Large-scale phylogenomics of aquatic bacteria reveal molecular mechanisms for adaptation to salinity

Authors: Krzysztof T Jurdzinski[1*], Maliheh Mehrshad[2*], Luis Fernando Delgado[1], Ziling Deng[1], Stefan Bertilsson[2], Anders F Andersson[1+]

* Equal contribution

Affiliations:
1. Department of Gene Technology, KTH Royal Institute of Technology, Science for Life Laboratory, Stockholm, Sweden
2. Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden

[+]**Correspondence:** anders.andersson@scilifelab.se

Abstract

The crossing of environmental barriers poses major adaptive challenges. Rareness of freshwater-marine transitions separates the bacterial communities, but how these are related to brackish counterparts remains elusive, as are molecular adaptations facilitating cross-biome transitions. Here, we conduct large-scale phylogenomic analysis of freshwater, brackish, and marine quality-filtered metagenome-assembled genomes (11,276 MAGs). Average nucleotide identity analyses showed that bacterial species rarely existed in multiple biomes. Distinct brackish basins co-hosted numerous species despite differences in salinity and geographic distance, the latter having stronger intra-species population structuring effects. We further identified the most recent cross-biome transitions, which were rare, ancient, and most commonly directed towards the brackish biome. Transitions were accompanied by changes in isoelectric point distribution and amino acid composition of inferred proteomes, as well as convergent gains or losses of specific gene functions. Therefore, adaptive challenges entailing proteome reorganization and specific changes in gene content result in species-level separation between aquatic biomes.

Introduction

Transitions between different environmental conditions have long been identified as one of the major drivers of evolutionary innovation[1]. Evidence of fast dispersal of bacteria across the globe suggests that physicochemical and biological factors constrain the distribution of bacteria to a higher extent than geography and dispersal limitation[2]. However, the highly dynamic nature of bacterial genomes[3], including the ability to acquire new genes through horizontal gene transfer[4], can potentially facilitate the crossing of environmental barriers.

Salinity, pH, temperature, and other factors important for structuring microbial diversity are predicted to change in coastal waters due to climate change[5]. Combined with the crucial role of bacteria in such ecosystems[6], there is a need to understand the potential and limitations of adaptation as well as dispersal in response to changing environmental conditions. Comparative genome analyses across major environmental barriers have the potential to uncover the molecular mechanisms of such adaptation, as phylogenetic frameworks allow to disentangle the effects of habitat from differences in taxonomic composition.

Salinity has long been considered the strongest physicochemical factor determining bacterial community composition[7,8]. Distinct bacterial lineages inhabit freshwater and marine biomes, and transitions between these biomes have been rare in the documented evolutionary history of aquatic bacteria[9–11], pinpointing differences in salinity as a particularly strong barrier for microbes to overcome. More recently, fragment-recruitment-based metagenomic analyses have indicated that brackish environments with intermediate salinity levels also host genetically distinct bacterial lineages[12,13]. However, this has not been confirmed in a rigorous phylogenomic framework and little is known about the biogeography of the potential brackish microbiome and its relationship with its freshwater and marine counterparts. Many questions remain unanswered, such as which groups of bacteria have transitioned between the aquatic biomes, how often, in which directions, and when have the transitions taken place.

Protein properties such as isoelectric points were recently shown to differ between aquatic biomes with a pattern that may reflect adaptation to different extra- and intracellular ion concentrations[14]. Detected changes are mainly seen for surfaces of soluble proteins and differ between certain groups of closely related freshwater, brackish, and marine lineages[14]. Likewise, the functional gene content of microbial communities changes in a salinity-dependent manner[11,15–17], though it is unclear to what extent these effects are merely attributable to differences in the taxonomic composition. The widespread patterns of convergent gene gain and loss accompanying the cross-biome transition, as well as evolutionary factors and dynamics driving

the shift in isoelectric point distribution, could be further elucidated through a systematic phylogenomic analysis across the bacterial tree of life.

Here we analyze an extensive and diverse set of freshwater, brackish, and marine metagenome-assembled genomes (MAGs) in a phylogenomic framework. We cluster MAGs by average nucleotide identity (ANI) and compare the similarities of genomes from the same and different biomes. We identify the most recent transitions between the aquatic biomes and estimate their evolutionary timeline and directionality. Finally, we show that large-scale changes in the properties of the predicted proteome and specific alterations in gene content accompany the transitions in diverse bacterial lineages.

Results

A comprehensive dataset of MAGs from freshwater, brackish and marine biomes

A total of 13,783 MAGs were collected from freshwater[18–20], brackish[17,21], and marine[22,23] biomes (Fig 1a) to analyze evolutionary histories and functional adaptations in aquatic bacteria. The MAGs were filtered at ≥75% completeness and ≤5% contamination thresholds, and the 11,509 MAGs passing these criteria were considered for the remaining analyses (11,276 bacterial and 233 archaeal). For further analyses, we focused on the bacterial MAGs, of which 7643 were from freshwater, 2268 from brackish, and 1365 from marine waters (Table 1). Collected MAGs contained representatives from 72 phyla distributed across 135 classes and 348 orders, with the majority (91%) belonging to only 12 bacterial phyla with >100 representatives per phylum (Supplementary Data S1).

Bacterial species are rarely shared between biomes

Clustering the bacterial MAGs using ANI at the threshold of the operational definition of species (95%[24,25]) resulted in 3561 genome clusters (Fig 1b). MAGs of the vast majority of genome clusters (n=3547) belonged to a single biome (2063 freshwater, 485 brackish, and 999 marine genome clusters), while no genome cluster had representatives from all three biomes. Only fourteen genome clusters harbored MAGs belonging to pairs of biomes (nine, four, and one genome clusters with freshwater-brackish (FB), brackish-marine (BM), and freshwater-marine (FM) MAGs, respectively) (Fig 1c). The only shared FM genome cluster was affiliated to the genus *Limonbacter* (phylum Proteobacteria). Shared FB clusters were affiliated to phyla Actinobacteriota, Proteobacteria, Verrucomicrobiota, and Bacteroidota. The shared BM clusters belonged to phyla Proteobacteria, Verrucomicrobiota, and Bacteroidota.

102 Fortyfive of the 485 brackish genome clusters had members from both included brackish
103 water bodies; the Baltic and the Caspian Sea, while 300 and 141 were unique to Baltic and
104 Caspian, respectively. The shared clusters belonged to 13 classes in 10 different phyla
105 (Planctomycetota, Verrucomicrobiota, Gemmatimonadota, Bacteroidota, Actinobacteriota,
106 Chloroflexota, Cyanobacteria, Firmicutes, Nitrospinota, and Proteobacteria). The extent of
107 overlap between Baltic and Caspian Sea genome clusters (18.9%) was significantly higher than
108 between brackish and freshwater (0.007%) and brackish and marine clusters (0.012%; $P < 10^{-16}$
109 for both comparisons), despite the geographic separation and lack of hydrological connectivity
110 between these ecosystems. This supports the previously postulated existence of a global
111 brackish-specific microbiome[12,21].
112 The clustering results above indicate that bacterioplankton belonging to the same species
113 (within the 95% ANI threshold) are inhabiting the geographically distinct and distant Baltic and
114 Caspian Seas. To investigate intra-specific (within-species) biogeographical patterns, we used a
115 population genomics approach, where allele frequencies in single nucleotide variant (SNV) loci
116 are obtained by aligning metagenomic sequences to the genomes. As previously reported[26],
117 principal coordinates (PCoA) analysis of pairwise fixation indices (F_{ST}) showed that for each of
118 the five genome clusters where we applied this approach in the present study, populations were
119 highly structured according to salinity within the Baltic Sea (Fig. 2a). However, adding the Caspian
120 Sea data to the analysis revealed even stronger population structuring between the two seas (Fig.
121 2b), indicating geographic separation at the strain level within these species.
122

123 Transitions between aquatic biomes are rare, ancient, and in most cases directed 124 into the brackish biome

125 We reconstructed the phylogeny of the aquatic MAGs using a set of conserved housekeeping
126 genes[27]. The resulting tree was pruned to contain a maximum of one MAG from each biome per
127 previously identified genome cluster (i.e. ~species) (Fig. 3). The most recent transitions between
128 the aquatic biomes, if detected, should be represented on the phylogenetic tree as pairs of closest
129 related monophyletic and biome-specific groups. We call these clades monobiomic sister groups
130 (MSGs, Fig. 4a). We identified 310 MSG pairs, with brackish and marine (BM; n=136) and
131 freshwater and brackish (FB; n=119) transitions being twice as numerous as freshwater and
132 marine (FM; n=55), despite the larger number of freshwater and marine MAGs in the pruned tree
133 (Fig. 4b). The number of FB and FM transitions were also significantly lower than would be
134 expected if phylogeny would not correlate with biome annotation (Fig. 4c). This corroborates
135 observations for the previously described genome clustering.

136 Most of the identified FB and BM MSG pairs belonged to the same genus (around 55% in
137 both cases), while for FM transitions the lowest-level shared taxonomic annotation was most often
138 (49%) at the family level (Fig. 4d, Supplementary Table S1). Only small proportions (1.8 - 4.2%)
139 of the MSG pairs belonged to the same species (defined either by GTDB annotation or as a >95%
140 ANI genome cluster). The taxonomic distribution of MSG pairs followed the overall distribution in
141 the corresponding biomes for BM and FM, but not for FB, probably due to the overrepresentation
142 of Actinobacteriota and Cyanobacteriia in this transition type (Fig. 4e, Supplementary Fig. S1).
143 We estimated the minimal time since divergence of the MSG pairs based on phylogenetic
144 distances, using a method which allows for variable evolutionary rates[28] (Fig. 3, Fig.4f,
145 Supplementary Fig. S2). As time constraints, we used minimal times since divergence of
146 genomes of endosymbiotic bacteria known from fossil records of their hosts[29–37] (Supplementary
147 Table S2). As obligate endosymbionts have high evolutionary rates that exceed those of free-
148 living bacteria[37], this model gives conservative estimates of time since the transitions started. The
149 timing differed depending on the transition type (Fig. 4f; $P < 10^{-6}$ for all comparisons), FB being
150 the most recent and FM the oldest. The difference in median times since FB and BM transitions
151 could not be attributed to the uneven representation of biomes in the dataset (Supplementary Fig.
152 S2b).
153 The most recent transition was estimated to have happened at least 44.8 kya, and all other
154 transitions millions of years ago (Supplementary Table S3). Therefore, all the detected transitions
155 started long before the current brackish conditions were established in the Baltic Sea (8 kya[38])
156 and only the six most recent ones to the period since the Caspian has continuously been brackish
157 (2-3 mya[39]). These results imply that the evolutionary history of the global brackish microbiome[12]
158 by far predates the formation of these two brackish basins.
159 To infer the directions of the biome transitions, we conducted ancestral (biome) state
160 reconstructions of the most recent common ancestors (MRCAs) of the MSG pairs, correcting for
161 the uneven representation of biomes in the pruned tree (see Methods). FB and BM transitions
162 were generally more likely to happen into than out of the brackish biome, while for FM transitions
163 no significant direction preference was observed (Fig. 4g-h, Supplementary Fig. S3 and 4).
164

165 **Large-scale changes in proteome properties follow cross-biome transitions**

166 We next assessed if changes in proteome properties accompany the biome transitions. The
167 average distributions of isoelectric points (pIs) of the proteins encoded in the genomes were
168 compared between MSGs in a pairwise manner. Significant differences in the proportion of neutral
169 (pI ∈ [5.5, 8.5)) proteins were observed for all the transition types (Fig. 5a-b, Supplementary Data

170 S3), with a higher proportion in the biomes of lower salinity. The frequencies of acidic proteins (pI
171 ∈ [3.0, 5.5)) showed the opposite pattern and were significantly different for the comparisons that
172 included freshwater MSGs, but not for the BM transition type. The results suggest that the higher
173 representation of neutral pIs observed in freshwater bacteria[14] gradually diminishes with
174 increasing salinity while low proportion of acidic proteins is a distinctive feature of freshwater
175 bacteria. However, the extent of transition-related changes in pI distribution highly depended on
176 taxonomy, with particularly small changes for Actinobacteria (Supplementary Fig. S5).
177 Polar, charged, and acidic amino acid frequencies broadly increased with salinity across
178 the MSG pairs (Fig 5c, Supplementary Data S3). Changes were most significant for acidic amino
179 acids, which is a subcategory of the other two categories and may thus also explain the changes
180 observed for those. In contrast, the frequency of basic amino acids differed significantly only for
181 BM transitions.
182 Among the acidic amino acids, the difference for glutamate was larger and more significant
183 than for aspartate across all the transition types (Supplementary Fig. S6, Supplementary Data
184 S3). For BM transitions, the higher proportion of basic residues in marine proteomes could be
185 attributed to lysine. Big and significant differences in the frequency of alanine across BM and FM
186 transitions suggest that this non-polar amino acid with a relatively "neutral" effect on protein
187 structure[40] is often substituted by charged residues in transitions to higher salinity. For all
188 transition types, there was significantly less proline in the predicted proteomes at higher salinity.
189 To investigate the dynamics in how the proteome is reshaped after biome transitions, we
190 compared the differences in pI distributions to inferred times since divergence. The overall change
191 in pI distributions followed different logarithmic relationships for within-MSG (no transition)
192 divergence times as compared to any of the transition types (Fig. 5d). Moreover, the differences
193 in pI distribution across transitions were consistently larger than changes resulting from within-
194 MSG (no transition) divergence events when the latter were at least 44 to 181 million years more
195 recent (Fig. 5g-i). This suggests that proteomes are under strong selection and continue to evolve
196 to match the new salinity conditions long after the switch to a new biome (Supplementary Fig.
197 S7).
198
199
200 **Specific changes in gene content accompany cross-biome transitions**
201 For each transition type, we were able to identify specific functional genes (KEGG orthology
202 groups [KOs]), significantly over-represented in the MSGs of one of the compared biomes after
203 correcting for differences in estimated completeness of the MAGs (Fig. 6, Supplementary Data

204 S4). However, transitions did not change the gain and loss rates (Fig 5e), even though gene
205 numbers significantly differed for BM transitions (Fig. 5f, Supplementary Data S3), with brackish
206 MAGs having a median of 386 more genes. Therefore, the gains or losses of the same gene
207 functions (KOs) we observed in multiple independent transition events most likely result from
208 analogous changes in selective pressures caused by the biome transitions.

209 Transporters were the only functional category (KO level C) over-represented in all
210 transition types (Fig. 6, Supplementary Figure S8), some with known roles in adaptation to altered
211 salinity. For FB transitions, brackish MAGs more frequently contained components of two (or of
212 the same[41]) potassium uptake systems (K03498 and K03499) co-acting to allow growth in higher
213 osmotic pressure[42]. A component of the uptake system with higher affinity, K03499[42], was also
214 more prevalent in marine genomes across FM transitions. Additionally, for both FB and FM MSG
215 pairs, a Na^+/Ca^{2+} antiporter (K07301) was more often found in the more saline biomes. Analogous
216 changes in presence of the same K^+ uptake systems and of Na^+/Ca^{2+} antiporters were reported
217 in a comparison of terrestrial and marine *Flavobacteriaceae*[43]. Finally, K03282, more frequently
218 present in brackish across BM transitions, is a mechanoreceptor involved in managing
219 hypoosmotic stress[44].

220 Two distinct magnesium uptake systems[45,46] were more frequently present in lower-salinity
221 biomes for FB and BM MSG pairs (K01531 and K03284, respectively). Moreover, for BM
222 transitions, brackish MAGs displayed a higher presence of chemotaxis (K03413) and two-
223 component system sensors (K02482 and K20974) involved in regulation of motility[47,48]. While
224 Mg^{2+} is a major component of sea salt[11] and salinity can alter chemotactic responses and
225 motility[49–51], to the best of our knowledge, there are no previous reports on the eco-evolutionary
226 relevance of these genes. For FB, a gene connected to either pigment synthesis (K15777) or
227 aromatic hydrocarbon degradation (K04100) was more often found in freshwater MAGs. Both of
228 these functions may be connected to salinity-related changes in dissolved organic carbon
229 concentration (see Supplementary Discussion for more details).

230 Other differentially present genes were less directly connected to physicochemical
231 differences between the biomes, suggesting more complex and/or cryptic biological mechanisms
232 associated with transitions. For BM transitions, genes connected to various metabolic pathways
233 (including modifications of charged amino acids), as well as transposases, antitoxins, and
234 competence proteins, were more often present in brackish MAGs than their marine counterparts.
235 Typical housekeeping genes were absent among the differentially present genes, which verifies
236 the specificity of the method.
237

Discussion

In this study, we combined several large-scale datasets, which allowed us to comparatively analyze bacterial genomes from freshwater, brackish and marine biomes at an unprecedented scale. Including brackish genomes allowed for a more detailed analysis of the evolutionary history of aquatic bacteria than in previous studies focusing on the two biomes at the extremes of the spectrum[9–11,16,43,52,53]. The brackish biome is a significant part of our global waters and combined, the two major brackish systems analyzed here have a similar volume (~99,000 km^3 [54]) as all freshwater lakes on earth (≤102,000 km^3 [55]), albeit only a fraction of the volume of the global ocean (~1.35 billion km^3 [56]).

We observed bacterial communities from the different biomes to be separated on at least the species level - nearly all >95% ANI genome clusters were restricted to a single biome. Brackish genome clusters often had representatives from both the Baltic and Caspian Sea. These results provide further support that the brackish biome hosts a distinct, globally distributed set of bacterial species[12]. However, the fact that populations of individual species were more genetically differentiated between the two brackish seas than across the strong salinity gradient within the Baltic Sea indicates that geographical barriers still have a structuring effect on the distribution of aquatic bacteria. This implies a limitation in microbial dispersal. While it might be of limited relevance on a large evolutionary timescale, it is likely to be of crucial importance when an ecosystem is disturbed, as it may take substantial amounts of time for migrating strains to fill the niches which open up as a consequence of the disturbance.

Alternatively, local bacteria could adapt to environmental changes, maintaining a diverse and functioning ecosystem throughout the process. This is unlikely to be the case when it comes to changes in salinity spanning across biome barriers however, since neither the formation of the Baltic nor the Caspian Sea caused an observable increase in transition rates, which is consistent with previous notions of major geological events having little impact on bacterial speciation rates[57]. Therefore, most of the local microdiversity which we uncovered through population genomics is likely to be lost as a consequence of such environmental changes. Ultimately, the destruction and disturbance of environments, enhanced in the times of global change, not only threaten local microbial diversity in the short run but also diminish genetic diversity even within global species.

Our analyses support previous concepts of freshwater-marine transitions being rare[9–11] and show more common but still rare freshwater-brackish and brackish-marine transitions, which most often are directed towards the brackish biome. However, as the three biomes have continuously existed throughout geological history and bacterial species appear to be globally

272 distributed within each of them[2], one might ask why any recent transitions have even happened.
273 A plausible scenario may be that genetic innovation (gene gain/loss or mutation) can make a
274 lineage competitive in a different biome, either by opening up a new niche or out-competing others
275 in an existing one, in accordance with the stable ecotype model[58]. However, the benefits of the
276 innovation need to outweigh the decrease in fitness due to switching to a new salinity regime.
277 Thus, the smaller salinity difference between the brackish and the other biomes would explain
278 why it is associated with more transitions.

279 The more frequent transitions into than out of the brackish environment are harder to
280 explain. For FB, the reason could simply be hydrological, i.e. that freshwater usually flows into
281 brackish water rather than the other way around. For BM, differences in population sizes may
282 instead be the explanation, as both the volume and surface of the ocean are much bigger than of
283 all brackish waters. Beneficial mutations are thus more likely to appear among the more numerous
284 marine bacteria. We also observed that there was a small fraction of FB and BM transitions with
285 unusually high third-biome ancestral states (Supplementary Fig. S4). These events might
286 correspond to FM transitions through brackish waters, where an intermediate period in brackish
287 water would allow for a more gradual adaptation than between FM directly.

288 Our results corroborate previous reports on differences in predicted proteomes at different
289 salinities[14], and support them with a systematic analysis across the bacterial tree of life. These
290 changes likely occur over long evolutionary timescales after the actual transition, meaning that in
291 the period after the transitions, bacteria may have a suboptimal protein repertoire in their new
292 environment. This environmental mismatch would potentially lower the fitness of newly
293 transitioned lineages in addition to the bottleneck effects, traces of which are present in LD12
294 clade genomes[59], and which can lead to fixation of slightly deleterious mutations[60].

295 We observed a shift towards more acidic protein pIs at higher salinity, also reported
296 before[14,61], and showed the underlying proteome-scale increase in proportion of acidic amino
297 acids, most notably glutamate. This likely reflects that charged amino acids increase protein
298 solubility at higher ionic strength[62,63]. However, an even stronger pattern observed was the
299 proportion of neutral proteins gradually diminishing with increasing salinity, which is also known
300 from previous studies but lacks explanation[14]. We hypothesize that this effect might be due to
301 diminishing extracellular pH variation as the concentration of buffering salts increases[11,64].
302 Moreover, Na^+-coupled transport is the key mechanism of pH homeostasis under alkaline
303 conditions[65], to which both freshwater[11] and brackish environments[64,66] transcend. Consistently,
304 lower NaCl concentrations have been connected with higher intracellular pH variation in alkaline
305 conditions[65]. Thus, neutral proteins could be soluble and active outside as well as inside the cell

306 under pH variations in less saline environments. Unlike freshwater, brackish pH rarely changes
307 to an acidic state[64,66], which may explain why the difference in frequency of acidic proteins was
308 less significant for BM than for other transition types, accompanied by a change in the frequency
309 of not only acidic but also basic amino acids.

310 For BM transitions, we identified genes involved in glutamate post-translational
311 modifications (K03412, K05844, K14940, K18310) to be more frequent in brackish genomes. All
312 these modifications increase the acidity of residues[67–70]. Thus, depending on the specificity of the
313 enzymes, brackish bacteria may change the charges (and pIs) of their proteins to adjust to
314 different salinities (and varying pH). We also observed that brackish bacteria were enriched in
315 gene functions for other physiological responses to environmental cues, including chemotaxis,
316 transcriptional regulation, as well as synthesis of trehalose and polyamines (both osmolytes) (see
317 Supplementary Discussion). Thus, phenotypic plasticity might be an important hallmark of
318 brackish bacteria, sustaining the populations across the physicochemical gradients typical for
319 many brackish environments.

320 BM transitions seem to expose bacteria to a different set of mobile genetic elements
321 (MGEs). This is suggested by a set of transposases (previously reported to be unusually abundant
322 in the Baltic Sea[71]), competence proteins, and, indirectly, antitoxins[72] being more often present in
323 the brackish MAGs (Fig. 6a). Consequently, more MGEs in the brackish MAGs could explain the
324 higher numbers of genes as compared to the marine relatives (Fig. 5f). MGEs tend to increase
325 gene gain and loss processes[73,74], and may thus facilitate fast adaptation to new physiochemical
326 conditions. It also promotes the establishment of a recombination barrier between biomes[75]. As
327 homologous recombination inhibits divergence[75–79], these effects would give a further mechanistic
328 explanation for the “species-level” separation.

329 We used a novel comparative genomics approach to leverage the large phylogenetic scale
330 of the analysis. It allows systematic comparison of close relatives differing in categorical states
331 and evolutionary dynamics of transitions between these states. Defining the transitions as pairs
332 of MSGs, we could meaningfully classify them without relying on often ambiguous ancestral state
333 inference (which we still perform, but only for inferring directions of transitions and in the process
334 correcting for imbalances in the dataset in an evolutionary history-informed manner). Moreover,
335 comparative genomics and evolutionary dynamics can be linked to each other with this approach.
336 That can be instrumental to move beyond mere correlations of ecological and genomic changes
337 toward their causal relationship, as we showed by unraveling the prolonged nature of post-
338 transition proteome reorganization. While correlation-based methods can also be adjusted for a
339 taxonomically balanced comparative genomics analysis [80], it is much harder to connect these

results to specific evolutionary events. Furthermore, with the new framework, we could correct for varying levels of genome completeness across MSG pairs, and not downsampling all the MAGs to the same, lowest, level, therefore minimizing the information loss. Consequently, we detected gains or losses of genes obviously related to adaptation to salinity but not any housekeeping genes. We were thus able to distill the traces of repeated gene gains and losses due to habitat change from the taxonomic effects which govern the differences in gene content of whole metagenomes [16]. At the same time, we identified convergent changes connected with cross-biome transitions in diverse bacterial lineages. These results, and the approach, can serve as a basis to disentangle the effects of transitions from consequences of parallel events and/or random changes when studying evolutionary histories of specific taxa.

Ultimately, we extend the view of transitions between aquatic biomes and highlight the brackish environment, showing trends specific for certain transition types and common across all of them. We report patterns prevalent across the bacterial tree of life, providing supplementary materials that allow investigation of the changes for chosen taxonomic groups. We observed many changes related to previously undescribed adaptation mechanisms and/or poorly characterized genes, both of which call for further exploration. In the future, phylogenomic analyses should be supplemented with experimental and ecological approaches. This would allow examining the relevance of the unraveled adaptation mechanisms in the context of ecosystem functioning and physiology.

Methods

MAG collection and sources

Representative metagenome-assembled genomes (MAGs) from freshwater, brackish, and marine biomes were collected from seven previous studies as specified in Table 1. Freshwater MAGs originate from two North American lakes (Lake Mendota and Trout Bog), 44 lakes of the StratfreshDB (mainly in Sweden, Finland and Canada), and Lake Baikal (Russia)[18–20]. Brackish MAGs originate from the Baltic Sea[17] and the Caspian Sea[21]. Finally, marine MAGs that were published by Tully et al.[22] and Delmont et al.[23] originate from the Tara Oceans Survey[81,82].

MAG statistics, ANI calculations and classification

Completeness and contamination of the collected MAGs were assessed using CheckM[83]. MAGs with ≥75% completeness and ≤5% contamination were considered for further analyses. Pairwise

374 Average Nucleotide Identity (ANI) between all MAGs was calculated using FastANI (V1.2) and
375 MAGs were clustered at the threshold of 95% [84]. Clustering of MAGs was done using hclust and
376 cuttree functions in Rstudio. To remove redundant information and reduce the complexity of the
377 phylogenetic tree we kept one random MAG per cluster per biome to generate a pruned tree.
378 Taxonomic affiliation of MAGs was assigned using the "classify_wf" workflow of the Genome
379 Taxonomy Database Toolkit (GTDB-tk) (v 0.3.2)[25,27]. To compare proportion of inter-biome and
380 brackish inter-basin clusters, we used Fisher's exact test conditioned on the lower number of
381 clusters (2 categories: the basin/biome with a lower number of clusters, and the number of shared
382 clusters).
383

384 **Population genomics**

385 Population genomics was conducted with the POGENOM software as described in Sjöqvist et
386 al.[26]. The same set of 22 MAGs (from[17]) were used as references for mapping of metagenomic
387 data from the Baltic[17] and Caspian[21] seas. Only five ese obtained sufficient coverage in the three
388 Caspian Sea samples and were included in the downstream analysis. The five MAGs all represent
389 different (>95% ANI) genome clusters. The Input_POGENOM pipeline was used for the automatic
390 generation of input files for POGENOM., The pairwise F_{ST} values output from POGENOM were
391 used to conduct Principal Coordinates Analysis (PCoA) using the Vegan package in *R*. Maps
392 were plotted with the rworldmap package. For parameter settings of Input_POGENOM and
393 POGENOM, see Sjöqvist et al.[26].
394

395 **Phylogeny reconstruction**

396 The "de_novo_wf" function of GTDB-tk (v 0.3.2) was used to reconstruct the phylogeny of
397 MAGs[25,27]. The WAG (Whelan and Goldman) amino acid evolution model was selected in the
398 infer step and phylum Patescibacteria was selected as the outgroup[85]. Most downstream
399 phylogenetic analyses were performed in RStudio (V3.6.2) with the R packages ape, phytools,
400 phangorn and ggplot. The multi2di function in R was used to resolve multichotomies (that appear
401 because some MAGs were nearly identical) in the phylogenetic tree. Since this function rendered
402 some branches with zero length, we arbitrarily changed all zero-length branches to 5×10^{-4} (which
403 was the smallest length of all other branches) to ensure all branch lengths were > 0.
404

405 **Gene prediction and annotation**

406 Prodigal (v2.6.3) was run with default settings to identify protein-coding genes[86]. Gene function
407 annotation was performed with eggNOG-mapper (v2.0.0), based on the eggNOG v5.0

408 database[87,88]. A custom Python script was used to count the number of occurrences of each Kegg
409 Ortholog (KO) in each annotated MAG.
410

411 **Identifying monobiomic sister groups (MSGs) for comparative analyses**

412 Pairs of MSGs (Fig. 4a) were identified based on the pruned tree. The most recent common
413 ancestors (MRCAs) of the monobiomic groups were identified by choosing nodes for which on
414 each of the two downstream branches there were MAGs from a different and only one biome.
415 Transition types were defined by the two biomes from which the MAGs originate and do not take
416 into consideration inferred direction of the transition. The biggest possible pairs of monophyletic
417 groups within MSGs (coming from the same biome; only one pair for each MSG with >1
418 representative MAG on the pruned tree) were used as "no transition" divergence events.
419

420 **Expected number of transitions**

421 Biome annotation of the tips of the pruned tree was randomly permuted 100 times and MSG pairs
422 were re-identified using the same tree structure and the random biome annotations. Rounded
423 averages of transitions obtained in all the iterations were used as the expected numbers. One-
424 sided P-values were calculated as the number of iterations in which a lower/higher number of
425 transitions of a type was observed. The fractions of each transition type among all expected
426 transitions were used as probabilities for *X*-squared test on the observed values.
427

428 **Minimal time since divergence**

429 Six divergence-time constraints were set as minimal estimates of time since divergence for pairs
430 of obligate endosymbiont taxa based on fossil record of host species[29–37] (Supplementary Table
431 S2, Supplementary Data S2). Divergence times were calculated using RelTime[28] based on the
432 unpruned trees, including GTDB reference sequences. Maximum relative rate ratio was set to
433 100. To correct for the difference between the placing of the root based on phylogenetic distances
434 only[85] and the most probable position of the last bacterial common ancestor (LBCA) on the tree[89],
435 *Fusobacteriota* were chosen as an outgroup. Only the tips present in the pruned tree were chosen
436 for further analyses.
437

438 **Transition directions**

439 One MAG from each MSG was randomly chosen. Additional MAGs from outside from the MSGs
440 were randomly chosen to downsample the dataset to the same number of genomes from each
441 biome, equal to the lowest number for any of the biomes (396 for brackish after removing multiple

442 MAGs within each MSG). “ace” function from the “ape” package in R was used to obtain maximum
443 likelihood ancestral states of the MRCAs of MSG pairs using an all-rates-different model[90].
444 Pairwise Wilcoxon rank-sum test of likelihoods of the ancestral states for the two biomes between
445 which transition occurred was performed for each transition type. Random selection of MAGs
446 within and outside the MSGs was iterated 100 times and mean likelihoods and p-values were
447 used for further statistical analyses.
448

449 **Comparison of isoelectric points and amino-acid frequencies**

450 Isoelectric points (pI) of all proteins within inferred bacterial proteomes were obtained using the
451 pepstats function from the EMBOSS package (v6.6.0)[91]. To get amino acid and amino acid
452 category frequencies the whole proteomes were concatenated into one sequence before running
453 separate pepstats analyses. pI frequencies within bins of pH width of 0.5 spanning the range from
454 3.0 to 12.5 were used to average, plot and compare the distributions of pIs between MSGs.
455 Average values over MAGs (95% ANI cluster representatives) were used for MSGs consisting of
456 more than 1 MAG.
457 For analysis of total changes in pI distribution, the absolute values of differences for each
458 0.5 pH wide bin were summed up. To compare transitions to divergence events within the same
459 biome, within each MSG bigger than 1 representative MAG, a pair of monophyletic groups which
460 diverged the longest time ago were identified. The differences between these “no transtion”
461 (within-MSG) monophyletic groups were compared with differences across the ancestral nodes
462 (MRCAs of MSG pairs) in a pairwise manner.
463

464 **Functional gene overlap and gene numbers**

465 Occurrences of KOs were transformed into a binary table of gene (KO) presence/absence. Gene
466 overlap was calculated as: $\frac{N_{gS}}{(N_{g1}+N_{g2})/2}$, where N_{gS} is the number of KOs shared between 2
467 genomes and N_{g1} and N_{g2} are the total numbers of KOs present in the respective genomes[75].
468 Formulas *log(gene overlap) ~ divergence times* and *gene overlap ~ divergence times* were used
469 to fit logarithmic and linear models for each transition type, including no transition (see
470 Comparison of isoelectric points and amino-acid frequencies). The models were compared using
471 a method based on the least-square means statistic[92]. Numbers of distinct polypeptide sequences
472 in the predicted proteomes were used as the gene numbers.

Gene presence/absence analysis

Occurrences of KOs were transformed into a binary table of gene (KO) presence/absence. Random pairs of MAGs from the MSGs were chosen for comparison for each transition type. To correct for differences in genome completeness, the presence count for the more complete MAG in each pair was randomly downsampled to the level of the less complete MAG. For each transition type, KOs present in at least one of the MAGs within any of the MSGs were analyzed. For each of the KOs, differences in presence between MAGs from each combination of two biomes were analyzed using pairwise Wilcoxon rank-sum test[93]. The procedure was iterated 100 times, starting from the step of randomly choosing pairs of MAGs. The Benjamini-Hochberg method[94] was used to obtain false-discovery rate (FDR)-adjusted p-values (q-values), and average q-values from the 100 iterations for each KO were used to determine the significance of difference in each KO's presence/absence. The identified KOs were annotated using KEGG orthology[95] table collected from the KEGG website on the 24th of November 2021.

Statistical analyses

If not stated otherwise, pairwise Wilcoxon rank-sum test[93] was used to compare values across MSG pairs of a certain type. If not stated otherwise, family-wise error rate (FWER) ≤ 0.05 was considered as the threshold of significance. The analyses were performed using R.

Data availability

All the MAGs used in this study are made publicly available via their original publications as referenced in the text and Table 1. Data files describing intermediate data analysis steps, including the phylogenetic trees, and the code for in-house data transformation procedures, are available at https://doi.org/10.17044/scilifelab.20732170.

Funding

This study was supported by a grant from the Swedish Research Council VR to A.F.A. (grant no. 2021-05563). Publication funding was provided by KTH Royal Institute of Technology.

Author contributions

A.F.A designed the study in consultation with all coauthors. K.T.J., M.M, A.F.A, Z.D., F.D. analyzed the data. K.T.J. and M.M. wrote the first draft of the manuscript. All authors contributed

506 to the interpretation of the data, revised the manuscript and approved the final version before
507 submission.
508

509 ## Acknowledgement
510 Authors would like to thank Dr. Moritz Buck for providing access to the StratFreshDB MAGs that
511 were included in this study.

512

513 **Conflict of interest**
514 Authors declare no conflict of interest.
515

516 ## Tables and Figures
517 **Table 1-** Number of MAGs originated from each dataset
518

Dataset	#metagenomes	#Total MAGs	#filtered MAGs	#filtered Bacteria	#filtered Archaea	Reference
North American lakes (F)	141	193	84	84	0	Linz et al. (2018)
StratfreshDB (F)	267	8554	7654	7554	100	Buck et al. (2021)
Lake Baikal (F)	2	35	7	5	2	Cabello-Yeves et.al. (2018)
Baltic Sea (B)	82	1989	1989	1954	35	Alneberg et al. (2020)
Caspian Sea (B)	3	324	324	314	10	Mehrshad et al. (2016)
Global Ocean T (M)	243*	2631	990	932	58	Tully et al. (2018)
Global Ocean D (M)	243*	957	461	433	28	Delmont et al. (2017)

519 * The same metagenomes were used to obtain MAGs by Delmont et al. (2017) and Tully et al.
520 (2018) using different approaches.

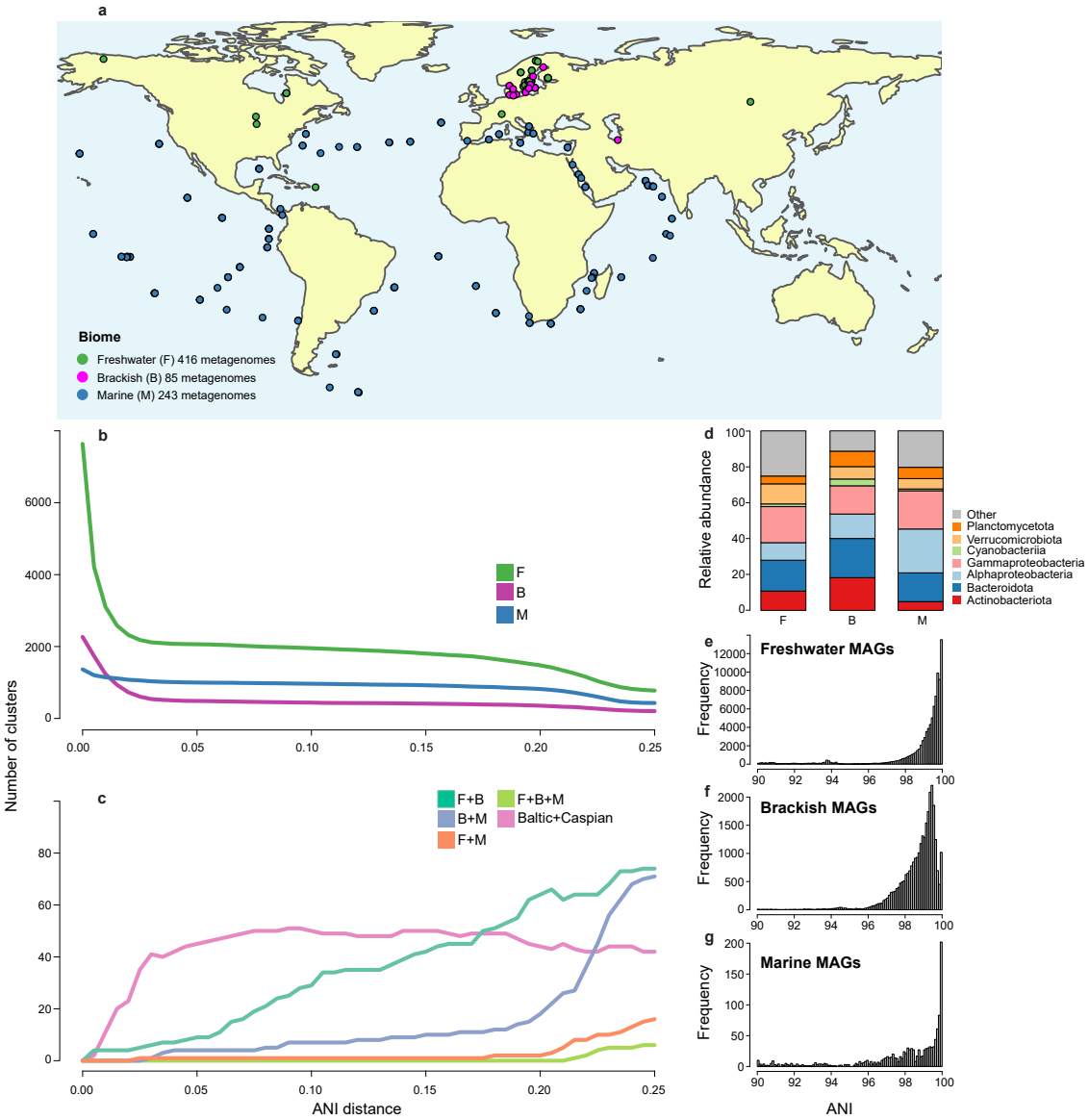
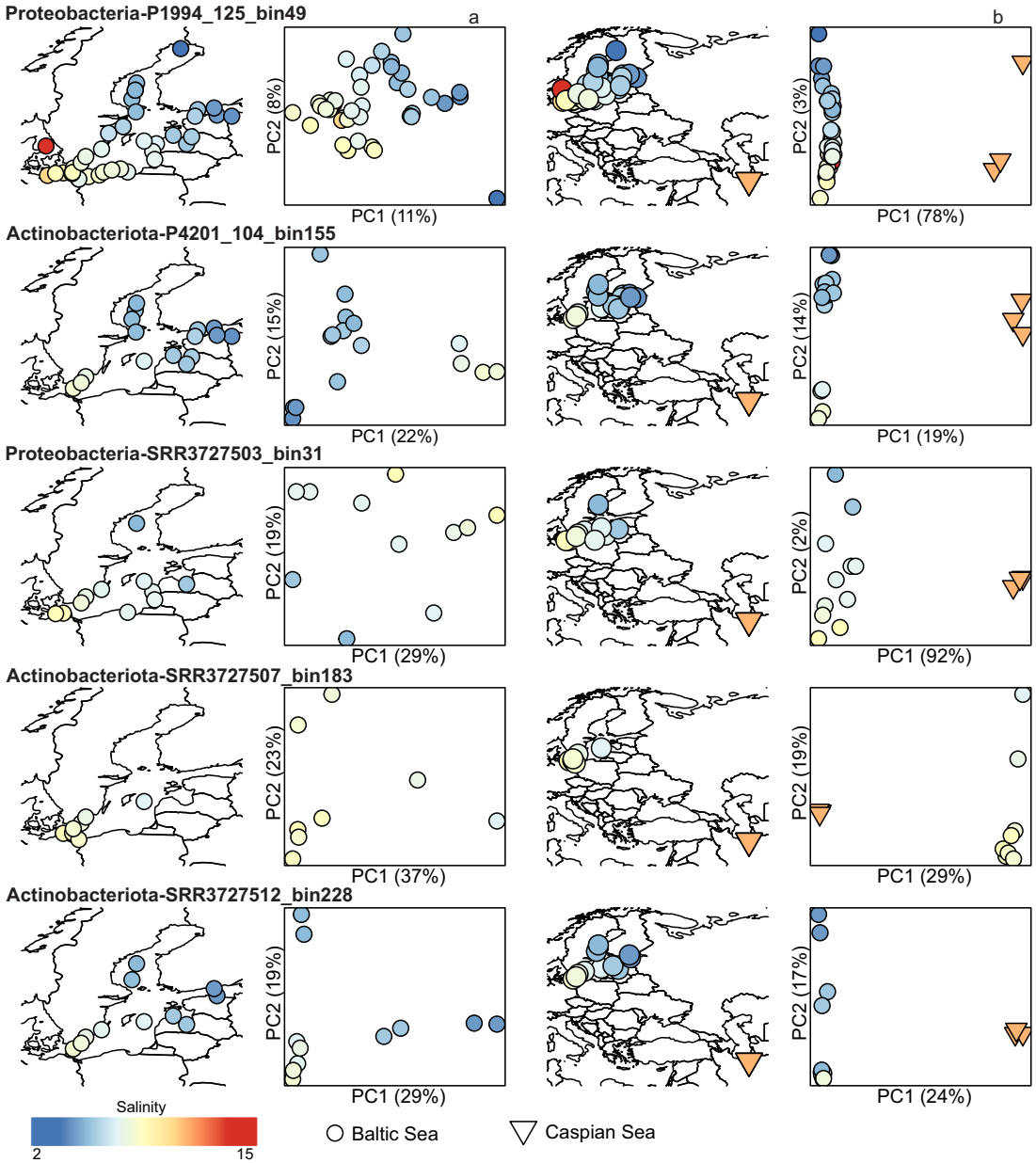


Fig. 1 | Status of shared bacterial clusters in different biomes. a. Map showing origins of MAGs [refs to papers from where we got the MAGs] **b.** Number of MAG clusters at different nucleotide distance cut-offs in each biome. **c.** Number of MAG clusters at different nucleotide distance cut-offs with members of different biome combinations. **d**. relative abundance of MAGs affiliated to different phyla in the three different biomes. Frequency distributions of pairwise inter-MAG ANI values in the range of 90 - 100% in **e.** freshwater, **f.** brackish, and **g.** marine genomes.



535
536 **Fig. 2 | Population structure of brackish genome clusters**. **a-b,** Geographic origin of each
537 metagenome sample included for the MAG is shown to the left and a PCoA based on pairwise
538 F_{ST} values to the right. Only Baltic Sea samples were included in **a**. whereas both Baltic Sea and
539 Caspian Sea samples were included in **b.**
540
541

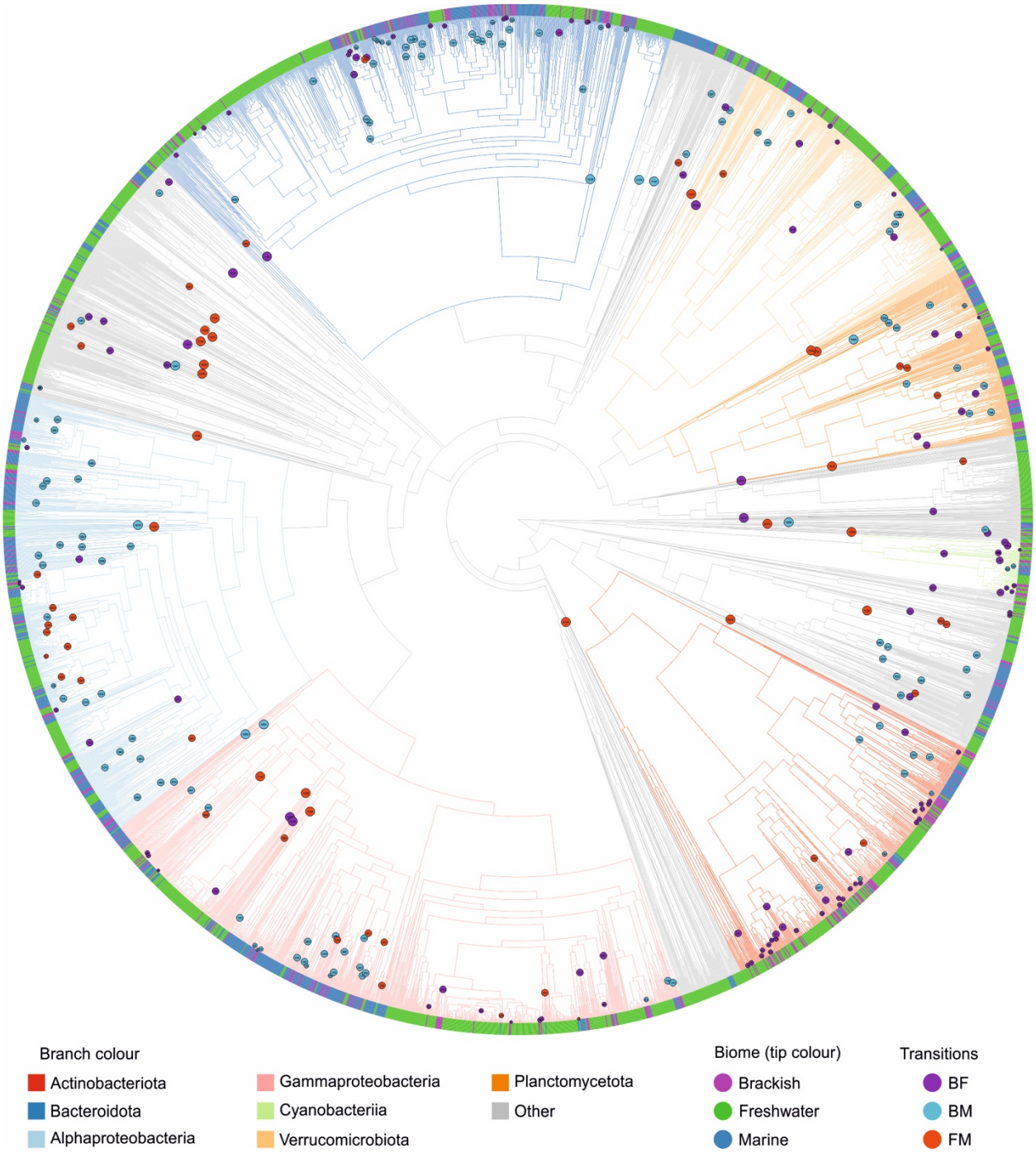


Fig. 3 | Reconstructed phylogeny of MAGs of the three aquatic biomes. Tree with branch lengths corresponding to estimated minimal times since divergence. Only tips present in the pruned tree are visualized. Identified transitions, i.e. nodes corresponding to most recent common ancestors of MSG pairs, are color-labeled by type.

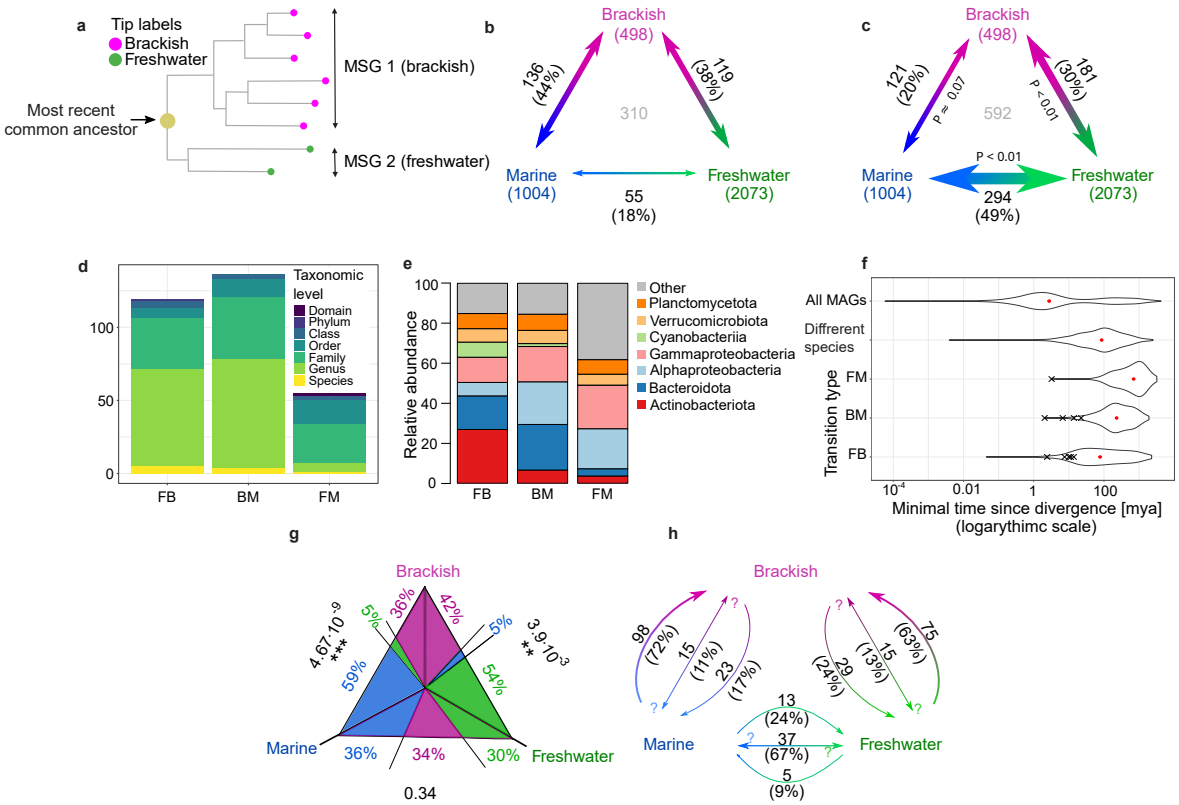


Fig. 4 | The most recent transitions between biomes. a. Illustration of an MSG pair corresponding to an FB transition **b**. Observed numbers of transitions between the biomes. Numbers of MAGs included in the pruned tree are mentioned in brackets under the biome name. The total number of all identified cross-biome transitions is shown in gray in the middle of the plot. **c.** Expected numbers of transitions between the biomes as predicted by random permutations of biome annotations. On the internal sides of the arrows are given P-values for the estimated numbers being smaller (for FM and FB) or bigger (for BM) than the observed numbers of transitions. **d.** The number of transitions distributed across the lowest shared taxonomic level to which both MSGs could be classified. If no annotation was available at a taxonomic level, a higher one was considered. **e**. Proportion of pairs of MSGs for broader taxonomic groups. **f.** The distributions of estimated minimal times since divergence for all nodes on the unpruned tree ("All MAGs"), nodes corresponding to most recent divergence events (pairs of bacterial species sharing an exclusive MRCA on the pruned tree, i.e. "Different species"), and nodes corresponding to transitions grouped by type. The timing of transitions within >95% ANI clusters is marked as "X". **g.** Mean likelihoods of ancestral biome-states for transition between the biomes. Next to the sides of the triangles are given p-values for difference in the likelihood for ancestral states of the 2 biomes in which the MSGs for the transition type were identified (for ex. between brackish and freshwater for FB). **h.** Numbers (and proportion) of transitions of each type for which the direction of the transition indicated by the arrows was more probable than the two other ancestral biome-states taken together. Both-sided arrows indicated transitions for which no dominant transition direction was identified.

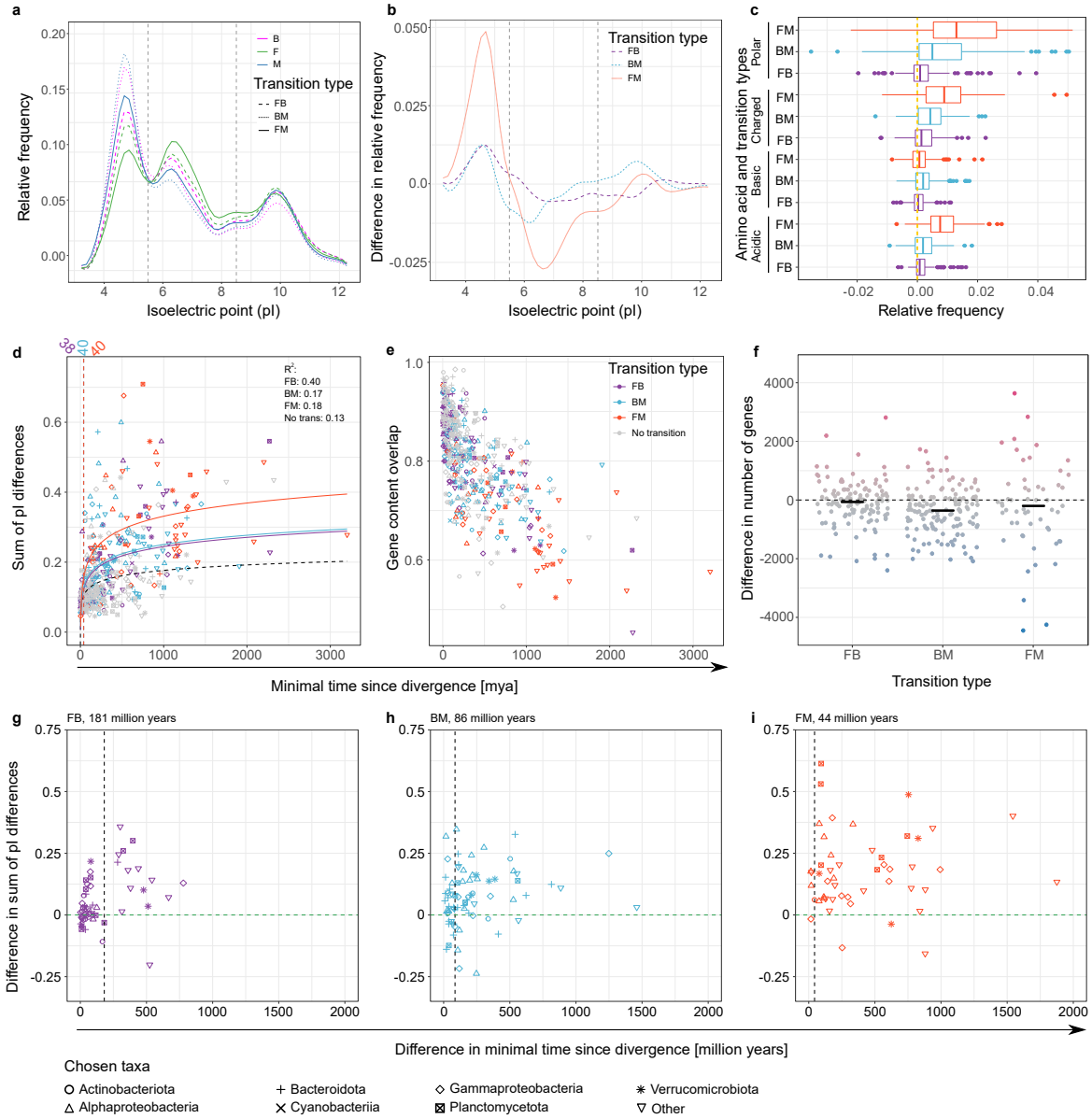


Fig. 5 | Changes in proteome properties and gene content with cross-biome transitions. a. Averaged distribution of pI values in predicted proteomes of MAGs. The figure is based on the mean frequencies of proteins within 0.5 pH wide bins (from pI ∈ [3.0, 3.5) to pI ∈ [12.0, 12.5)). Vertical lines mark pI values set to distinguish between acidic, neutral, and basic proteins. **b.** Differences in pI frequencies (as in **a.**) across transitions. **c.** Differences in fractions of the proteomes which the chosen categories of amino acids make up. The lower categories are subsets of the higher ones: polar ⊃ (charged = acidic + basic). **d.** Sum of all differences in pI frequency vs. minimal estimated divergence times, for different transition types and for no transition pairs. For no transition (same biome), the most distantly related pairs of monophyletic groups within MSGs were used. A logarithmic relationship was fitted for each transition type and no transition events. The vertical lines and values above correspond to the timepoint when the fit

for a transition and no transition events diverge significantly (no overlap in +/- standard error ranges thereafter). **e.** Overlap in gene content vs minimal time since divergence (grouped as in **d**). **f.** Differences in numbers of genes between the pairs of MSGs. The horizontal bars represent median values. **g-i.** Differences in the sum of pI differences between MSG pairs and between the biggest monophyletic groups within the MSGs formed after the transitions vs. differences in minimal times since transition and subsequent within-MSG divergence events for FB (**g**), BM (**h**), and FM (**i**). The vertical lines correspond to the highest values in mya for which the differences of sums of pI differences are not significantly higher than zero. In **b**, **c** and **f**, the value differences are presented as value in the more saline - value in the less saline biome.

Fig. 6 | Functional genes differentially present across transitions ("gain-loss maps"). Genes present in only one MSG from BM (**a**), FB (**b**), and FM (**c**) pairs of MSGs. Each column is a pair of MSGs, with taxonomic annotation marked above, and each row is an orthologous group,

601 with its KO number and a descriptive name given on the right and a broader (KO level C)
602 functional annotation marked to the left. To correct for labels inconsistent with the literature
603 reference, the annotations of K01531[45] and K07301[96] were manually curated. The category
604 "Enzymes with EC numbers" (represented by K00809) has been merged with "General function
605 prediction only". The genes are marked as present in an MSG if they were found in >0.5 of the
606 MAGs of the MSG. Only KOs found significantly (FDR adjusted $P < 0.1$) differentially present
607 across transitions in a pairwise analysis are displayed. Descriptions of genes identified in more
608 than one transition type were marked in bold. Groups of genes with at least one gene being in
609 >50% of the cases being annotated also as the other(s) (Supplementary Data S4) were marked
610 by adding #, $, ^ or * before the KO number.
611

612 ## References

613 1. Gould & Eldredge. Punctuated equilibria: an alternative to phyletic gradualism. *Models in*
614 *paleobiology*.
615 2. Louca, S. The rates of global bacterial and archaeal dispersal. *ISME J.* **16**, 159–167 (2022).
616 3. Puigbò, P., Lobkovsky, A. E., Kristensen, D. M., Wolf, Y. I. & Koonin, E. V. Genomes in
617 turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol.* **12**, 66
618 (2014).
619 4. Vos, M., Hesselman, M. C., Te Beek, T. A., van Passel, M. W. J. & Eyre-Walker, A. Rates
620 of Lateral Gene Transfer in Prokaryotes: High but Why? *Trends Microbiol.* **23**, 598–605
621 (2015).
622 5. Intergovernmental Panel on Climate Change. Coastal Systems and Low-Lying Areas. in
623 *Climate Change 2014 – Impacts, Adaptation and Vulnerability: Part A: Global and Sectoral*
624 *Aspects: Working Group II Contribution to the IPCC Fifth Assessment Report* 361–410
625 (Cambridge University Press, 2014).
626 6. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's
627 biogeochemical cycles. *Science* **320**, 1034–1039 (2008).
628 7. Ali, S. S., Yu, Y., Pfosser, M. & Wetschnig, W. Inferences of biogeographical histories
629 within subfamily Hyacinthoideae using S-DIVA and Bayesian binary MCMC analysis
630 implemented in RASP (Reconstruct Ancestral State in Phylogenies). *Ann. Bot.* **109**, 95–107

631 (2012).
632 8. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity.
633 *Nature* **551**, 457–463 (2017).
634 9. Logares, R. *et al.* Infrequent marine-freshwater transitions in the microbial world. *Trends*
635 *Microbiol.* **17**, 414–422 (2009).
636 10. Barberán, A. & Casamayor, E. O. Global phylogenetic community structure and β-diversity
637 patterns in surface bacterioplankton metacommunities. *Aquat. Microb. Ecol.* **59**, 1–10
638 (2010).
639 11. Walsh, D. A., Lafontaine, J. & Grossart, H.-P. On the Eco-Evolutionary Relationships of
640 Fresh and Salt Water Bacteria and the Role of Gene Transfer in Their Adaptation. in *Lateral*
641 *Gene Transfer in Evolution* (ed. Gophna, U.) 55–77 (Springer New York, 2013).
642 12. Hugerth, L. W. *et al.* Metagenome-assembled genomes uncover a global brackish
643 microbiome. *Genome Biol.* **16**, 279 (2015).
644 13. Delgado, L. F. & Andersson, A. F. Evaluating metagenomic assembly approaches for
645 biome-specific gene catalogues. *Microbiome* **10**, 72 (2022).
646 14. Cabello-Yeves, P. J. & Rodriguez-Valera, F. Marine-freshwater prokaryotic transitions
647 require extensive changes in the predicted proteome. *Microbiome* **7**, 117 (2019).
648 15. Dupont, C. L. *et al.* Functional tradeoffs underpin salinity-driven divergence in microbial
649 community composition. *PLoS One* **9**, e89549 (2014).
650 16. Eiler, A. *et al.* Productivity and salinity structuring of the microplankton revealed by
651 comparative freshwater metagenomics. *Environ. Microbiol.* **16**, 2682–2698 (2014).
652 17. Alneberg, J. *et al.* Ecosystem-wide metagenomic binning enables prediction of ecological
653 niches from genomes. *Commun Biol* **3**, 119 (2020).
654 18. Linz, A. M. *et al.* Freshwater carbon and nutrient cycles revealed through reconstructed
655 population genomes. *PeerJ* **6**, e6075 (2018).
656 19. Cabello-Yeves, P. J. *et al.* Genomes of Novel Microbial Lineages Assembled from the Sub-

657 Ice Waters of Lake Baikal. *Appl. Environ. Microbiol.* **84**, (2018).
658 20. Buck, M. *et al.* Comprehensive dataset of shotgun metagenomes from oxygen stratified
659 freshwater lakes and ponds. *Sci Data* **8**, 131 (2021).
660 21. Mehrshad, M., Amoozegar, M. A., Ghai, R., Shahzadeh Fazeli, S. A. & Rodriguez-Valera,
661 F. Genome Reconstruction from Metagenomic Data Sets Reveals Novel Microbes in the
662 Brackish Waters of the Caspian Sea. *Appl. Environ. Microbiol.* **82**, 1599–1612 (2016).
663 22. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft
664 metagenome-assembled genomes from the global oceans. *Sci Data* **5**, 170203 (2018).
665 23. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are
666 abundant in surface ocean metagenomes. *Nat Microbiol* **3**, 804–813 (2018).
667 24. Varghese, N. J. *et al.* Microbial species delineation using whole genome sequences.
668 *Nucleic Acids Res.* **43**, 6761–6771 (2015).
669 25. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat.*
670 *Biotechnol.* **38**, 1079–1086 (2020).
671 26. Sjöqvist, C., Delgado, L. F., Alneberg, J. & Andersson, A. F. Ecologically coherent
672 population structure of uncultivated bacterioplankton. *ISME J.* **15**, 3034–3049 (2021).
673 27. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify
674 genomes with the Genome Taxonomy Database. *Bioinformatics* (2019)
675 doi:10.1093/bioinformatics/btz848.
676 28. Battistuzzi, F. U., Tao, Q., Jones, L., Tamura, K. & Kumar, S. RelTime Relaxes the Strict
677 Molecular Clock throughout the Phylogeny. *Genome Biol. Evol.* **10**, 1631–1636 (2018).
678 29. Kim, H., Lee, S. & Jang, Y. Macroevolutionary patterns in the Aphidini aphids (Hemiptera:
679 Aphididae): diversification, host association, and biogeographic origins. *PLoS One* **6**,
680 e24749 (2011).
681 30. Moran, N. A., Munson, M. A., Baumann, P. & Ishikawa, H. A molecular clock in
682 endosymbiotic bacteria is calibrated using the insect hosts. *Proceedings of the Royal*

683 *Society of London. Series B: Biological Sciences* **253**, 167–171 (1993).
684 31. Thao, M. L. *et al.* Cospeciation of psyllids and their primary prokaryotic endosymbionts.
685 *Appl. Environ. Microbiol.* **66**, 2898–2905 (2000).
686 32. Cockerell, T. D. A. The Third Fossil Tsetse-Fly. *Nature* **98**, 70–70 (1916).
687 33. Lo, N., Bandi, C., Watanabe, H., Nalepa, C. & Beninati, T. Evidence for cocladogenesis
688 between diverse dictyopteran lineages and their intracellular endosymbionts. *Mol. Biol.*
689 *Evol.* **20**, 907–913 (2003).
690 34. Thorne, B. L., Grimaldi, D. A. & Krishna, K. Early Fossil History of the Termites. in *Termites:*
691 *Evolution, Sociality, Symbioses, Ecology* (eds. Abe, T., Bignell, D. E. & Higashi, M.) 77–93
692 (Springer Netherlands, 2000).
693 35. Moran, N. A., Tran, P. & Gerardo, N. M. Symbiosis and insect diversification: an ancient
694 symbiont of sap-feeding insects from the bacterial phylum Bacteroidetes. *Appl. Environ.*
695 *Microbiol.* **71**, 8802–8810 (2005).
696 36. Shcherbakov, D. The 270 million year history of Auchenorrhyncha (Homoptera).
697 https://www.academia.edu/download/33480266/Denisia__Biologiezentrum_Linz__Austria_
698 Vol_0004-0029-0036.pdf.
699 37. Kuo, C.-H. & Ochman, H. Inferring clocks when lacking rocks: the variable rates of
700 molecular evolution in bacteria. *Biol. Direct* **4**, 35 (2009).
701 38. Björck, S. A review of the history of the Baltic Sea, 13.0-8.0 ka BP. *Quat. Int.* **27**, 19–40
702 (1995).
703 39. Dumont, H. J. The Caspian Lake: History, biota, structure, and function. *Limnol. Oceanogr.*
704 **43**, 44–52 (1998).
705 40. Morrison, K. L. & Weiss, G. A. Combinatorial alanine-scanning. *Curr. Opin. Chem. Biol.* **5**,
706 302–307 (2001).
707 41. Bossemeyer, D. *et al.* K+-transport Protein TrkA of Escherichia coli Is a Peripheral
708 Membrane Protein That Requires other trk Gene Products for Attachment to the

709 Cytoplasmic Membrane*. *J. Biol. Chem.* **264**, 16403–16410 (1989).
710 42. Holtmann, G., Bakker, E. P., Uozumi, N. & Bremer, E. KtrAB and KtrCD: two K+ uptake
711 systems in Bacillus subtilis and their role in adaptation to hypertonicity. *J. Bacteriol.* **185**,
712 1289–1298 (2003).
713 43. Zhang, H. *et al.* Repeated evolutionary transitions of flavobacteria from marine to non-
714 marine habitats. *Environ. Microbiol.* **21**, 648–666 (2019).
715 44. Levina, N. *et al.* Protection of Escherichia coli cells against extreme turgor by activation of
716 MscS and MscL mechanosensitive channels: identification of genes required for MscS
717 activity. *EMBO J.* **18**, 1730–1737 (1999).
718 45. Tao, T., Snavely, M. D., Farr, S. G. & Maguire, M. E. Magnesium transport in Salmonella
719 typhimurium: mgtA encodes a P-type ATPase and is regulated by Mg2+ in a manner similar
720 to that of the mgtB P-type ATPase. *J. Bacteriol.* **177**, 2654–2662 (1995).
721 46. Papp, K. M. & Maguire, M. E. The CorA Mg2+ transporter does not transport Fe2+. *J.*
722 *Bacteriol.* **186**, 7653–7658 (2004).
723 47. Hsu, J.-L., Chen, H.-C., Peng, H.-L. & Chang, H.-Y. Characterization of the Histidine-
724 containing Phosphotransfer Protein B-mediated Multistep Phosphorelay System in
725 Pseudomonas aeruginosa PAO1*. *J. Biol. Chem.* **283**, 9933–9944 (2008).
726 48. Wösten, M. M. S. M., Wagenaar, J. A. & van Putten, J. P. M. The FlgS/FlgR Two-
727 component Signal Transduction System Regulates the fla Regulon in Campylobacter
728 jejuni*. *J. Biol. Chem.* **279**, 16214–16222 (2004).
729 49. Futamata, H. *et al.* Chemotactic response to amino acids of fluorescent pseudomonads
730 isolated from spinach roots grown in soils with different salinity levels. *Soil Sci. Plant Nutr.*
731 **44**, 1–7 (1998).
732 50. Larsen, M. H., Blackburn, N., Larsen, J. L. & Olsen, J. E. Influences of temperature, salinity
733 and starvation on the motility and chemotactic response of Vibrio anguillarum. *Microbiology*
734 **150**, 1283–1290 (2004).

735 51. Xu, X. *et al.* cheA, cheB, cheR, cheV, and cheY Are Involved in Regulating the Adhesion of
736 Vibrio harveyi. *Front. Cell. Infect. Microbiol.* **10**, 591751 (2020).
737 52. Eiler, A. *et al.* Tuning fresh: radiation through rewiring of central metabolism in streamlined
738 bacteria. *ISME J.* **10**, 1902–1914 (2016).
739 53. Ramachandran, A., McLatchie, S. & Walsh, D. A. A Novel Freshwater to Marine
740 Evolutionary Transition Revealed within Methylophilaceae Bacteria from the Arctic Ocean.
741 *MBio* **12**, e0130621 (2021).
742 54. Snoeijs-Leijonmalm, P. & Andrén, E. Why is the Baltic Sea so special to live in? in
743 *Biological Oceanography of the Baltic Sea* (eds. Snoeijs-Leijonmalm, P., Schubert, H. &
744 Radziejewska, T.) 23–84 (Springer Netherlands, 2017).
745 55. Gleick, P. H. *Water in Crisis: A Guide to the World's Fresh Water Resources*. (Oxford
746 University Press, 1993).
747 56. Garrison, T. S. *Oceanography: An Invitation to Marine Science*. (Cengage Learning, 2012).
748 57. Louca, S. *et al.* Bacterial diversification through geological time. *Nat Ecol Evol* **2**, 1458–
749 1467 (2018).
750 58. Cohan, F. M. Bacterial species and speciation. *Syst. Biol.* **50**, 513–524 (2001).
751 59. Zaremba-Niedzwiedzka, K. *et al.* Single-cell genomics reveal low recombination
752 frequencies in freshwater bacteria of the SAR11 clade. *Genome Biol.* **14**, R130 (2013).
753 60. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98 (1973).
754 61. Kiraga, J. *et al.* The relationships between the isoelectric point and: length of proteins,
755 taxonomy and ecology of organisms. *BMC Genomics* **8**, 163 (2007).
756 62. Lanyi, J. K. Salt-dependent properties of proteins from extremely halophilic bacteria.
757 *Bacteriol. Rev.* **38**, 272–290 (1974).
758 63. Mevarech, M., Frolow, F. & Gloss, L. M. Halophilic enzymes: proteins with a grain of salt.
759 *Biophys. Chem.* **86**, 155–164 (2000).
760 64. Carstensen, J. & Duarte, C. M. Drivers of pH Variability in Coastal Ecosystems. *Environ.*

761 *Sci. Technol.* **53**, 4020–4029 (2019).

762 65. Slonczewski, J. L., Fujisawa, M., Dopson, M. & Krulwich, T. A. Cytoplasmic pH
763 Measurement and Homeostasis in Bacteria and Archaea. in *Advances in Microbial*
764 *Physiology* (ed. Poole, R. K.) vol. 55 1–317 (Academic Press, 2009).

765 66. Schneider, B. *et al.* Biogeochemical cycles. in *Biological Oceanography of the Baltic Sea*
766 (eds. Snoeijs-Leijonmalm, P., Schubert, H. & Radziejewska, T.) 87–122 (Springer
767 Netherlands, 2017).

768 67. Kirsch, M. L., Peters, P. D., Hanlon, D. W., Kirby, J. R. & Ordal, G. W. Chemotactic
769 methylesterase promotes adaptation to high concentrations of attractant in Bacillus subtilis.
770 *J. Biol. Chem.* **268**, 18610–18616 (1993).

771 68. Collard, F. *et al.* Molecular Identification of N-Acetylaspartylglutamate Synthase and β-
772 Citrylglutamate Synthase*. *J. Biol. Chem.* **285**, 29826–29833 (2010).

773 69. Li, H., Xu, H., Graham, D. E. & White, R. H. Glutathione synthetase homologs encode α-l-
774 glutamate ligases for methanogenic coenzyme F_{420} and tetrahydrosarcinapterin
775 biosyntheses. *Proceedings of the National Academy of Sciences* **100**, 9785–9790 (2003).

776 70. Kino Kuniki, Arai Toshinobu & Arimura Yasuhiro. Poly-α-Glutamic Acid Synthesis Using a
777 Novel Catalytic Activity of RimK from Escherichia coli K-12. *Appl. Environ. Microbiol.* **77**,
778 2019–2025 (2011).

779 71. Vigil-Stenman, T., Ininbergs, K., Bergman, B. & Ekman, M. High abundance and
780 expression of transposases in bacteria from the Baltic Sea. *ISME J.* **11**, 2611–2623 (2017).

781 72. Rankin, D. J., Rocha, E. P. C. & Brown, S. P. What traits are carried on mobile genetic
782 elements, and why? *Heredity* **106**, 1–10 (2011).

783 73. Koonin, E. V. & Wolf, Y. I. Genomics of bacteria and archaea: the emerging dynamic view
784 of the prokaryotic world. *Nucleic Acids Res.* **36**, 6688–6719 (2008).

785 74. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life.
786 *Nat. Rev. Genet.* **16**, 472–482 (2015).

787 75. Iranzo, J., Wolf, Y. I., Koonin, E. V. & Sela, I. Gene gain and loss push prokaryotes beyond
788 the homologous recombination barrier and accelerate genome sequence divergence. *Nat.*
789 *Commun.* **10**, 5376 (2019).
790 76. Hoetzinger, M. & Hahn, M. W. Genomic divergence and cohesion in a species of pelagic
791 freshwater bacteria. *BMC Genomics* **18**, 794 (2017).
792 77. Hanage, W. P., Spratt, B. G., Turner, K. M. E. & Fraser, C. Modelling bacterial speciation.
793 *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 2039–2044 (2006).
794 78. Marttinen, P., Croucher, N. J., Gutmann, M. U., Corander, J. & Hanage, W. P.
795 Recombination produces coherent bacterial species clusters in both core and accessory
796 genomes. *Microb Genom* **1**, e000038 (2015).
797 79. Whitaker, R. J., Grogan, D. W. & Taylor, J. W. Recombination shapes the natural
798 population structure of the hyperthermophilic archaeon Sulfolobus islandicus. *Mol. Biol.*
799 *Evol.* **22**, 2354–2361 (2005).
800 80. Ives, A. R. & Garland, T., Jr. Phylogenetic logistic regression for binary dependent
801 variables. *Syst. Biol.* **59**, 9–26 (2010).
802 81. Sunagawa, S. *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev.*
803 *Microbiol.* **18**, 428–445 (2020).
804 82. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic
805 through eastern tropical Pacific. *PLoS Biol.* **5**, e77 (2007).
806 83. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM:
807 assessing the quality of microbial genomes recovered from isolates, single cells, and
808 metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
809 84. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High
810 throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.
811 *Nat. Commun.* **9**, 5114 (2018).
812 85. Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between

813 domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).
814 86. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
815 identification. *BMC Bioinformatics* **11**, 119 (2010).
816 87. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology
817 Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
818 88. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically
819 annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids*
820 *Res.* **47**, D309–D314 (2019).
821 89. Coleman, G. A. *et al.* A rooted phylogeny resolves early bacterial evolution. *Science* **372**,
822 (2021).
823 90. Pagel, M. Detecting correlated evolution on phylogenies: a general method for the
824 comparative analysis of discrete characters. *Proceedings of the Royal Society of London.*
825 *Series B: Biological Sciences* **255**, 37–45 (1994).
826 91. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open
827 Software Suite. *Trends Genet.* **16**, 276–277 (2000).
828 92. Lenth, R. V. Least-Squares Means: The R Package lsmeans. *J. Stat. Softw.* **69**, 1–33
829 (2016).
830 93. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1**, 80–83
831 (1945).
832 94. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful
833 approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
834 95. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference
835 resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–62 (2016).
836 96. Besserer, G. M., Nicoll, D. A., Abramson, J. & Philipson, K. D. Characterization and
837 Purification of a Na+/Ca2+ Exchanger from an Archaebacterium*. *J. Biol. Chem.* **287**,
838 8652–8659 (2012).