont
701 Buchnera aphidicola. Genome Biol 2, research0054.1. https://doi.org/10.1186/gb-
702 2001-2-12-research0054
703 Morris, J.J., Lenski, R.E., Zinser, E.R., 2012. The Black Queen Hypothesis: Evolution of
704 Dependencies through Adaptive Gene Loss. mBio 3, e00036-12.
705 https://doi.org/10.1128/mBio.00036-12
706 Nayfach, S., Pollard, K.S., 2015. Average genome size estimation improves comparative
707 metagenomics and sheds light on the functional ecology of the human microbiome.
708 Genome Biol 16, 51. https://doi.org/10.1186/s13059-015-0611-7
709 Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-
710 Espino, D., Chen, I.-M., Huntemann, M., Palaniappan, K., Ladau, J., Mukherjee, S.,
711 Reddy, T.B.K., Nielsen, T., Kirton, E., Faria, J.P., Edirisinghe, J.N., Henry, C.S.,
712 Jungbluth, S.P., Chivian, D., Dehal, P., Wood-Charlson, E.M., Arkin, A.P., Tringe,
713 S.G., Visel, A., IMG/M Data Consortium, Abreu, H., Acinas, S.G., Allen, E., Allen,

M.A., Alteio, L.V., Andersen, G., Anesio, A.M., Attwood, G., Avila-Magaña, V., Badis, Y., Bailey, J., Baker, B., Baldrian, P., Barton, H.A., Beck, D.A.C., Becraft, E.D., Beller, H.R., Beman, J.M., Bernier-Latmani, R., Berry, T.D., Bertagnolli, A., Bertilsson, S., Bhatnagar, J.M., Bird, J.T., Blanchard, J.L., Blumer-Schuette, S.E., Bohannan, B., Borton, M.A., Brady, A., Brawley, S.H., Brodie, J., Brown, S., Brum, J.R., Brune, A., Bryant, D.A., Buchan, A., Buckley, D.H., Buongiorno, J., Cadillo-Quiroz, H., Caffrey, S.M., Campbell, A.N., Campbell, B., Carr, S., Carroll, J., Cary, S.C., Cates, A.M., Cattolico, R.A., Cavicchioli, R., Chistoserdova, L., Coleman, M.L., Constant, P., Conway, J.M., Mac Cormack, W.P., Crowe, S., Crump, B., Currie, C., Daly, R., DeAngelis, K.M., Denef, V., Denman, S.E., Desta, A., Dionisi, H., Dodsworth, J., Dombrowski, N., Donohue, T., Dopson, M., Driscoll, T., Dunfield, P., Dupont, C.L., Dynarski, K.A., Edgcomb, V., Edwards, E.A., Elshahed, M.S., Figueroa, I., Flood, B., Fortney, N., Fortunato, C.S., Francis, C., Gachon, C.M.M., Garcia, S.L., Gazitua, M.C., Gentry, T., Gerwick, L., Gharechahi, J., Girguis, P., Gladden, J., Gradoville, M., Grasby, S.E., Gravuer, K., Grettenberger, C.L., Gruninger, R.J., Guo, J., Habteselassie, M.Y., Hallam, S.J., Hatzenpichler, R., Hausmann, B., Hazen, T.C., Hedlund, B., Henny, C., Herfort, L., Hernandez, M., Hershey, O.S., Hess, M., Hollister, E.B., Hug, L.A., Hunt, D., Jansson, J., Jarett, J., Kadnikov, V.V., Kelly, C., Kelly, R., Kelly, W., Kerfeld, C.A., Kimbrel, J., Klassen, J.L., Konstantinidis, K.T., Lee, L.L., Li, W.-J., Loder, A.J., Loy, A., Lozada, M., MacGregor, B., Magnabosco, C., Maria da Silva, A., McKay, R.M., McMahon, K., McSweeney, C.S., Medina, M., Meredith, L., Mizzi, J., Mock, T., Momper, L., Moran, M.A., Morgan-Lang, C., Moser, D., Muyzer, G., Myrold, D., Nash, M., Nesbø, C.L., Neumann, A.P., Neumann, R.B., Noguera, D., Northen, T., Norton, J., Nowinski, B., Nüsslein, K., O'Malley, M.A., Oliveira, R.S., Maia de Oliveira, V., Onstott, T., Osvatic, J., Ouyang, Y., Pachiadaki, M., Parnell, J., Partida-Martinez, L.P., Peay, K.G., Pelletier, D., Peng, X., Pester, M., Pett-Ridge, J., Peura, S., Pjevac, P., Plominsky, A.M., Poehlein, A., Pope, P.B., Ravin, N., Redmond, M.C., Reiss, R., Rich, V., Rinke, C., Rodrigues, J.L.M., Rodriguez-Reillo, W., Rossmassler, K., Sackett, J., Salekdeh, G.H., Saleska, S., Scarborough, M., Schachtman, D., Schadt, C.W., Schrenk, M., Sczyrba, A., Sengupta, A., Setubal, J.C., Shade, A., Sharp, C., Sherman, D.H., Shubenkova, O.V., Sierra-Garcia, I.N., Simister, R., Simon, H., Sjöling, S., Slonczewski, J., Correa de Souza, R.S., Spear, J.R., Stegen, J.C., Stepanauskas, R., Stewart, F., Suen, G., Sullivan, M., Sumner, D., Swan, B.K., Swingley, W., Tarn, J., Taylor, G.T., Teeling, H., Tekere, M., Teske, A., Thomas, T., Thrash, C., Tiedje, J., Ting, C.S., Tully, B., Tyson, G., Ulloa, O., Valentine, D.L., Van Goethem, M.W., VanderGheynst, J., Verbeke, T.J., Vollmers, J., Vuillemin, A., Waldo, N.B., Walsh, D.A., Weimer, B.C., Whitman, T., van der Wielen, P., Wilkins, M., Williams, T.J., Woodcroft, B., Woolet, J., Wrighton, K., Ye, J., Young, E.B., Youssef, N.H., Yu, F.B., Zemskaya, T.I., Ziels, R., Woyke, T., Mouncey, N.J., Ivanova, N.N., Kyrpides, N.C., Eloe-Fadrosh, E.A., 2021. A genomic catalog of Earth's microbiomes. Nat Biotechnol 39, 499–509. https://doi.org/10.1038/s41587-020-0718-6

Neuenschwander, S.M., Ghai, R., Pernthaler, J., Salcher, M.M., 2018. Microdiversification in genome-streamlined ubiquitous freshwater Actinobacteria. ISME J 12, 185–198. https://doi.org/10.1038/ismej.2017.156

Nielsen, D.A., Fierer, N., Geoghegan, J.L., Gillings, M.R., Gumerov, V., Madin, J.S., Moore, L., Paulsen, I.T., Reddy, T.B.K., Tetu, S.G., Westoby, M., 2021. Aerobic bacteria and archaea tend to have larger and more versatile genomes. Oikos 130, 501–511. https://doi.org/10.1111/oik.07912

Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., Hugenholtz, P., 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol 38, 1079–1086. https://doi.org/10.1038/s41587-020-0501-8

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25, 1043–1055. https://doi.org/10.1101/gr.186072.114

769 Pereira-Flores, E., Glöckner, F.O., Fernandez-Guerra, A., 2019. Fast and accurate average
770 genome size and 16S rRNA gene average copy number computation in
771 metagenomic data. BMC Bioinformatics 20, 453. https://doi.org/10.1186/s12859-019-
772 3031-y
773 Peters, J.W., Schut, G.J., Boyd, E.S., Mulder, D.W., Shepard, E.M., Broderick, J.B., King,
774 P.W., Adams, M.W.W., 2015. [FeFe]- and [NiFe]-hydrogenase diversity, mechanism,
775 and maturation. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research
776 1853, 1350–1369. https://doi.org/10.1016/j.bbamcr.2014.11.021
777 R Core Team, 2020. R: A language and environment for statistical computing.
778 Rodríguez-Gijón, A., Nuy, J.K., Mehrshad, M., Buck, M., Schulz, F., Woyke, T., Garcia, S.L.,
779 2022. A genomic perspective across Earth's microbiomes reveals that genome size
780 in Archaea and Bacteria is linked to ecosystem type and trophic strategy. Frontiers in
781 Microbiology 12. https://doi.org/10.3389/fmicb.2021.761869
782 Salcher, M.M., Schaefle, D., Kaspar, M., Neuenschwander, S.M., Ghai, R., 2019. Evolution
783 in action: habitat transition from sediment to the pelagial leads to genome
784 streamlining in Methylophilaceae. ISME J 13, 2764–2777.
785 https://doi.org/10.1038/s41396-019-0471-3
786 Sánchez-Baracaldo, P., Bianchini, G., Di Cesare, A., Callieri, C., Chrismas, N.A.M., 2019.
787 Insights Into the Evolution of Picocyanobacteria and Phycoerythrin Genes (mpeBA
788 and cpeBA). Front. Microbiol. 10, 45. https://doi.org/10.3389/fmicb.2019.00045
789 Scanlan, D.J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W.R., Post,
790 A.F., Hagemann, M., Paulsen, I., Partensky, F., 2009. Ecological Genomics of
791 Marine Picocyanobacteria. MMBR 73, 249–299.
792 https://doi.org/10.1128/MMBR.00035-08
793 Scheffold, M.I.E., Hense, I., 2020. Quantifying Contemporary Organic Carbon Stocks of the
794 Baltic Sea Ecosystem. Front. Mar. Sci. 7, 571956.
795 https://doi.org/10.3389/fmars.2020.571956
796 Simonsen, A.K., 2022. Environmental stress leads to genome streamlining in a widely
797 distributed species of soil bacteria. ISME J 16, 423–434.
798 https://doi.org/10.1038/s41396-021-01082-x
799 Sinkko, H., 2013. Sediment bacterial communities in nutrient cycling and in the history of the
800 Baltic Sea (Doctoral dissertation). University of Helsinki, Helsinki, Finland.
801 Swan, B.K., Tupper, B., Sczyrba, A., Lauro, F.M., Martinez-Garcia, M., Gonzalez, J.M., Luo,
802 H., Wright, J.J., Landry, Z.C., Hanson, N.W., Thompson, B.P., Poulton, N.J.,
803 Schwientek, P., Acinas, S.G., Giovannoni, S.J., Moran, M.A., Hallam, S.J.,
804 Cavicchioli, R., Woyke, T., Stepanauskas, R., 2013. Prevalent genome streamlining
805 and latitudinal divergence of planktonic bacteria in the surface ocean. Proceedings of
806 the National Academy of Sciences 110, 11463–11468.
807 https://doi.org/10.1073/pnas.1304246110
808 Titus Brown, C., Irber, L., 2016. sourmash: a library for MinHash sketching of DNA. JOSS 1,
809 27. https://doi.org/10.21105/joss.00027
810 van Ham, R.C.H.J., Kamerbeek, J., Palacios, C., Rausell, C., Abascal, F., Bastolla, U.,
811 Fernández, J.M., Jiménez, L., Postigo, M., Silva, F.J., Tamames, J., Viguera, E.,
812 Latorre, A., Valencia, A., Morán, F., Moya, A., 2003. Reductive genome evolution in
813 *Buchnera aphidicola*. Proc. Natl. Acad. Sci. U.S.A. 100, 581–586.
814 https://doi.org/10.1073/pnas.0235981100
815 Zhou, Z., Tran, P.Q., Breister, A.M., Liu, Y., Kieft, K., Cowley, E.S., Karaoz, U.,
816 Anantharaman, K., 2022. METABOLIC: high-throughput profiling of microbial
817 genomes for functional traits, metabolism, biogeochemistry, and community-scale
818 functional networks. Microbiome 10, 33. https://doi.org/10.1186/s40168-021-01213-8
819 Zhou, Z., Tran, P.Q., Kieft, K., Anantharaman, K., 2020. Genome diversification in globally
820 distributed novel marine Proteobacteria is linked to environmental adaptation. ISME J
821 14, 2060–2077. https://doi.org/10.1038/s41396-020-0669-4
822

FIGURES LEGENDS

Figure 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 Supplementary Table 1. Shape type indicates the reference and shape size indicates the AGS of the given metagenome.

Figure 2. Boxplots showing the AGS distribution of Baltic metagenomes. **Panel A** indicates the AGS distribution in both water column and sediments. **Panel 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).

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

Figure 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. **Panel A** shows the genome size distribution of Archaea and Bacteria obtained from Baltic water column and sediments for a total of 397 representative genomes. **Panel 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). **Panel C** shows the estimated genome size per phylum. We selected only phyla with at least 2 MAGs in each environment. **Panel 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).

Figure 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 (**Panel A**), only phylum Actinobacteriota (**Panel B**), phylum Bacteroidota (**Panel C**), phylum Cyanobacteria (**Panel D**), class Alphaproteobacteria (**Panel E**) and class Gammaproteobacteria (**Panel 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.

Figure 5. Overview of the presence of module steps for six metabolic categories: amino acid metabolism (**Panels A-C**), aminoacyl tRNAs (**Panels D-F**), carbon fixation (**Panels G-I**), central carbohydrate metabolism (**Panels J-L**), nitrogen metabolism (**Panels M-O**) and sulfur metabolism (**Panels 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). Panels B, E, H, K, N and Q indicate total number of module steps. Panels C, F, I, L, O and R indicate the number of module steps per Mbp.

Figure 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).

SUPPLEMENTARY MATERIAL

Figure S1. Overview of the average genome size (AGS) of metagenomes across gradients of salinity and depth.

Supplementary Material 1. Statistical analysis of the effect of abiotic variables and their interactions on the average genome size (AGS) of full Baltic microbial communities (**Figure 1**). Tables show the results of ANOVA type II analysis for both sediments and water column.

Supplementary Material 2. Overview of the presence of module steps for several metabolic categories, complementing **Figure 3**. We also provide a list of all functional categories used with the different module categories included, and the number of module steps used in the analysis.

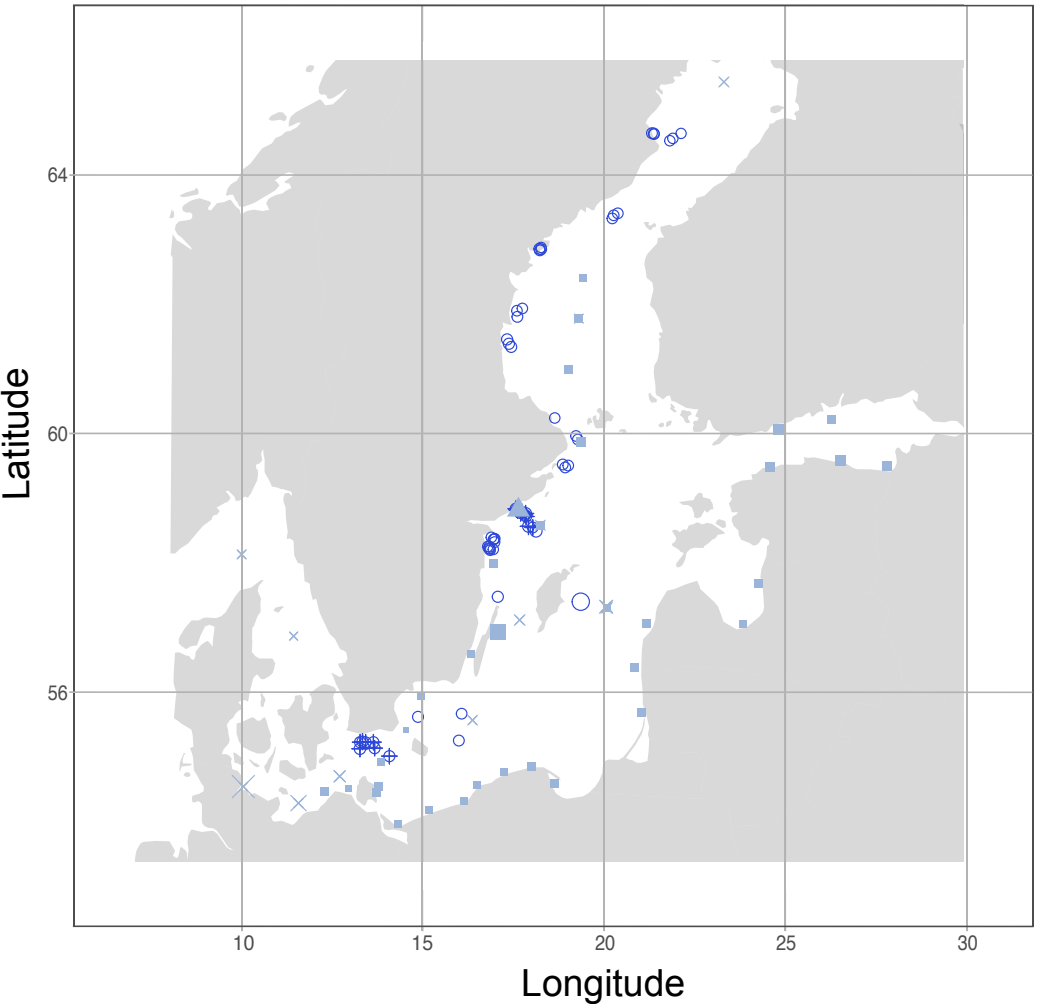
Supplementary Material 3. Table with all Actinobacteriota genomes used to calculate marker genes to use in CheckM quality assessment. We include information on GenBank Accession, lake, tribe, assembly size (Mbp), GC content (%) and reference. Two figures are included. First a boxplot showing the differences on completeness of the of the 8 phyla with most bins in the StratfreshDB database. We have included the completeness estimations using both default CheckM parameters and specific set of marker genes in both phyla Patescibacteria and Actinobacteriota. Stars indicate significant differences $p < 0.05$ (Wilcoxon non-parametric test). We also provide a figure showing the markers used by CheckM with default parameters for the 8 analyzed phyla.

Supplementary Table 1. Table with information of all metagenomes used in this research project, including: sample run in NCBI/ENA, estimated average genome size (Mbp), depth (m), salinity (PSU), temperature (C) and oxygen concentration (mg/L), sample material processing (μm), environment, latitude, longitude, study accession, online database, sample accession and reference.

Supplementary Table 2. Table with information of all metagenome assembled genomes (MAGs) used in this research project, including: bin id, mOTU, completeness (%), contamination (%), GC (%), coding density (%), bin length, estimated genome size (Mbp), environment, domain, phylum, class, order, family, genus and species.

Supplementary Table 3. Table with information of all pelagic metagenome assembled genomes (MAGs) used to compare genome size across major pelagic environments (marine, freshwater and brackish). All brackish MAGs belong (Alneberg et al., 2020), while the genomic information for marine and freshwater MAGs was retrieved from a previous study (Rodríguez-Gijón et al., 2022) that compiled them from (Buck et al., 2021; Nayfach et al.,

923 2021). We include bin id, completeness (%), contamination (%), bin length, estimated
924 genome size (Mbp), ecosystem type, domain, phylum, class, order, family, genus and species.
925



Reference

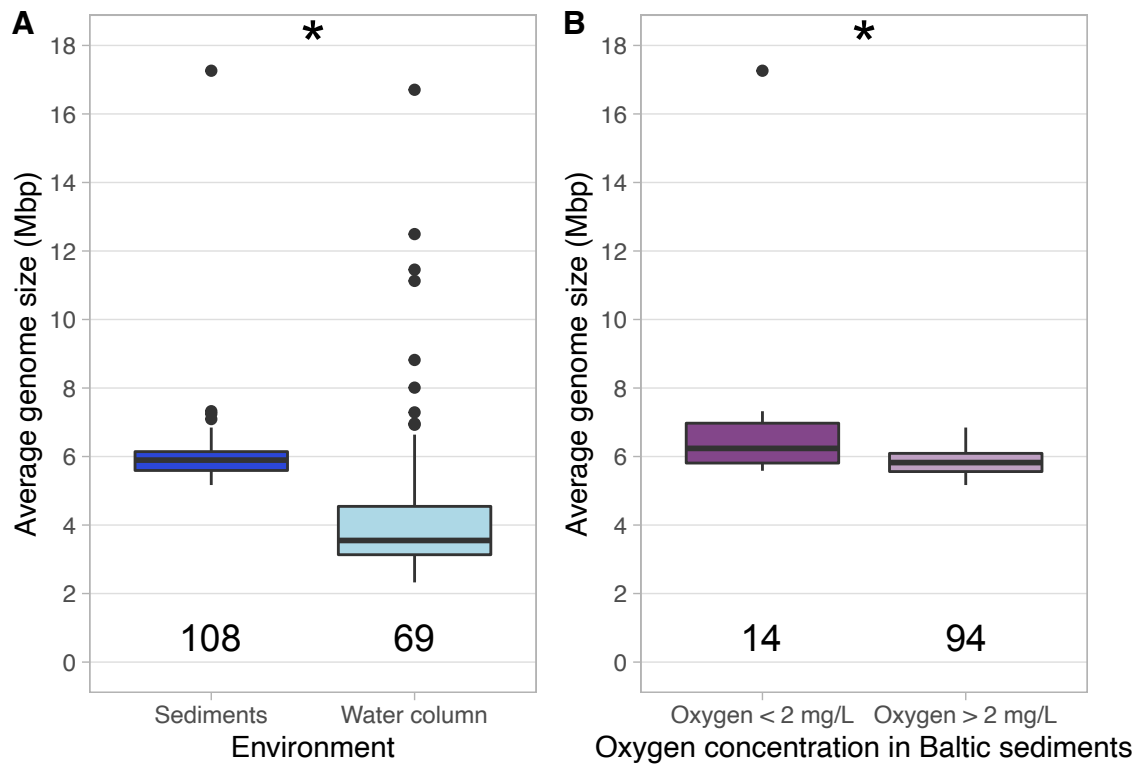
- Broman et al., 2022
- Larsson et al., 2014
- Alneberg et al., 2020
- Alneberg et al., 2018
- This study

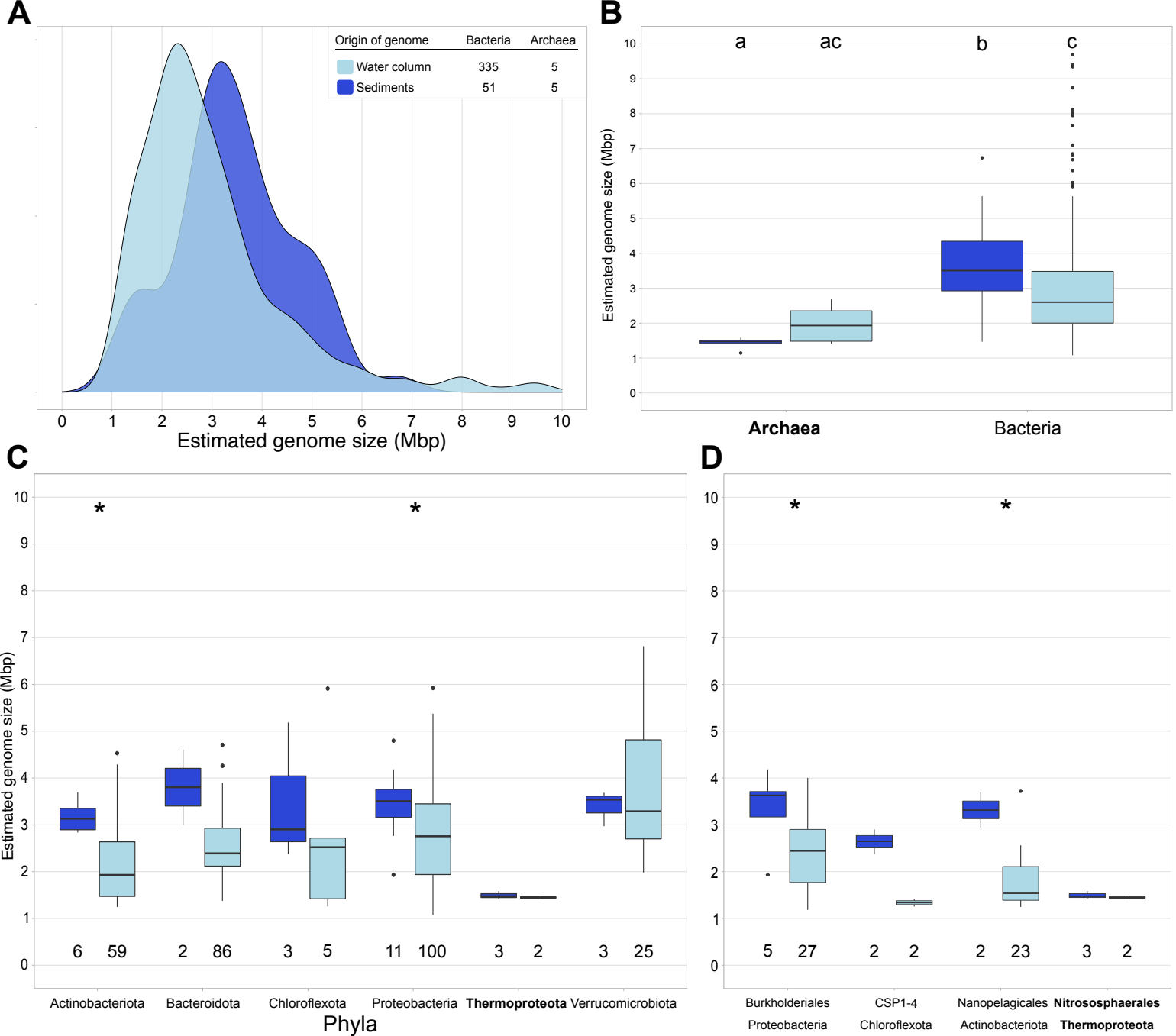
Average genome size (Mbp)

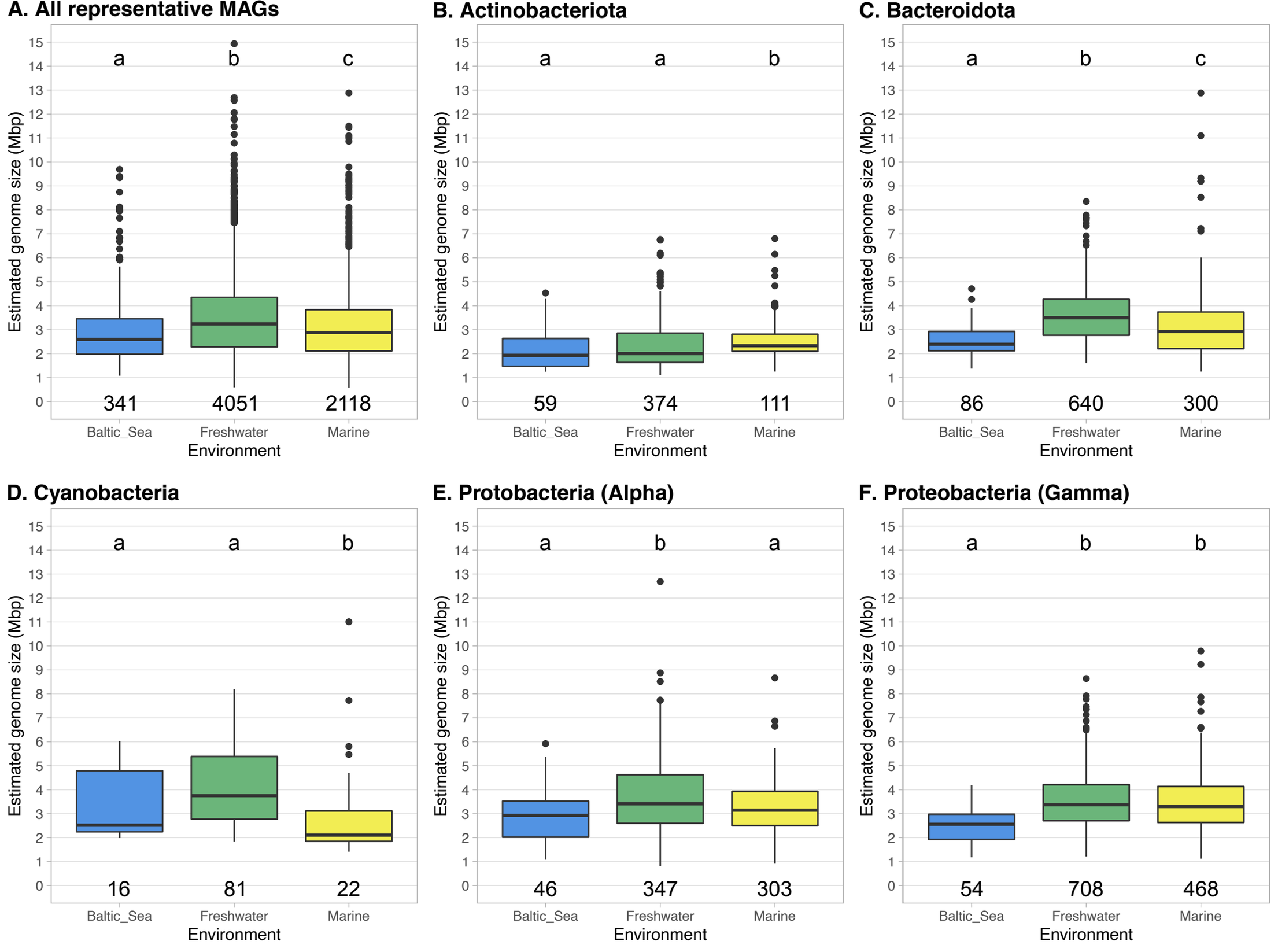
- 10
- 20
- 30

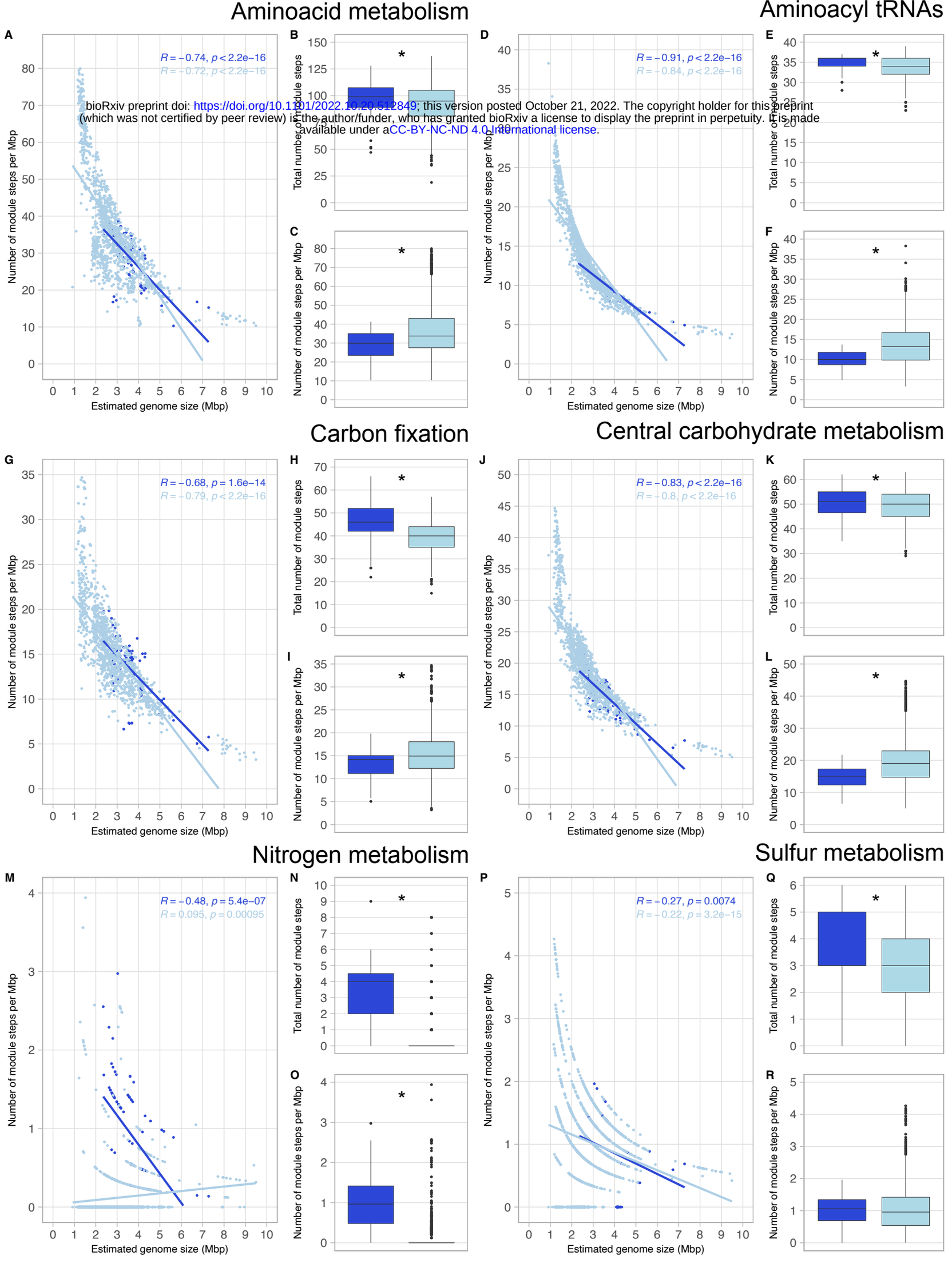
Environment

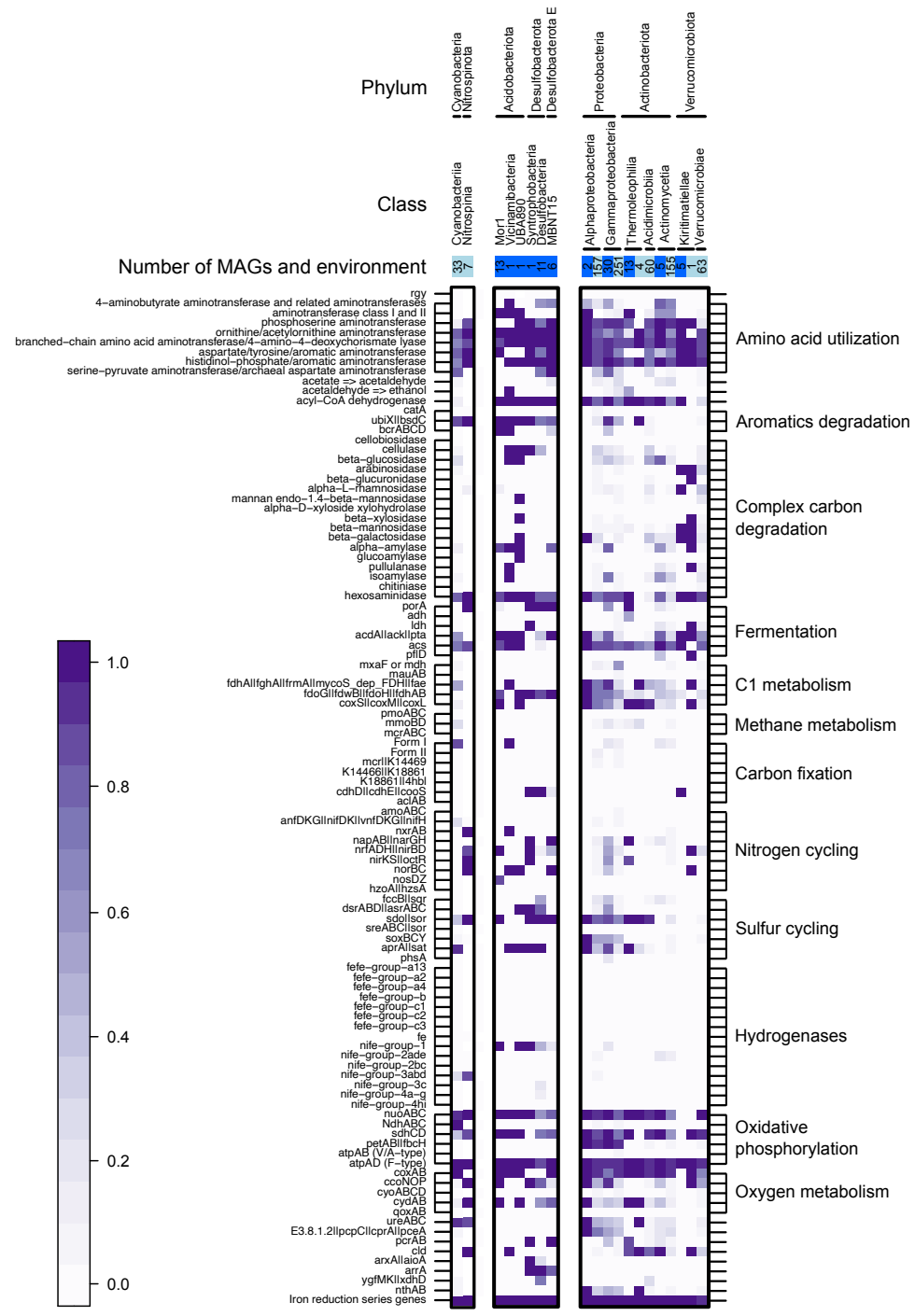
- Sediments
- Water column


A
B
*
*
Average genome size (Mbp)
0
2
4
6
8
10
12
14
16
18
108
69
14
94
Sediments
Water column
Environment
Oxygen < 2 mg/L
Oxygen > 2 mg/L
Oxygen concentration in Baltic sediments


A
Origin of genome
Bacteria
Archaea
Water column
335
5
Sediments
51
5
0 1 2 3 4 5 6 7 8 9 10
Estimated genome size (Mbp)
B
a
ac
b
c
Estimated genome size (Mbp)
0 1 2 3 4 5 6 7 8 9 10
Archaea
Bacteria
C
*
*
Estimated genome size (Mbp)
0 1 2 3 4 5 6 7 8 9 10
6 59 2 86 3 5 11 100 3 2 3 25
Actinobacteriota
Bacteroidota
Chloroflexota
Proteobacteria
Thermoproteota
Verrucomicrobiota
Phyla
D
*
*
0 1 2 3 4 5 6 7 8 9 10
5 27 2 2 2 23 3 2
Burkholderiales
Proteobacteria
CSP1-4
Chloroflexota
Nanopelagicales
Actinobacteriota
Nitrososphaerales
Thermoproteota


A. All representative MAGs
B. Actinobacteriota
C. Bacteroidota
D. Cyanobacteria
E. Protobacteria (Alpha)
F. Proteobacteria (Gamma)
Estimated genome size (Mbp)
Environment
Baltic_Sea
Freshwater
Marine
a
b
c
341
4051
2118
59
374
111
86
640
300
16
81
22
46
347
303
54
708
468




Phylum
Cyanobacteria
Nitrospinota
Acidobacteriota
Desulfobacterota
Desulfobacterota E
Proteobacteria
Actinobacteriota
Verrucomicrobiota
Class
Cyanobacteria
Nitrospinia
Mor1
UBA1800
Syntrophobacteria
Desulfobacteria
MBNT15
Alphaproteobacteria
Gammaproteobacteria
Thermoleophilia
Acidimicrobiia
Actinomycetia
Kiritimatiellae
Verrucomicrobiae
Number of MAGs and environment
33
7
13
1
1
11
6
27
157
301
251
13
4
60
155
5
63
rgy
4−aminobutyrate aminotransferase and related aminotransferases
aminotransferase class I and II
phosphoserine aminotransferase
ornithine/acetylornithine aminotransferase
branched−chain amino acid aminotransferase/4−amino−4−deoxychorismate lyase
aspartate/tyrosine/aromatic aminotransferase
histidinol−phosphate/aromatic aminotransferase
serine−pyruvate aminotransferase/archaeal aspartate aminotransferase
acetate => acetaldehyde
acetaldehyde => ethanol
acyl−CoA dehydrogenase
catA
ubiXIIbsdC
bcrABCD
cellobiosidase
cellulase
beta−glucosidase
arabinosidase
beta−glucuronidase
alpha−L−rhamnosidase
mannan endo−1,4−beta−mannosidase
alpha−D−xyloside xylohydrolase
beta−xylosidase
beta−mannosidase
beta−galactosidase
alpha−amylase
glucoamylase
pullulanase
isoamylase
chitinase
hexosaminidase
porA
adh
ldh
acdAIIackIIpta
acs
pflD
mxaF or mdh
mauAB
fdhAIIfghAIIfrmAIImycoS_dep_FDHIIfae
fdoGIIfdwBIIfdoHIIfdhAB
coxSIIcoxMIIcoxL
pmoABC
mmoBD
mcrABC
Form I
Form II
mcrIIK14469
K14466IIK18861
K18861II4hbl
cdhDIIcdhEIIcooS
aclAB
amoABC
anfDKGIInifDKIIvnfDKGIInifH
nxrAB
napABIInarGH
nrfADHInirBD
nirKSIIoqtR
norBC
nosDZ
hzoAIIhzsA
fccBIIsqr
dsrABDIIasrABC
sdoIIsor
sreABCIIsor
soxBCY
aprAIIsat
phsA
fefe−group−a13
fefe−group−a2
fefe−group−a4
fefe−group−b
fefe−group−c1
fefe−group−c2
fefe−group−c3
fe
nife−group−1
nife−group−2ade
nife−group−2bc
nife−group−3abd
nife−group−3c
nife−group−4a−g
nife−group−4hi
nuoABC
NdhABC
sdhCD
petABIIfbcH
atpAB (V/A−type)
atpAD (F−type)
coxAB
ccoNOP
cyoABCD
cydAB
qoxAB
ureABC
E3.8.1.2IIpcpCIIcprAIIpceA
pcrAB
cld
arxAIIaioA
arrA
ygfMKIIxdhD
nthAB
Iron reduction series genes
Amino acid utilization
Aromatics degradation
Complex carbon degradation
Fermentation
C1 metabolism
Methane metabolism
Carbon fixation
Nitrogen cycling
Sulfur cycling
Hydrogenases
Oxidative phosphorylation
Oxygen metabolism
1.0
0.8
0.6
0.4
0.2
0.0

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