Linking prokaryotic genome size variation to metabolic potential and environment

RUNNING TITLE: Genome size links to metabolic potential

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

ABSTRACT (250 words)

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.

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. 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 (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

102 capabilities in Bacteria in the Baltic Sea, and iii) which taxa and metabolic pathways contrast
103 between pelagic and benthic Bacteria. To achieve this, we selected 111 pelagic and 59 benthic
104 metagenomic samples that were previously published, and we provide one new and
105 unpublished benthic dataset with 49 metagenomic samples. We use these 219 metagenomes to
106 study genome size distribution across sediments and water column in the Baltic Sea (**Figure 1**
107 and **Table S1**). For this, we use two different approaches to study genome size: the estimated
108 average genome size (AGS) per metagenomic sample, and the estimated genome size of
109 bacterial and archaeal metagenome-assembled genomes (MAGs). Our results show that
110 Bacteria and Archaea with larger genomes in the sediments present lower coding density than
111 the smaller genomes in the water column. We also find that microorganisms inhabiting the
112 Baltic benthos have more metabolically-versatile genomes than pelagic prokaryotes, which
113 mean they code for a wider range of metabolic capabilities. Finally, we also find that functions
114 involved in the nitrogen metabolism are disproportionately more detected in benthic bacteria
115 in the Baltic Sea.
116
117 RESULTS AND DISCUSSION
118
119 **While depth is significantly associated with genome size in benthic and pelagic**
120 **metagenomes, salinity is only correlated to genome size in benthic metagenomes.** First, we
121 calculated the average genome size (AGS) of metagenomes across the latitudinal gradient of
122 the Baltic Sea comparing benthic and pelagic metagenomes. We observed that sediment-
123 dwelling microbial communities present significantly larger AGS (mean = 6.01 Mbp, n = 108)
124 than pelagic communities (mean AGS = 5.40 Mbp, n = 69) (Wilcoxon test, $p < 0.01$) (**Figure**
125 **2A**). We then evaluated the relationships between AGS of metagenomes and each of the
126 environmental variables (depth, salinity, temperature, and oxygen concentration)
127 independently and their interactions (ANOVA type II). The AGS in the pelagic metagenomes
128 is significantly associated only with depth and shows a negative correlation (**Supplementary**
129 **Material 1** and **Figure S1**). Previous marine analysis have found the opposite effect, bigger
130 genome sizes in deeper areas (19). However, our pelagic analysis only covers 5 meters of depth
131 and explains 19% of the genome size variation in pelagic metagenomes. AGS in sediment
132 metagenomes is significantly associated with both salinity and depth and most of the
133 interactions of these variables (**Supplementary Material 1**). Both depth and salinity have a
134 weak positive correlation with AGS in the sediment samples (**Figure S1**). To our knowledge,

135 this is the first time where a positive correlation between genome size and salinity is reported
136 for brackish sediment microorganisms.
137
138 Although the ANOVA did not show a significant effect of water oxygen concentration
139 on genome size, we further investigate the effect of bottom water O_2 concentration in the AGS
140 of metagenomes from sediments. We separated these metagenomes into those from oxygen
141 concentration 0 to 2 mg/L (mostly metagenomes from the dead zone) and those metagenomes
142 with oxygen concentration 2 to 12.45 mg/L. We observe that benthic metagenomes from lower
143 oxygen concentration (mean AGS = 7.08 Mbp, n = 14) present significantly bigger genome
144 sizes than those in sediments with higher oxygen concentration (mean AGS = 5.85 Mbp, n =
145 94) (Wilcoxon test, $p < 0.01$) (**Figure 2B**). Complementarily, previous results show that the
146 dead zone bacteria tend to be more metabolically similar to each other when compared to
147 bacteria from oxic sediments (29). Our observation confirms a recent study that observed
148 obligate anaerobes from diverse environments present bigger genome sizes than
149 microaerophilic microorganisms (30). Moreover, previous studies show a positive correlation
150 between genome size and nutrient concentration (19,31,32). Dead zones in the Baltic Sea are
151 characterized by anthropogenic eutrophication (33), which would also promote bigger genome
152 sizes. Altogether, these results indicate that high nutrient concentration and low oxygen
153 concentrations in Baltic Sea dead zones may select for prokaryotes with bigger genome sizes.
154
155 **While average benthic estimated genome size is bigger than pelagic, the biggest**
156 **genomes in the Baltic Sea were found in the water column.** To compare genome size
157 between pelagic and benthic bacteria and archaea, we used metagenome-assembled genomes
158 (MAGs; >75% completeness and <5% contamination) from our metagenome datasets. Our
159 dataset compiles 216 MAGs from the sediments that dereplicated into 56 representative
160 genomospecies (95% average nucleotide identity). Additionally, 1920 pelagic MAGs were
161 dereplicated into 340 representative genomospecies. We observe that 12 phyla were detected
162 in both sediments and water column, while 16 phyla were specific to either habitat. Seven phyla
163 were found specific to the sediment (14 representative MAGs) and nine phyla were specific to
164 the water column (29 representative MAGs) (**Table 1**). Interestingly, only one genomospecies
165 representative was binned from both sediments and water column. This genome representative
166 belongs to genus *Mycobacterium* (phylum Actinobacteriota, mOTU_124/pelagic and
167 mOTU_027/sediments in **Table S2**), a genus that is not commonly found on brackish surface
168 waters (34). However, this genus was found to be abundant in sediments in some regions of

169 the Baltic Sea, especially in anoxic areas close to Landsort (35). These differences on
170 taxonomical composition in microbial communities between water column and sediments
171 altogether with latitudinal changes in microbial biodiversity in the Baltic Sea (34) show how
172 heterogeneous is the microbial composition of brackish environments.
173
174 Additionally, the eight largest representative genomes of the dataset were observed in
175 the water column and belonged to phylum Planctomycetota (family Planctomycetaceae) (7.95
176 – 9.69 Mbp). It has been previously observed that Planctomycetota is the phylum containing
177 the aquatic MAG with the biggest known estimated genome size (14.93 Mbp) (2). These large
178 genomes contain large collections of genes that could be linked to the very complex cell
179 structures and chromosomes observed in this phylum (36). Still, further research is necessary
180 to understand if extant genome size in phylum Plactomycetota is the result of ecological
181 adaptation to abiotic gradients. On the other side of the genome size spectrum, the
182 representative MAG with the smallest estimated genome size belongs to class
183 Alphaproteobacteria (family Pelagibacteraceae, 1.08 Mbp; **Table 1**). Bacteria from this family
184 have been widely reported to be streamlined, abundant and ubiquitous across all salinity
185 gradients (37,38).
186
187 Altogether, representative MAGs from sediments presented an average estimated
188 genome size of 3.47 Mbp, which was significantly higher than for water column MAGs (2.96
189 Mbp) ($p < 0.01$) (**Figure 3A**). This was true also at the phyla, class, and order level (**Figure**
190 **3C** and **D**). Bacteria from sediments presented bigger estimated genome size on average (3.67
191 Mbp), followed by pelagic Bacteria (2.98 Mbp), pelagic Archaea (1.97 Mbp) and sediment
192 Archaea (1.43 Mbp) (**Figure 3B**). This is supported by previous results, as Bacteria show
193 bigger genome sizes than Archaea regardless of the environment (2). Moreover, the average
194 estimated genome size of Baltic sediments (3.47 Mbp) is more similar to that of terrestrial
195 microbial genomes (3.7 Mbp) (2). Additionally, genomes in the Baltic sediments have also
196 lower coding density than pelagic (**Figure S2**). Our results corroborate previous findings that
197 streamlining is common in pelagic marine environments (7,37,39,40) and pelagic brackish
198 environments (23).
199
200 In our study, the average estimated AGS for the sediments (6.01 Mbp) and water
201 column (4.44 Mbp) is larger than the average estimated genome size of the MAGs assembled
202 and binned from the sediments (3.47 Mbp) and water column (2.96 Mbp), respectively (**Figure**

203 **2A** and **3A**). We calculated the AGS to estimate the average genome size of the whole
204 microbial community. We used MicrobeCensus as a robust and accurate tool to calculate AGS
205 (41). However, this AGS of metagenomes overestimates the genome size because of the viral
206 and eukaryotic reads that might be present in the sample (42,43). On the other hand, there are
207 two biases with looking at the average estimated genome size of MAGs. One, assembly and
208 binning biases make a MAG an average 3.7% smaller than genomes from isolates in the same
209 genomospecies (2). Second, a bias of overlooking all those bacteria and archaea that are hard
210 to assemble and bin due to high genomic intrapopulation diversity (44). For these reasons, in
211 our study, we use two methods that have different biases to answer the same question; how is
212 the genome size of microorganisms distributed in the Baltic Sea. Irrespective of the method
213 used, the average estimated genome size in sediments is larger than in the water column.
214
215 The available metadata for benthic and pelagic carbon concentration was not
216 comparable due to methods and metrics used for analysis. This made it not possible to look for
217 a clear ecological link between bigger genome size in benthic zones and nutrient concentration
218 (**Table S1**). Luckily a previous study has compiled information on organic carbon stocks in the
219 Baltic Sea (45): by collecting information from many different studies and years, they show
220 that in average the top 10 cm of sediments contain between 2 and 4 times more organic carbon
221 per area than the water column in the Baltic Sea. More organic carbon available for bacteria
222 and archaea would also pose a lower pressure on the genome to streamline. From the results of
223 our study, we hypothesize that one of the reasons pelagic microbial genomes are smaller than
224 benthic microbial genomes is the difference in organic carbon availability. Altogether, a
225 positive correlation between genome size and nutrient concentration has been shown before
226 (19,31,32).
227
228 **Brackish pelagic microorganisms tend to show smaller genome sizes than marine**
229 **and freshwater**. In further analysis of genome size variation across salinity gradients, we
230 compared the average estimated genome size of pelagic Baltic Sea MAGs to previously
231 published genome size estimations (2). This comparison includes all taxonomic groups found
232 in three large MAG-datasets (17,46,47) (completeness >75%) that includes 4051 freshwater
233 representative MAGs, 2118 marine representative MAGs and 340 pelagic representative
234 MAGs from the Baltic Sea. We found that the average estimated genome size of the brackish
235 pelagic MAGs (2.96 Mbp) is lower than in marine MAGs (3.10 Mbp) (Kruskal-Wallis test, p
236 < 0.01). Furthermore, we observed the largest average estimated genome size in freshwater

237 MAGs (3.48 Mbp) (Kruskal-Wallis test, $p < 0.01$) (**Figure 4A**). These observed differences in
238 genome sizes across different pelagic environments together with the previously observed
239 differential functions (48) suggests that genome size has a potential signature across aquatic
240 environments.
241
242 We then divided all MAGs into phyla and focus on the most common from aquatic
243 environments: phyla Actinobacteriota, Bacteroidota, Cyanobacteria and Proteobacteria
244 (classes Alpha and Gammaproteobacteria) (**Figure 4B-F**). We test if the genome size
245 differences are consistent across phyla. We observe that only phylum Bacteroidota follows the
246 same trend as the full dataset in the average estimated genome size (**Figure 4C**). On the other
247 hand, MAGs from phylum Actinobacteriota present the biggest genome sizes for marine
248 environments, while no difference on average genome size is observed between freshwater and
249 brackish (**Figure 4B**). This result updates previous observations on aquatic Actinobacteriota
250 genome size variation (23). Opposite to Actinobacteriota, MAGs from phylum Cyanobacteria
251 show the smallest average genome size for marine environments, while we do not observe
252 statistical differences between brackish and freshwater MAGs (**Figure 4D**). Similar trends
253 were observed for isolates and MAGs of picocyanobacteria (22,49). However, it is important
254 to remark that DNA extraction, assembly, binning and/or quality check of aquatic
255 cyanobacterial MAGs is still a big challenge that needs to be addressed (**Supplementary**
256 **Material 2**) (50,51). All in all, these results hint at the complicated ecological role of genome
257 size in pelagic bacterial groups, where environment (2,15,16), lifestyle (4,13,14) and taxonomy
258 (11,12) are intertwined.
259
260 **Smaller bacterial genomes in the Baltic Sea tend to lack certain functional**
261 **categories.** We selected the bacterial MAGs with >90% completeness for metabolic
262 annotation and analyze which functional categories correlate with genome size in brackish
263 sediments and water column (**Figure 5 A-R** and **Supplementary Material 3**). Functional
264 categories include different but related metabolic pathways (KEGG modules), and each module
265 comprises multiple specific metabolic reactions (module steps) (52). In our results, we observe
266 patterns of negative correlation between estimated genome size and number of module steps
267 per Mbp in most of the functional categories analyzed (**Figure 5**). For example, given the core
268 functions amino acid metabolism, aminoacyl tRNAs and central carbohydrate metabolism, we
269 observe a higher number of module steps per Mbp at lower genome sizes in both environments
270 both in pelagic and benthic MAGs (**Figures 5A, D** and **J**). However, genome size does not

271 explain the number of module steps per Mbp in all functional categories and metabolisms,
272 especially in the case of non-essential functions such as drug resistance and transport systems
273 (**Supplementary Material 3F** and **N**). These results suggest that streamlining of genomes
274 select for specific functions and not the whole genome. This could be explained by the Black
275 Queen Hypothesis (53); when the fitness cost of a function is higher than its benefit, microbes
276 might lose it and, instead, obtain benefit from leaky metabolites from neighboring cells,
277 establishing interdependencies. The loss of those specific functions in the Baltic Sea might
278 have consequential long-lasting metabolic partnerships within the community.
279
280 When considering the total number of module steps, we observe that genomes of
281 benthic bacteria code for a higher number of module steps than pelagic bacteria in amino acid
282 metabolism, aminoacyl tRNAs, carbon fixation, nitrogen and sulfur metabolism (**Figures 5B,**
283 **E, H, N** and **Q**) (Wilcoxon test, $p < 0.01$). This could be the result of a confounding effect of
284 genome size: microorganisms from sediments have bigger genome sizes than those in the water
285 column (**Figure 2**) and therefore, code for a higher number of functions. Hence, the total
286 number of module steps per Mbp in all six categories was analyzed. If streamlining affects all
287 functional categories similarly, coding density trends would be similar for all functional
288 categories. Indeed, we observe that pelagic bacteria present a greater number of module steps
289 per Mbp than sediment bacteria in amino acid metabolism (**Figure 5C**), aminoacyl tRNAs
290 (**Figure 5F**), carbon fixation (**Figure 5I**) and central carbohydrate metabolism (**Figure 5L**).
291 However, nitrogen metabolism shows the opposite trend: sediment bacteria have a greater
292 number of module steps per Mbp than water column bacteria (**Figure 5O)**. Literature report
293 that Baltic sediments contain about 95% of the total pool of nitrogen while the water column
294 only 5%, hence the water column only carries a small part of the overall nitrogen cycle (54).
295 As mentioned above for carbon, a higher availability of resources, including nitrogen-derived
296 compounds, would also imply a lower evolutionary pressure for streamlined genome sizes in
297 Baltic sediments. These results indicate that benthic bacteria potentially play a bigger role than
298 pelagic bacteria in nitrogen cycling of autotropic systems like the Baltic Sea (55,56).
299
300 Although most of the MAGs presented at least 1 module step related to drug resistance
301 (96.97% sediment MAGs and 98.81% pelagic MAGs), bacteria from sediments presented a
302 higher number of antibiotic resistance module steps than pelagic bacteria. This applies both to
303 total number of module steps and the number of module steps per Mbp (**Supplementary**
304 **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 (**Figure 6**).

We observed the presence of the genes *cdhH* | *cdhE* | *cooS* from the Wood-Ljungdhal pathway exclusively in marine sediments (**Figure 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 (**Figure 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, sulphate 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).

338 **Conclusions**. In this research article, we provide a comprehensive analysis to investigate
339 the ecological implications of microbial genome in the Baltic Sea. We show that genome size
340 in Bacteria and Archaea is linked to the environment (**Figures 2**, **3, 4** and **S1** and
341 **Supplementary Material 1**), taxonomic identity (**Table 1** and **Figure 3**) and metabolic
342 potential (**Figure 5** and **Supplementary Material 3**). We also provide some insights on how
343 distinct pelagic and benthic microbial communities in the Baltic Sea are: not only microbial
344 MAGs retrieved from these two environments are different in taxonomy (**Table 1**), but also in
345 genome size (**Figure 3**) and metabolism (**Figures 5** and **6**). This highlights water bodies are
346 highly heterogeneous biomes, with highly distinct microbial communities between micro-
347 niches. With the continuous progression of aquatic microbial ecology and the development of
348 new isolation, omics and bioinformatic techniques, future research should provide a more
349 complete and unbiased view of genome sizes distribution in nature and its ecological
350 implication in microbial populations.
351
352 MATERIAL AND METHODS
353
354 **Baltic Sea metagenomes collection**. For this study we compiled new and public Baltic
355 Sea metagenomes from the water column and the sediments. The final dataset consisted of 219
356 metagenome samples from a wide range of locations in the Baltic Sea that include 5
357 independent datasets (**Table S1** and **Figure 1**). We compiled 108 sediment metagenomes, of
358 which 59 were collected in 2019 and recently published (29), and we collected 49 metagenomes
359 from 2016 to 2018. The pelagic dataset consists of 118 pelagic samples collected from 2011 to
360 2015 published in three different studies (46,64,65). All five datasets have abiotic metadata of
361 depth (m), salinity (PSU), temperature (C) and oxygen concentration (mg/L) (**Table S1**).
362
363 **Environmental sampling.** The top 2 cm of sediment was collected at soft bottom clay-
364 muddy habitats from 59 stations from north to south in the Baltic Sea in 2019, following the
365 sampling described in Broman et al., (2022) (**Table S1** for coordinates). Briefly, one sediment
366 core was collected per station using a Kajak gravity corer (surface area: 50 cm^2, one core per
367 station) and the top 0–2 cm layer was sliced into a 215 ml polypropylene container
368 (207.0215PP, Noax Lab, Sweden). The sediment was homogenized and stored at -20°C on the
369 boat, kept on an iced cooler without thawing for ~2 h during transportation to the university,
370 and finally stored again at -20 °C until DNA extraction. Bottom water (~20 cm above the
371 sediment surface) was collected at each station with a Niskin bottle. This was followed by on

372 deck measurements of bottom water temperature, salinity, and dissolved O2 using a portable
373 multimeter (HQ40D, Hach).
374
375 **DNA extraction and sequencing.** The sediment samples were thawed, homogenized,
376 and a subsample of 0.25 g was used for DNA extraction using the DNeasy PowerSoil kit
377 (Qiagen) according to the manufacturer's protocol. The quantity and quality of eluted DNA
378 were measured using NanoDrop One spectrophotometer and Qubit 2 (both by ThermoFisher
379 Scientific) to ensure that samples meet the minimum requirements for sequencing. The samples
380 were then sequenced at the Science for Life Laboratories facility on one NovaSeq 6000 S4
381 lanes using a 2 × 150 bp setup. Sequencing yielded on average 53.0 million reads per sample.
382
383 **MAGs collection**. Assembling and binning of the 108 sediment metagenomes resulted
384 in 2248 bins. To obtain bins from metagenomes, we followed the 0053_metasssnake2 pipeline
385 (https://github.com/moritzbuck/0053_metasssnake2) (v0.0.2). In this pipeline we used
386 Sourmash (66) to create signatures, Megahit (67) to obtain single-sample assemblies and
387 Metabat2 (68) for the binning of the assemblies. We used default parameters throughout the
388 whole pipeline. Quality of the bins was assessed using CheckM (v1.1.3) (69): we used the
389 typical workflow (*lineage_wf*) with default parameters, and selected only bins with quality of
390 completeness >75% and contamination <5%. From all the bins, only 216 passed our quality
391 threshold and we named those MAGs (metagenome-assembled genomes). All MAGs were
392 taxonomically classified using GTDB-tk (v1.5.0) (70) according to the GTDB classification
393 (data version r202) (71). The quality of the MAGs belonging to phyla Actinobacteriota and
394 Patescibacteria were assessed separately using a custom set of marker genes. Preliminary
395 quality check of Actinobacteriota genomes in a publicly available freshwater dataset (47) show
396 that default parameters underestimate the quality of the MAGs that are classified as
397 Actinobacteriota compared to using a custom marker gene set (**Supplementary Material 2**).
398 We obtained the custom marker gene set using Ac1 circular genomes (72,73). In the case of
399 phylum Patescibacteria, we used a custom set of maker genes provided by CheckM (69,74).
400
401 Complementarily, we used 1920 pelagic MAGs that were published (46) and passed
402 the >75% completeness and <5% contamination threshold. All high-quality MAGs from the
403 sediments and water column were dereplicated using fastANI (95% ANI threshold as estimator
404 of genetic unit) and mOTUlizer (v0.3.2) (75,76). From the 216 with >75% completeness
405 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) (77). 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) (78). 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 (**Figures 2, 3** and **5** and **Supplementary Material 3**). Asterisks 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 (79) (**Supplementary Material 1**). We obtained the correlation coefficients on scatterplots

440 (**Figures 5** and **S1**) to test the fit of our data to linear regressions using the function *stat_cor*
441 from the R package *ggpubr* v0.4.0 (80).
442
443 **Data availability**. All metagenome datasets are available in public repositories under
444 NCBI project accession number SRP077551 and ENA accession numbers PRJEB34883,
445 PRJEB22997 and PRJEB41834. Specific accession numbers of all metagenomes are available
446 in **Table S1**. Pelagic MAGs can be found on project PRJEB34883 and benthic MAGs on
447 project PRJNA891615. Assembly and binning of the dataset provided in this paper used scripts
448 available at https://github.com/moritzbuck/0053_metasssnake2. Supplemental material can
449 also be accessed 10.17044/scilifelab.21378294.
450
451 ACKNOWLEDGEMENTS
452
453 We thank Elias Broman for his help on the acquisition of the benthic dataset of
454 metagenomes and metadata and Ola Svensson, Caroline Raymond and Jonas Gunnarsson for
455 assistance during sampling. We also thank Juanita Gutiérrez-Valencia for her suggestions in
456 statistical analysis.
457
458 This work was supported by SciLifeLab. The authors acknowledge support from the
459 National Genomics Infrastructure in Stockholm funded by Science for Life Laboratory, the
460 Knut and Alice Wallenberg Foundation and the Swedish Research Council, and SNIC/Uppsala
461 Multidisciplinary Center for Advanced Computational Science for assistance with massively
462 parallel sequencing and access to the UPPMAX computational infrastructure. Computational
463 work and data handling were enabled by resources in the projects SNIC 2020/5-159, 2021/5-
464 133, 2022/5-137, 2020-6-60 and 2022/6-77 provided by the Swedish National Infrastructure
465 for Computing (SNIC) at UPPMAX, partially funded by the Swedish Research Council
466 through grant agreement no. 2018-05973.
467
468 AR-G and SLG conceptualized and designed research project. AR-G, MB and SLG
469 refined the project idea. AR-G, AFA and FJAN participated in data collection. AR-G and DIS
470 did the molecular work. AR-G and MB performed bioinformatic analysis to obtain MAGs from
471 sediment raw sequences. AR-G and SLG performed data analysis. AR-G and SLG drafted the
472 first manuscript. AR-G, DIS and SLG did literature searches. All authors contributed to the
473 writing and editing of the manuscript.

CONFLICTS OF INTEREST

The authors declare that there are no competing interests.

REFERENCES

1. Kirchberger PC, Schmidt ML, Ochman H. The Ingenuity of Bacterial Genomes. Annu Rev Microbiol. 2020 Sep 8;74(1):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. Frontiers in Microbiology. 2022 Jan 5;12(761869).

3. Lynch M. Streamlining and Simplification of Microbial Genome Architecture. 2006;25.

4. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. ISME J. 2014 Aug;8(8):1553–65.

5. Kuo CH, Moran NA, Ochman H. The consequences of genetic drift for bacterial genome complexity. Genome Research. 2009 Aug 1;19(8):1450–4.

6. Wolf YI, Koonin EV. Genome reduction as the dominant mode of evolution. BioEssays. 2013 Sep;35(9):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 Dec;12(12):841–50.

8. Bobay LM, Ochman H. Factors driving effective population size and pan-genome evolution in bacteria. BMC Evol Biol. 2018 Dec;18(1):153.

9. Konstantinidis KT, Tiedje JM. Trends between gene content and genome size in prokaryotic species with larger genomes. Proceedings of the National Academy of Sciences. 2004 Mar 2;101(9):3160–5.

10. Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. 2014;19.

11. Biller SJ, Berube PM, Lindell D, Chisholm SW. Prochlorococcus: the structure and function of collective diversity. Nat Rev Microbiol. 2015 Jan;13(1):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. Casadesús J, editor. PLoS Genet. 2022 May 23;18(5):e1010220.

13. Moran NA, Mira A. The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. Genome Biol. 2001;2(12):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 Jan 21;100(2):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 [Internet]. 2020 Feb 11 [cited 2020 Apr 1]; Available from: http://www.nature.com/articles/s41396-020-0600-z

16. Simonsen AK. Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria. ISME J. 2022 Feb;16(2):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 Apr;39(4):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 Aug 15;75(16):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 Oct;2(10):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 Nov;13(11):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 Jan;15(1):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 Aug 8;20(1):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 Dec;16(1):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 Jan 13;601(7892):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 Aug;14(8):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 Dec;4(1):604.

27. Louca S, Parfrey LW, Doebeli M. Decoupling function and taxonomy in the global ocean microbiome. Science. 2016 Sep 16;353(6305):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 Dec;8(1):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 Aug 15;10(1):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 Apr;130(4):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 Jul;6(7):1403–14.

32. Aylward FO, Santoro AE. Heterotrophic Thaumarchaea with Small Genomes Are Widespread in the Dark Ocean. Bouskill N, editor. mSystems. 2020 Jun 16;5(3):e00415-20, /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 May 15;43(10):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 Oct;5(10):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 Sep 23;8(9):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 May 4;12:643045.

37. Giovannoni SJ. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. Annu Rev Mar Sci. 2017 Jan 3;9(1):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 [Internet]. 2023 Feb 4 [cited 2023 Feb 6]; Available from: https://www.nature.com/articles/s41396-023-01376-2

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. Proceedings of the National Academy of Sciences. 2013 Jul 9;110(28):11463–8.

40. Giovannoni SJ. Genome Streamlining in a Cosmopolitan Oceanic Bacterium. Science. 2005 Aug 19;309(5738):1242–5.

589 41. Pereira-Flores E, Glöckner FO, Fernandez-Guerra A. Fast and accurate average
590 genome size and 16S rRNA gene average copy number computation in metagenomic
591 data. BMC Bioinformatics. 2019 Dec;20(1):453.

592 42. Kristensen DM, Mushegian AR, Dolja VV, Koonin EV. New dimensions of the virus world
593 discovered through metagenomics. Trends in Microbiology. 2010 Jan;18(1):11–9.

594 43. Lind AL, Pollard KS. Accurate and sensitive detection of microbial eukaryotes from
595 whole metagenome shotgun sequencing. Microbiome. 2021 Dec;9(1):58.

596 44. Ghurye JS. Metagenomic Assembly: Overview, Challenges and Applications. 2016;10.

597 45. Scheffold MIE, Hense I. Quantifying Contemporary Organic Carbon Stocks of the Baltic
598 Sea Ecosystem. Front Mar Sci. 2020 Dec 17;7:571956.

599 46. Alneberg J, Bennke C, Beier S, Bunse C, Quince C, Ininbergs K, et al. Ecosystem-wide
600 metagenomic binning enables prediction of ecological niches from genomes. Commun
601 Biol. 2020 Dec;3(1):119.

602 47. Buck M, Garcia SL, Fernandez L, Martin G, Martinez-Rodriguez GA, Saarenheimo J, et
603 al. Comprehensive dataset of shotgun metagenomes from oxygen stratified freshwater
604 lakes and ponds. Sci Data. 2021 Dec;8(1):131.

605 48. Jurdzinski KT, Mehrshad M, Delgado LF, Deng Z, Bertilsson S, Andersson AF. Large-
606 scale phylogenomics of aquatic bacteria reveal molecular mechanisms for adaptation to
607 salinity [Internet]. Microbiology; 2022 Oct [cited 2023 Jan 17]. Available from:
608 http://biorxiv.org/lookup/doi/10.1101/2022.10.03.510577

609 49. Sánchez-Baracaldo P, Bianchini G, Di Cesare A, Callieri C, Chrismas NAM. Insights Into
610 the Evolution of Picocyanobacteria and Phycoerythrin Genes (mpeBA and cpeBA). Front
611 Microbiol. 2019 Jan 30;10:45.

612 50. Alvarenga DO, Fiore MF, Varani AM. A Metagenomic Approach to Cyanobacterial
613 Genomics. Front Microbiol. 2017 May 9;8:809.

614 51. Kim Tiam S, Comte K, Dalle C, Duval C, Pancrace C, Gugger M, et al. Development of a
615 new extraction method based on high-intensity ultra-sonication to study RNA regulation
616 of the filamentous cyanobacteria Planktothrix. Kalendar R, editor. PLoS ONE. 2019 Sep
617 6;14(9):e0222029.

618 52. Zhou Z, Tran PQ, Breister AM, Liu Y, Kieft K, Cowley ES, et al. METABOLIC: high-
619 throughput profiling of microbial genomes for functional traits, metabolism,
620 biogeochemistry, and community-scale functional networks. Microbiome. 2022
621 Dec;10(1):33.

622 53. Morris JJ, Lenski RE, Zinser ER. The Black Queen Hypothesis: Evolution of
623 Dependencies through Adaptive Gene Loss. mBio. 2012 Mar 23;3(2):e00036-12.

624 54. Lønborg C, Markager S. Nitrogen in the Baltic Sea: Long-term trends, a budget and
625 decadal time lags in responses to declining inputs. Estuarine, Coastal and Shelf
626 Science. 2021 Oct;261:107529.

627 55. Albert S, Bonaglia S, Stjärnkvist N, Winder M, Thamdrup B, Nascimento FJA. Influence
628 of settling organic matter quantity and quality on benthic nitrogen cycling. Limnol
629 Oceanogr. 2021 May;66(5):1882–95.

630 56. Griffiths JR, Kadin M, Nascimento FJA, Tamelander T, Törnroos A, Bonaglia S, et al.
631 The importance of benthic-pelagic coupling for marine ecosystem functioning in a
632 changing world. Glob Change Biol. 2017 Jun;23(6):2179–96.

633 57. Guo X pan, Zhao S, Chen Y ru, Yang J, Hou L jun, Liu M, et al. Antibiotic resistance
634 genes in sediments of the Yangtze Estuary: From 2007 to 2019. Science of The Total
635 Environment. 2020 Nov;744:140713.

636 58. Marti E, Variatza E, Balcazar JL. The role of aquatic ecosystems as reservoirs of
637 antibiotic resistance. Trends in Microbiology. 2014 Jan;22(1):36–41.

638 59. Esposito A, Tamburini S, Triboli L, Ambrosino L, Chiusano ML, Jousson O. Insights into
639 the genome structure of four acetogenic bacteria with specific reference to the Wood–
640 Ljungdahl pathway. MicrobiologyOpen [Internet]. 2019 Dec [cited 2022 Aug 26];8(12).
641 Available from: https://onlinelibrary.wiley.com/doi/10.1002/mbo3.938

642 60. Lever MA. Acetogenesis in the Energy-Starved Deep Biosphere – A Paradox? Front
643 Microbio [Internet]. 2012 Jan [cited 2022 Aug 26];2(284). Available from:
644 http://journal.frontiersin.org/article/10.3389/fmicb.2011.00284/abstract

645 61. Marshall IPG, Karst SM, Nielsen PH, Jørgensen BB. Metagenomes from deep Baltic
646 Sea sediments reveal how past and present environmental conditions determine
647 microbial community composition. Marine Genomics. 2018 Feb;37:58–68.

648 62. Peters JW, Schut GJ, Boyd ES, Mulder DW, Shepard EM, Broderick JB, et al. [FeFe]-
649 and [NiFe]-hydrogenase diversity, mechanism, and maturation. Biochimica et Biophysica
650 Acta (BBA) - Molecular Cell Research. 2015 Jun;1853(6):1350–69.

651 63. Sinkko H. Sediment bacterial communities in nutrient cycling and in the history of the
652 Baltic Sea [Doctoral dissertation]. [Helsinki, Finland]: University of Helsinki; 2013.

653 64. Larsson J, Celepli N, Ininbergs K, Dupont CL, Yooseph S, Bergman B, et al.
654 Picocyanobacteria containing a novel pigment gene cluster dominate the brackish water
655 Baltic Sea. ISME J. 2014 Sep;8(9):1892–903.

656 65. Alneberg J, Sundh J, Bennke C, Beier S, Lundin D, Hugerth LW, et al. BARM and
657 BalticMicrobeDB, a reference metagenome and interface to meta-omic data for the
658 Baltic Sea. Sci Data. 2018 Dec;5(1):180146.

659 66. Titus Brown C, Irber L. sourmash: a library for MinHash sketching of DNA. JOSS. 2016
660 Sep 14;1(5):27.

661 67. Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, et al. MEGAHIT v1.0: A fast and
662 scalable metagenome assembler driven by advanced methodologies and community
663 practices. Methods. 2016 Jun;102:3–11.

664 68. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive
665 binning algorithm for robust and efficient genome reconstruction from metagenome
666 assemblies. PeerJ. 2019 Jul 26;7:e7359.

667 69. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the
668 quality of microbial genomes recovered from isolates, single cells, and metagenomes.
669 Genome Res. 2015 Jul;25(7):1043–55.

70. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Hancock J, editor. Bioinformatics. 2019 Nov 15;btz848.

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 Sep 1;38(9):1079–86.

72. Neuenschwander SM, Ghai R, Pernthaler J, Salcher MM. Microdiversification in genome-streamlined ubiquitous freshwater Actinobacteria. ISME J. 2018 Jan;12(1):185–98.

73. Kang I, Kim S, Islam MdR, Cho JC. The first complete genome sequences of the acI lineage, the most abundant freshwater Actinobacteria, obtained by whole-genome-amplification of dilution-to-extinction cultures. Sci Rep. 2017 Sep 27;7(1):42252.

74. 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 Jul;523(7559):208–11.

75. Buck M, Mehrshad M, Bertilsson S. mOTUpan: a robust Bayesian approach to leverage metagenome-assembled genomes for core-genome estimation. NAR Genomics and Bioinformatics. 2022 Jul 9;4(3):lqac060.

76. 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 Dec;9(1):5114.

77. 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 Dec;16(1):51.

78. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010 Dec;11(1):119.

79. R Core Team. R: A language and environment for statistical computing. [Internet]. R Foundation for Statistical Computing, Vienna, Austria; 2020. Available from: https://www.R-project.org/.

80. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. [Internet]. 2020. Available from: https://CRAN.R-project.org/package=ggpubr

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 **Table S1**. Shape type indicates the reference and shape size indicates the AGS of the given metagenome.

710

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

716

717 **Figure 3**. Overview of the estimated genome size in bacteria and archaea obtained from Baltic
718 Sea sediments (dark blue) and water column (light blue) using only the 397 representative
719 MAGs (95% average nucleotide identity) with >75% completeness. **Panel A** shows the
720 genome size distribution of archaea and bacteria obtained from Baltic water column and
721 sediments for a total of 397 representative genomes. **Panel B** shows the estimated genome size
722 per domain and environment. Different letters indicate significant differences $p < 0.05$
723 (Kruskal-Wallis non-parametric test; multiple testing corrected with Benjamini-Hochberg).
724 **Panel C** shows the estimated genome size per phylum. We selected only phyla with at least 2
725 MAGs in each environment. **Panel D** shows the estimated genome size per order. We selected
726 only orders with at least 2 MAGs in each environment. Numbers below the boxes indicate the
727 number of MAGs per environment. Stars in Panel C and D indicate significant differences $p <$
728 0.05 (Wilcoxon non-parametric test).

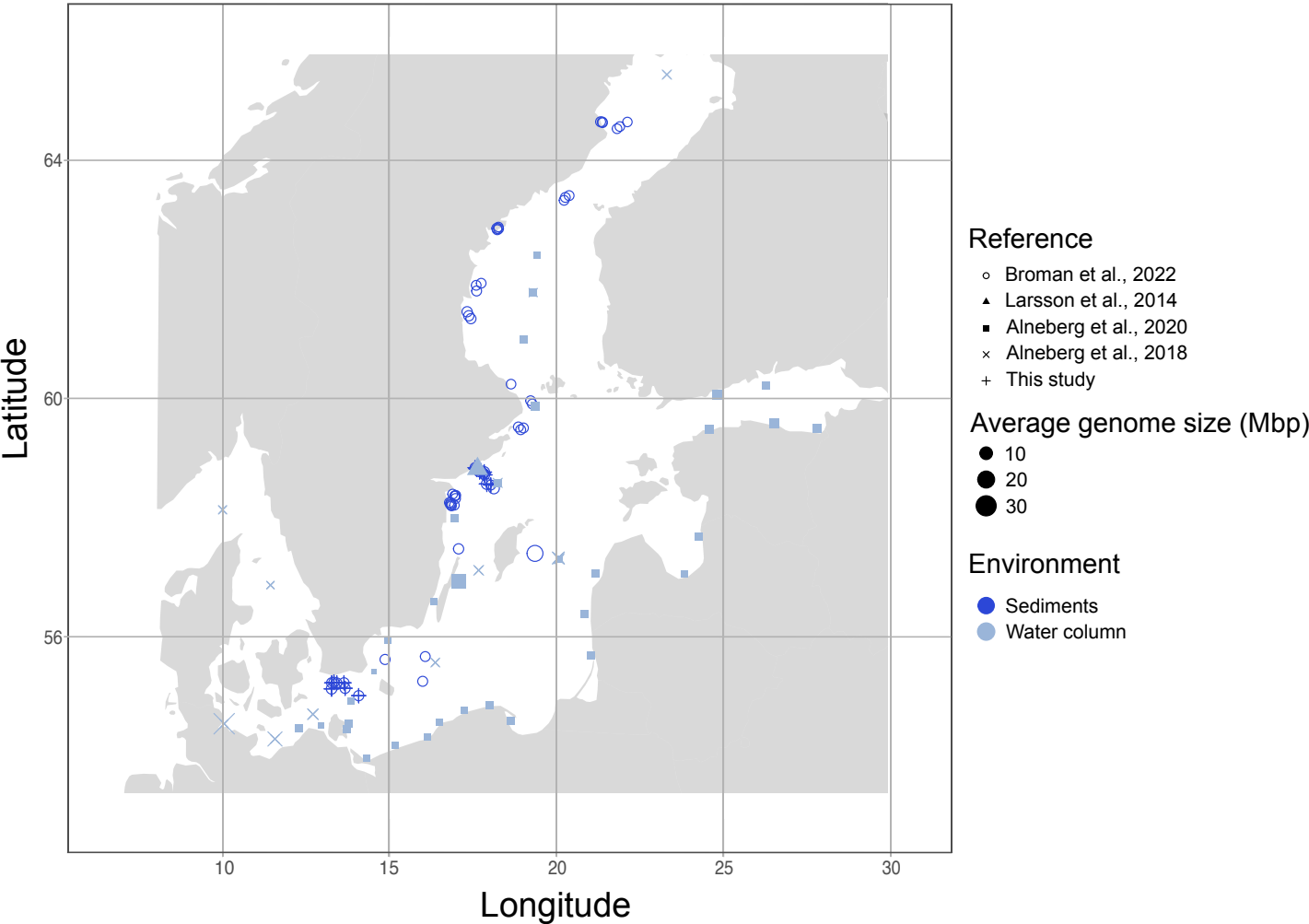
729

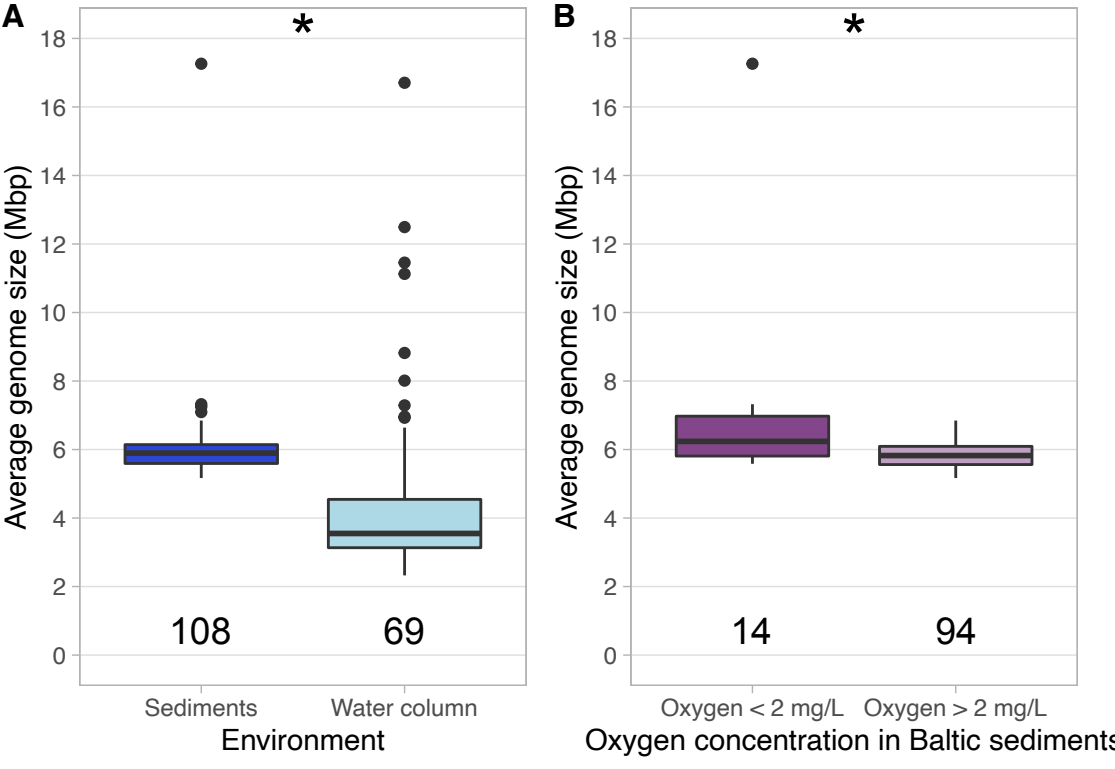
730 **Figure 4**. Overview of the estimated genome size of pelagic Bacteria and Archaea obtained
731 from Baltic Sea (blue), freshwater (green) and marine (yellow) using only representative
732 MAGs (calculates using 95% average nucleotide identity) with >75% completeness. We
733 compare all representative MAGs (**Panel A**), only phylum Actinobacteriota (**Panel B**), phylum
734 Bacteroidota (**Panel C**), phylum Cyanobacteria (**Panel D**), class Alphaproteobacteria (**Panel**
735 **E**) and class Gammaproteobacteria (**Panel F**). Different letters indicate significant differences
736 $p < 0.05$ (Kruskal-Wallis non-parametric test; multiple testing corrected with Benjamini-
737 Hochberg). Numbers below the boxes indicate the number of MAGs per environment.

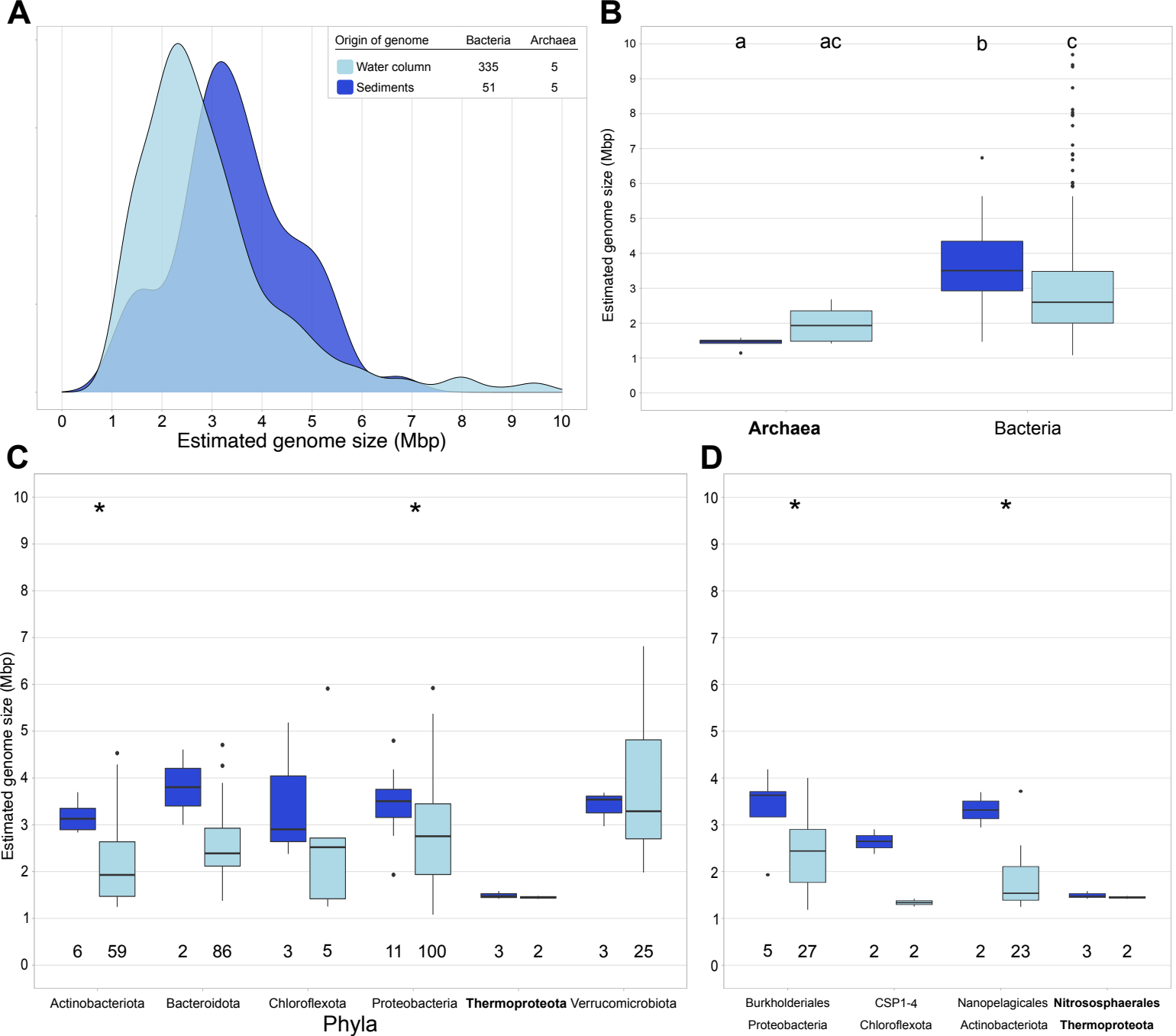
738

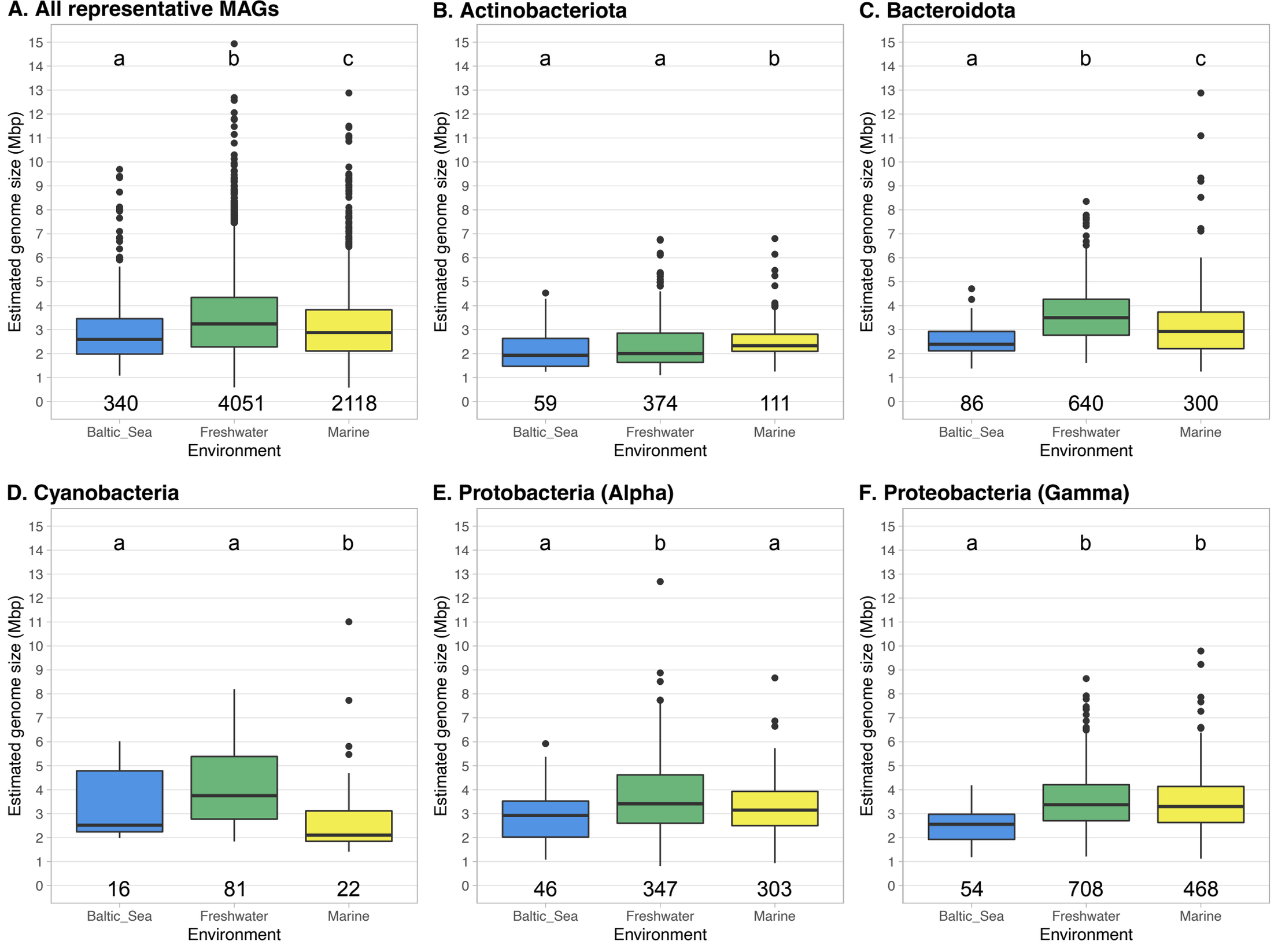
739 **Figure 5**. Overview of the presence of module steps for six metabolic categories: amino acid
740 metabolism (**Panels A-C**), aminoacyl tRNAs (**Panels D-F**), carbon fixation (**Panels G-I**),
741 central carbohydrate metabolism (**Panels J-L**), nitrogen metabolism (**Panels M-O**) and sulfur
742 metabolism (**Panels P-R**). We used all bacterial MAGs with very-high quality (>90%
743 completeness and <5% contamination), from both environments (99 MAGs from sediments

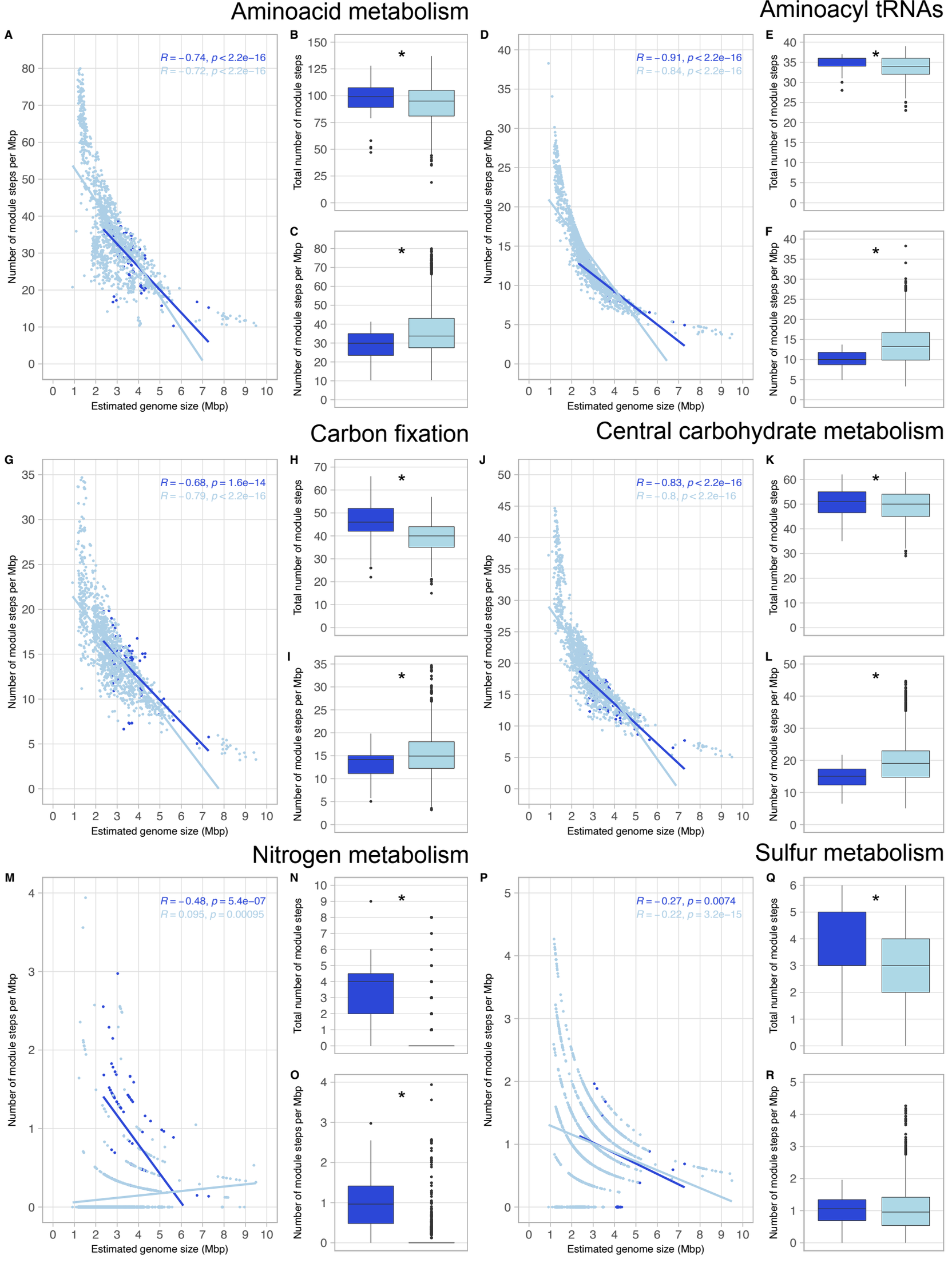
744 and 1215 MAGs from water column). Stars in boxplots indicate significant differences $p < 0.05$
745 (Wilcoxon non-parametric test). **Panels B, E, H, K, N** and **Q** indicate total number of module
746 steps. **Panels C, F, I, L, O** and **R** indicate the number of module steps per Mbp.
747
748 **Figure 6**. Metabolic potential of all high-quality MAGs (>90% completeness and <5%
749 contamination). We selected all phyla with at least 5 MAGs, and then divided by class. Boxes
750 on the top of the figure indicate environment (pelagic in light blue and sediments in dark blue),
751 and inner numbers indicate the number of MAGs per category. In the heatmap, white squares
752 indicate absence of a given gene in all MAGs, and the darkest purple indicates presence in all
753 of them (gradient scale on the bottom-left part of the figure for reference).
754
755 **Table 1**. Summary of all 56 sediment and 340 water column representative MAGs (95%
756 average nucleotide identity) with >75% completeness. Table includes phyla, environment
757 (either water column or sediments), number of representative genomes (*n*), smallest and largest
758 estimated genome sizes (Mbp) observed for each phylum, average GC content (%) and average
759 coding density (%). When only one MAG is indicated, estimated genome size of that MAG is
760 expressed in the sixth column.

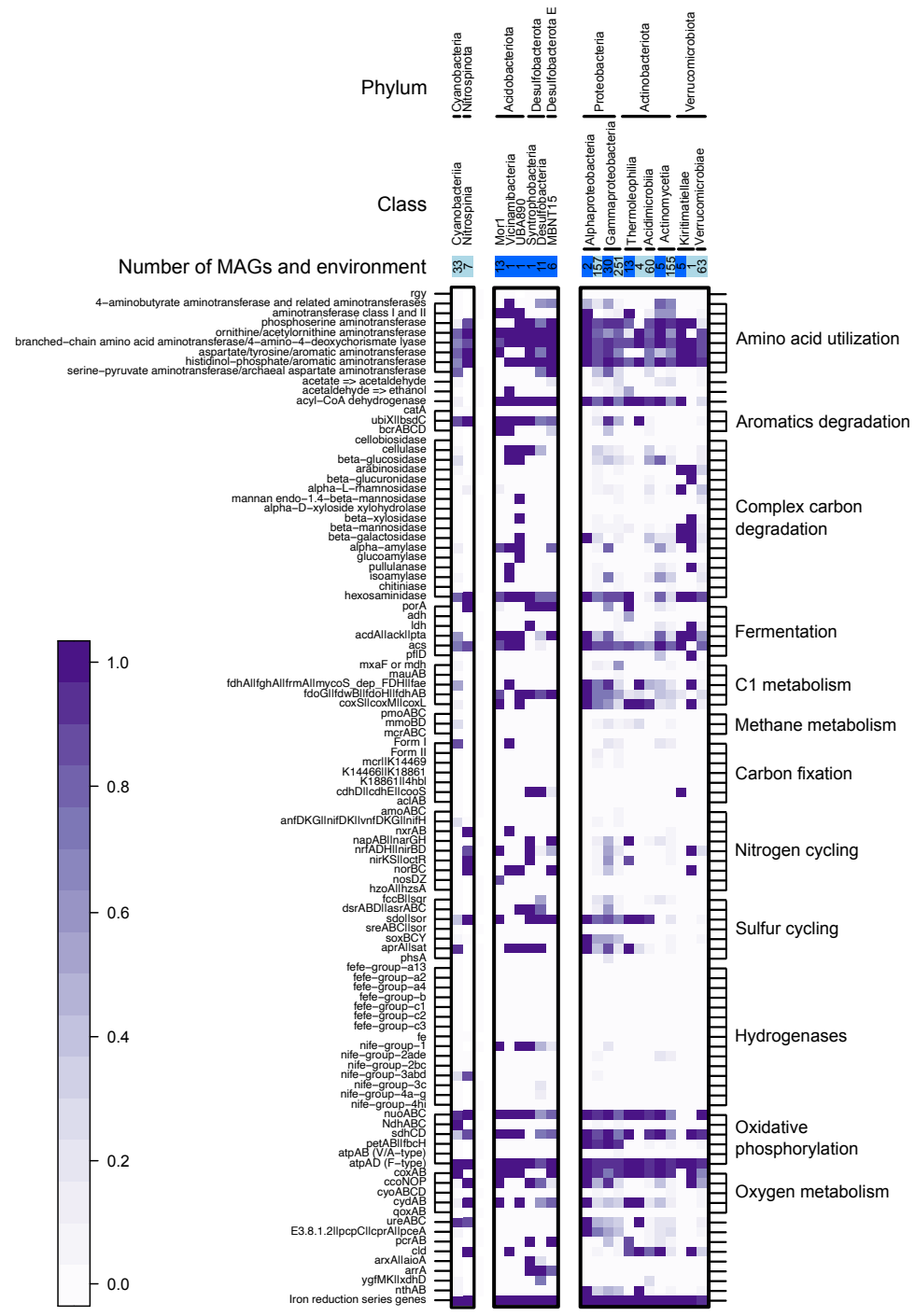

Latitude
64
60
56
10
15
20
25
30
Longitude
Reference
Broman et al., 2022
Larsson et al., 2014
Alneberg et al., 2020
Alneberg et al., 2018
This study
Average genome size (Mbp)
10
20
30
Environment
Sediments
Water column


A
*
Average genome size (Mbp)
18
16
14
12
10
8
6
4
2
0
108
69
Sediments
Water column
Environment
B
*
Average genome size (Mbp)
18
16
14
12
10
8
6
4
2
0
14
94
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)
Archaea
Bacteria
C
*
*
Estimated genome size (Mbp)
6
59
2
86
3
5
11
100
3
2
3
25
Actinobacteriota
Bacteroidota
Chloroflexota
Proteobacteria
Thermoproteota
Verrucomicrobiota
Phyla
D
*
*
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)
Baltic_Sea
Freshwater
Marine
Environment
a
b
c
340
4051
2118
59
374
111
86
640
300
16
81
22
46
347
303
54
708
468


Aminoacid metabolism
Aminoacyl tRNAs
Carbon fixation
Central carbohydrate metabolism
Nitrogen metabolism
Sulfur metabolism
A
R = −0.74, p < 2.2e−16
R = −0.72, p < 2.2e−16
Number of module steps per Mbp
Estimated genome size (Mbp)
B
Total number of module steps
C
Number of module steps per Mbp
D
R = −0.91, p < 2.2e−16
R = −0.84, p < 2.2e−16
E
F
G
R = −0.68, p = 1.6e−14
R = −0.79, p < 2.2e−16
H
I
J
R = −0.83, p < 2.2e−16
R = −0.8, p < 2.2e−16
K
L
M
R = −0.48, p = 5.4e−07
R = 0.095, p = 0.00095
N
O
P
R = −0.27, p = 0.0074
R = −0.22, p = 3.2e−15
Q
R



Phylum
Cyanobacteria
Nitrospinota
Acidobacteriota
Desulfobacterota
Desulfobacterota E
Proteobacteria
Actinobacteriota
Verrucomicrobiota
Class
Cyanobacteriia
Nitrospinia
Mor1
Vicinamibacteria
UBA890
Syntrophobacteria
Desulfobacteria
MBNT15
Alphaproteobacteria
Gammaproteobacteria
Thermoleophilia
Acidimicrobiia
Actinomycetia
Kiritimatiellae
Verrucomicrobiae
Number of MAGs and environment
33 7
13 1 1 1 11 6
2 157 30 251 13 4 60 5 155 5 1 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
chitiniase
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
nrfADHIInirBD
nirKSIIoctR
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