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 their 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 patterns of change accompanying bacterial transitions between aquatic biomes remain unresolved, with regards to both functional gene repertoires and proteome properties.

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 (Extended Data 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.

Fortyfive of the 485 brackish genome clusters had members from both included brackish water bodies; the Baltic and the Caspian Sea, while 300 and 141 were unique to Baltic and Caspian, respectively. The shared clusters belonged to 13 classes in 10 different phyla

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

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

124 We reconstructed the phylogeny of the aquatic MAGs using a set of conserved housekeeping
125 genes[27]. The resulting tree was pruned to contain a maximum of one MAG from each biome per
126 previously identified genome cluster (i.e. ~species) (Fig. 3). The most recent transitions between
127 the aquatic biomes, if detected, should be represented on the phylogenetic tree as pairs of closest
128 related monophyletic and biome-specific groups. We call these clades monobiomic sister groups
129 (MSGs, Fig. 4a). We identified 310 MSG pairs, with brackish and marine (BM; n=136) and
130 freshwater and brackish (FB; n=119) transitions being twice as numerous as freshwater and
131 marine (FM; n=55), despite the larger number of freshwater and marine MAGs in the pruned tree
132 (Fig. 4b). The number of FB and FM transitions were also significantly lower than would be
133 expected if phylogeny would not correlate with biome annotation (Fig. 4c). This corroborates
134 observations for the previously described genome clustering.
135 Most of the identified FB and BM MSG pairs belonged to the same genus (around 55% in
136 both cases), while for FM transitions the lowest-level shared taxonomic annotation was most often
137 (49%) at the family level (Fig. 4d, Extended Data Table 2). The taxonomic distribution of MSG

138 pairs followed the overall distribution in the corresponding biomes for BM and FM, but not for FB,
139 probably due to the overrepresentation of Actinobacteriota and Cyanobacteriia in this transition
140 type (Fig. 4e, Extended Data Fig. 1).

141 We estimated the minimal time since divergence of the MSG pairs based on phylogenetic
142 distances, using a method which allows for variable evolutionary rates[28] (Fig. 3, Fig.4f, Extended
143 Data Fig. 2). As time constraints, we used minimal times since divergence of genomes of
144 endosymbiotic bacteria known from fossil records of their hosts[29–37] (Supplementary Table S1).
145 As obligate endosymbionts have high evolutionary rates that exceed those of free-living
146 bacteria[37], this model gives conservative estimates of time since the transitions started. The timing
147 differed depending on the transition type (Fig. 4f; $P < 10^{-6}$ for all comparisons), FB being the most
148 recent and FM the oldest. The difference in median times since FB and BM transitions could not
149 be attributed to the uneven representation of biomes in the dataset (Extended Data Fig. 2b).

150 The most recent transition was estimated to have happened at least 44.8 kya, and all other
151 transitions millions of years ago (Supplementary Table S2). Therefore, all the detected transitions
152 started long before the current brackish conditions were established in the Baltic Sea (8 kya[38])
153 and only the six most recent ones to the period since the Caspian has continuously been brackish
154 (2-3 mya[39]). These results imply that the evolutionary history of the global brackish microbiome[12]
155 by far predates the formation of these two brackish basins.

156 To infer the directions of the biome transitions, we conducted ancestral (biome) state
157 reconstructions of the most recent common ancestors (MRCAs) of the MSG pairs, correcting for
158 the uneven representation of biomes in the pruned tree (see Methods). FB and BM transitions
159 were generally more likely to happen into than out of the brackish biome, while for FM transitions
160 no significant direction preference was observed (Fig. 4g-h, Extended Data Fig. 3 and 4).

161

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

163 We next assessed if changes in proteome properties accompany the biome transitions. The
164 average distributions of isoelectric points (pIs) of the proteins encoded in the genomes were
165 compared between MSGs in a pairwise manner. Significant differences in the proportion of neutral
166 proteins were observed for all the transition types (Fig. 5a-b, Supplementary Data S3), with a
167 higher proportion in the biomes of lower salinity. The frequencies of acidic proteins showed the
168 opposite pattern and were significantly different for the comparisons that included freshwater
169 MSGs, but not for the BM transition type. The results suggest that the higher representation of
170 neutral pIs observed in freshwater bacteria[14] gradually diminishes with increasing salinity while
171 low proportion of acidic proteins is a distinctive feature of freshwater bacteria. However, the extent

172 of transition-related changes in pI distribution highly depended on taxonomy, with particularly
173 small changes for Actinobacteria (Extended Data Fig. 5).
174 Polar, charged, and acidic amino acid frequencies broadly increased with salinity across
175 the MSG pairs (Fig 5c, Supplementary Data S3). Changes were most significant for acidic amino
176 acids, which is a subcategory of the other two categories and may thus also explain the changes
177 observed for those. In contrast, the frequency of basic amino acids differed significantly only for
178 BM transitions.
179 Among the acidic amino acids, the difference for glutamate was larger and more significant
180 than for aspartate across all the transition types (Extended Data Fig. 6, Supplementary Data S3).
181 For BM transitions, the higher proportion of basic residues in marine proteomes could be
182 attributed to lysine. Big and significant differences in the frequency of alanine across BM and FM
183 transitions suggest that this non-polar amino acid with a relatively "neutral" effect on protein
184 structure[40] is often substituted by charged residues in transitions to higher salinity. For all
185 transition types, there was significantly less proline in the predicted proteomes at higher salinity.
186 To investigate the dynamics in how the proteome is reshaped after biome transitions, we
187 compared the differences in pI distributions to inferred times since divergence. The overall change
188 in pI distributions followed different logarithmic relationships for within-MSG (no transition)
189 divergence times as compared to any of the transition types (Fig. 5d). Moreover, the differences
190 in pI distribution across transitions were consistently larger than changes resulting from within-
191 MSG (no transition) divergence events when the latter were at least 44 to 181 million years more
192 recent (Fig. 5g-i). This suggests that proteomes are under strong selection and continue to evolve
193 to match the new salinity conditions long after the switch to a new biome (Extended Data Fig. 7).
194
195
196 **Specific changes in gene content accompany cross-biome transitions**
197 For each transition type, we were able to identify specific functional genes (KEGG orthology
198 groups [KOs]), significantly over-represented in the MSGs of one of the compared biomes after
199 correcting for differences in estimated completeness of the MAGs (Fig. 6, Supplementary Data
200 S4). However, transitions did not change the gain and loss rates (Fig 5e), even though gene
201 numbers significantly differed for BM transitions (Fig. 5f, Supplementary Data S3), with brackish
202 MAGs having a median of 386 more genes. Therefore, the gains or losses of the same gene
203 functions (KOs) we observed in multiple independent transition events most likely result from
204 analogous changes in selective pressures caused by the biome transitions.

205 Transporters were the only functional category (KO level C) over-represented in all
206 transition types (Fig. 6, Extended Data Figure S8), some with known roles in adaptation to altered
207 salinity. For FB transitions, brackish MAGs more frequently contained components of two (or of
208 the same[41]) potassium uptake systems (K03498 and K03499) co-acting to allow growth in higher
209 osmotic pressure[42]. A component of the uptake system with higher affinity, K03499[42], was also
210 more prevalent in marine genomes across FM transitions. Additionally, for both FB and FM MSG
211 pairs, a Na^+/Ca^{2+} antiporter (K07301) was more often found in the more saline biomes. Analogous
212 changes in presence of the same K^+ uptake systems and of Na^+/Ca^{2+} antiporters were reported
213 in a comparison of terrestrial and marine *Flavobacteriaceae*[43]. Finally, K03282, more frequently
214 present in brackish across BM transitions, is a mechanoreceptor involved in managing
215 hypoosmotic stress[44].
216 Two distinct magnesium uptake systems[45,46] were more frequently present in lower-salinity
217 biomes for FB and BM MSG pairs (K01531 and K03284, respectively). Moreover, for BM
218 transitions, brackish MAGs displayed a higher presence of chemotaxis (K03413) and two-
219 component system sensors (K02482 and K20974) involved in regulation of motility[47,48]. While
220 Mg^{2+} is a major component of sea salt[11] and salinity can alter chemotactic responses and
221 motility[49–51], to the best of our knowledge, there are no previous reports on the eco-evolutionary
222 relevance of these genes. For FB, a gene connected to either pigment synthesis (K15777) or
223 aromatic hydrocarbon degradation (K04100) was more often found in freshwater MAGs. Both of
224 these functions may be connected to salinity-related changes in dissolved organic carbon
225 concentration (see Supplementary Discussion for more details).
226 Other differentially present genes were less directly connected to physicochemical
227 differences between the biomes, suggesting more complex and/or cryptic biological mechanisms
228 associated with transitions. For BM transitions, genes connected to various metabolic pathways
229 (including modifications of charged amino acids), as well as transposases, antitoxins, and
230 competence proteins, were more often present in brackish MAGs than their marine counterparts.
231 Typical housekeeping genes were absent among the differentially present genes, which verifies
232 the specificity of the method.
233

234 Discussion

235 In this study, we combined several large-scale datasets, which allowed us to comparatively
236 analyze bacterial genomes from freshwater, brackish and marine biomes at an unprecedented
237 scale. Including brackish genomes allowed for a more detailed analysis of the evolutionary history
238 of aquatic bacteria than in previous studies focusing on the two biomes at the extremes of the

239 spectrum[9–11,16,43,52]. The brackish biome is a significant part of our global waters and combined,
240 the two major brackish systems analyzed here have a similar volume (~99,000 km^3 [53]) as all
241 freshwater lakes on earth (≤102,000 km^3 [54]), albeit only a fraction of the volume of the global
242 ocean (~1.35 billion km^3 [55]).

243 We observed bacterial communities from the different biomes to be separated on at least
244 the species level - nearly all >95% ANI genome clusters were restricted to a single biome.
245 Brackish genome clusters often had representatives from both the Baltic and Caspian Sea. These
246 results provide further support that the brackish biome hosts a distinct, globally distributed set of
247 bacterial species[12]. However, the fact that populations of individual species were more genetically
248 differentiated between the two brackish seas than across the strong salinity gradient within the
249 Baltic Sea, indicates that geographical barriers still have a structuring effect on the distribution of
250 aquatic bacteria.

251 Our analyses support previous concepts of freshwater-marine transitions being rare[9–11]
252 and show more common but still rare freshwater-brackish and brackish-marine transitions, which
253 most often are directed towards the brackish biome. However, as the three biomes have
254 continuously existed throughout geological history and bacterial species appear to be globally
255 distributed within each of them[2], one might ask why any recent transitions have even happened.
256 A plausible scenario may be that genetic innovation (gene gain/loss or mutation) can make a
257 lineage competitive in a different biome, either by opening up a new niche or out-competing others
258 in an existing one, in accordance with the stable ecotype model[56]. However, the benefits of the
259 innovation need to outweigh the decrease in fitness due to switching to a new salinity regime.
260 Thus, the smaller salinity difference between the brackish and the other biomes would explain
261 why it is associated with more transitions.

262 The more frequent transitions into than out of the brackish environment are harder to
263 explain. For FB, the reason could simply be hydrological, i.e. that freshwater usually flows into
264 brackish water rather than the other way around. For BM, differences in population sizes may
265 instead be the explanation, as both the volume and surface of the ocean are much bigger than of
266 all brackish waters. Beneficial mutations are thus more likely to appear among the more numerous
267 marine bacteria. As the formation of neither the Baltic nor the Caspian Sea caused an observable
268 increase in transition rates, which is consistent with previous notions of major geological events
269 having little impact on bacterial speciation rates[57], formation of new brackish habitats does not
270 seem to play a key role in promoting cross-biome transitions.

271 Our results corroborate previous reports on differences in predicted proteomes at different
272 salinities[14], and support them with a systematic analysis across the bacterial tree of life. These

273 changes likely occur over longer evolutionary timescales after the actual transition, meaning that
274 in the period after the transitions, bacteria may have a suboptimal protein repertoire in their new
275 environment. This environmental mismatch would potentially lower the fitness of newly
276 transitioned lineages in addition to the bottleneck effects, traces of which are present in LD12
277 clade genomes[58], and which can lead to fixation of slightly deleterious mutations[59].

278 We observed a shift towards more acidic protein pIs at higher salinity, also reported
279 before[14,60], and showed the underlying proteome-scale increase in proportion of acidic amino
280 acids, most notably glutamate. This likely reflects that charged amino acids increase protein
281 solubility at higher ionic strength[61,62]. However, an even stronger pattern observed was the
282 proportion of neutral proteins gradually diminishing with increasing salinity, which is also known
283 from previous studies but lacks explanation[14]. We hypothesize that this effect might be due to
284 diminishing extracellular pH variation as the concentration of buffering salts increases[11,63].
285 Moreover, Na^+-coupled transport is the key mechanism of pH homeostasis under alkaline
286 conditions[64], to which both freshwater[11] and brackish environments[63,65] transcend. Consistently,
287 lower NaCl concentrations have been connected with higher intracellular pH variation in alkaline
288 conditions[64]. Thus, neutral proteins could be soluble and active outside as well as inside the cell
289 under pH variations in less saline environments. Unlike freshwater, brackish pH rarely changes
290 to an acidic state[63,65], which may explain why the difference in frequency of acidic proteins was
291 less significant for BM than for other transition types, accompanied by a change in the frequency
292 of not only acidic but also basic amino acids.

293 For BM transitions, we identified genes involved in glutamate post-translational
294 modifications (K03412, K05844, K14940, K18310) to be more frequent in brackish genomes. All
295 these modifications increase the acidity of residues[66–69]. Thus, depending on the specificity of the
296 enzymes, brackish bacteria may change the charges (and pIs) of their proteins to adjust to
297 different salinities (and varying pH). We also observed that brackish bacteria were enriched in
298 gene functions for other physiological responses to environmental cues, including chemotaxis,
299 transcriptional regulation, as well as synthesis of trehalose and polyamines (both osmolytes) (see
300 Supplementary Discussion). Thus, phenotypic plasticity might be an important hallmark of
301 brackish bacteria, sustaining the populations across the physicochemical gradients typical for
302 many brackish environments.

303 BM transitions seem to expose bacteria to a different set of mobile genetic elements
304 (MGEs). This is suggested by a set of transposases (previously reported to be unusually abundant
305 in the Baltic Sea[70]), competence proteins, and, indirectly, antitoxins[71] being more often present in
306 the brackish MAGs (Fig. 6a). Consequently, more MGEs in the brackish MAGs could explain the

307 higher numbers of genes as compared to the marine relatives (Fig. 5f). MGEs tend to increase
308 gene gain and loss processes[72,73], and may thus facilitate fast adaptation to new physiochemical
309 conditions. It also promotes the establishment of a recombination barrier between biomes[74]. As
310 homologous recombination inhibits divergence[74–78], these effects would give a further mechanistic
311 explanation for the "species-level" separation.
312 Ultimately, we extend the view of transitions between aquatic biomes and highlight the
313 brackish environment, showing trends specific for certain transition types and common across all
314 of them. We report patterns prevalent across the bacterial tree of life, providing supplementary
315 materials that allow investigation of the changes for chosen taxonomic groups. We observed
316 many changes related to previously undescribed adaptation mechanisms and/or poorly
317 characterized genes, both of which call for further exploration. In the future, phylogenomic
318 analyses should be supplemented with experimental and ecological approaches. This would allow
319 examining the relevance of the unraveled adaptation mechanisms in the context of ecosystem
320 functioning and physiology.
321
322
323 # Methods
324
325 ## MAG collection and sources
326 Representative metagenome-assembled genomes (MAGs) from freshwater, brackish, and
327 marine biomes were collected from seven previous studies as specified in Extended Data Table
328 1. Freshwater MAGs originate from two North American lakes (Lake Mendota and Trout Bog), 44
329 lakes of the StratfreshDB (mainly in Sweden, Finland and Canada), and Lake Baikal (Russia)[18–
330 20]. Brackish MAGs originate from the Baltic Sea[17] and the Caspian Sea[21]. Finally, marine MAGs
331 that were published by Tully et al.[22] and Delmont et al.[23] originate from the Tara Oceans
332 Survey[79,80].
333
334 ## MAG statistics, ANI calculations and classification
335 Completeness and contamination of the collected MAGs were assessed using CheckM[81]. MAGs
336 with ≥75% completeness and ≤5% contamination were considered for further analyses. Pairwise
337 Average Nucleotide Identity (ANI) between all MAGs was calculated using FastANI (V1.2) and
338 MAGs were clustered at the threshold of 95% [82]. Clustering of MAGs was done using hclust and
339 cuttree functions in Rstudio. To remove redundant information and reduce the complexity of the
340 phylogenetic tree we kept one random MAG per cluster per biome to generate a pruned tree.

341 Taxonomic affiliation of MAGs was assigned using the "classify_wf" workflow of the Genome
342 Taxonomy Database Toolkit (GTDB-tk) (v 0.3.2)[25,27]. To compare proportion of inter-biome and
343 brackish inter-basin clusters, we used Fisher's exact test conditioned on the lower number of
344 clusters (2 categories: the basin/biome with a lower number of clusters, and the number of shared
345 clusters).
346

347 **Population genomics**

348 Population genomics was conducted with the POGENOM software as described in Sjöqvist et
349 al.[26]. The same set of 22 MAGs (from[17]) were used as references for mapping of metagenomic
350 data from the Baltic[17] and Caspian[21] seas. Only five ese obtained sufficient coverage in the three
351 Caspian Sea samples and were included in the downstream analysis. The five MAGs all represent
352 different (>95% ANI) genome clusters. The Input_POGENOM pipeline was used for the automatic
353 generation of input files for POGENOM., The pairwise F_{ST} values output from POGENOM were
354 used to conduct Principal Coordinates Analysis (PCoA) using the Vegan package in *R*. Maps
355 were plotted with the rworldmap package. For parameter settings of Input_POGENOM and
356 POGENOM, see Sjöqvist et al.[26].
357

358 **Phylogeny reconstruction**

359 The "de_novo_wf" function of GTDB-tk (v 0.3.2) was used to reconstruct the phylogeny of
360 MAGs[25,27]. The WAG (Whelan and Goldman) amino acid evolution model was selected in the
361 infer step and phylum Patescibacteria was selected as the outgroup[83]. Most downstream
362 phylogenetic analyses were performed in RStudio (V3.6.2) with the R packages ape, phytools,
363 phangorn and ggplot. The multi2di function in R was used to resolve multichotomies (that appear
364 because some MAGs were nearly identical) in the phylogenetic tree. Since this function rendered
365 some branches with zero length, we arbitrarily changed all zero-length branches to 5×10^{-4} (which
366 was the smallest length of all other branches) to ensure all branch lengths were > 0.
367

368 **Gene prediction and annotation**

369 Prodigal (v2.6.3) was run with default settings to identify protein-coding genes[84]. Gene function
370 annotation was performed with eggNOG-mapper (v2.0.0), based on the eggNOG v5.0
371 database[85,86]. A custom Python script was used to count the number of occurrences of each Kegg
372 Ortholog (KO) in each annotated MAG.
373

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

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

382

383 **Expected number of transitions**

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

390

391 **Minimal time since divergence**

392 Six divergence-time constraints were set as minimal estimates of time since divergence for pairs
393 of obligate endosymbiont taxa based on fossil record of host species[29–37] (Supplementary Table
394 S1, Supplementary Data S2). Divergence times were calculated using RelTime[28] based on the
395 unpruned trees, including GTDB reference sequences. Maximum relative rate ratio was set to
396 100. To correct for the difference between the placing of the root based on phylogenetic distances
397 only[83] and the most probable position of the last bacterial common ancestor (LBCA) on the tree[87],
398 *Fusobacteriota* were chosen as an outgroup. Only the tips present in the pruned tree were chosen
399 for further analyses.

400

401 **Transition directions**

402 One MAG from each MSG was randomly chosen. Additional MAGs from outside from the MSGs
403 were randomly chosen to downsample the dataset to the same number of genomes from each
404 biome, equal to the lowest number for any of the biomes (396 for brackish after removing multiple
405 MAGs within each MSG). "ace" function from the "ape" package in R was used to obtain maximum
406 likelihood ancestral states of the MRCAs of MSG pairs using an all-rates-different model[88].
407 Pairwise Wilcoxon rank-sum test of likelihoods of the ancestral states for the two biomes between

408 which transition occurred was performed for each transition type. Random selection of MAGs
409 within and outside the MSGs was iterated 100 times and mean likelihoods and p-values were
410 used for further statistical analyses.
411

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

413 Isoelectric points (pI) of all proteins within inferred bacterial proteomes were obtained using the
414 pepstats function from the EMBOSS package (v6.6.0)[89]. To get amino acid and amino acid
415 category frequencies the whole proteomes were concatenated into one sequence before running
416 separate pepstats analyses. pI frequencies within bins of pH width of 0.5 spanning the range from
417 3.0 to 12.5 were used to average, plot and compare the distributions of pIs between MSGs.
418 Average values over MAGs (95% ANI cluster representatives) were used for MSGs consisting of
419 more than 1 MAG.
420 For analysis of total changes in pI distribution, the absolute values of differences for each
421 0.5 pH wide bin were summed up. To compare transitions to divergence events within the same
422 biome, within each MSG bigger than 1 representative MAG, a pair of monophyletic groups which
423 diverged the longest time ago were identified. The differences between these "no transtion"
424 (within-MSG) monophyletic groups were compared with differences across the ancestral nodes
425 (MRCAs of MSG pairs) in a pairwise manner.
426

427 **Functional gene overlap and gene numbers**

428 Occurrences of KOs were transformed into a binary table of gene (KO) presence/absence. Gene
429 overlap was calculated as: $\frac{N_{gS}}{(N_{g1}+N_{g2})/2}$, where N_{gS} is the number of KOs shared between 2
430 genomes and N_{g1} and N_{g2} are the total numbers of KOs present in the respective genomes[74].
431 Formulas *log(gene overlap) ~ divergence times* and *gene overlap ~ divergence times* were used
432 to fit logarithmic and linear models for each transition type, including no transition (see
433 Comparison of isoelectric points and amino-acid frequencies). The models were compared using
434 a method based on the least-square means statistic[90]. Numbers of distinct polypeptide sequences
435 in the predicted proteomes were used as the gene numbers.
436

437 **Gene presence/absence analysis**

438 Occurrences of KOs were transformed into a binary table of gene (KO) presence/absence.
439 Random pairs of MAGs from the MSGs were chosen for comparison for each transition type. To
440 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[91]. The procedure was iterated 100 times, starting from the step of randomly choosing pairs of MAGs. The Benjamini-Hochberg method[92] 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[93] table collected from the KEGG website on the 24th of November 2021.

Statistical analyses

If not stated otherwise, pairwise Wilcoxon rank-sum test[91] 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 Extended Data 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 to the interpretation of the data, revised the manuscript and approved the final version before submission.

Acknowledgements

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

Conflict of interest

Authors declare no conflict of interest.

Figures

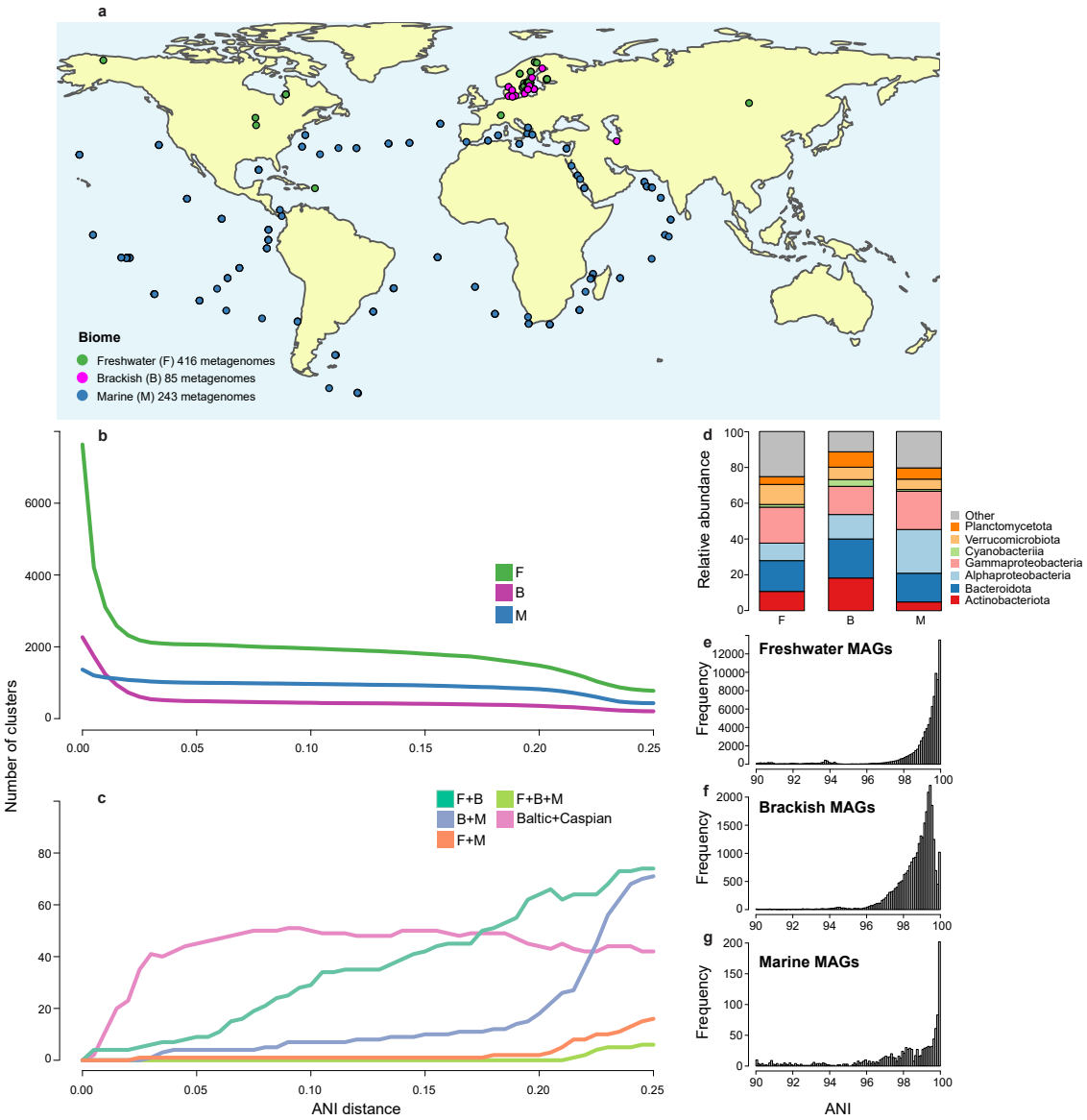


Fig. 1 | Status of shared bacterial clusters in different biomes. a. Map showing origins of MAGs [refs to papers from where we got the MAGs] **b.** Number of MAG clusters at different nucleotide distance cut-offs in each biome. **c.** Number of MAG clusters at different nucleotide distance cut-offs with 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.

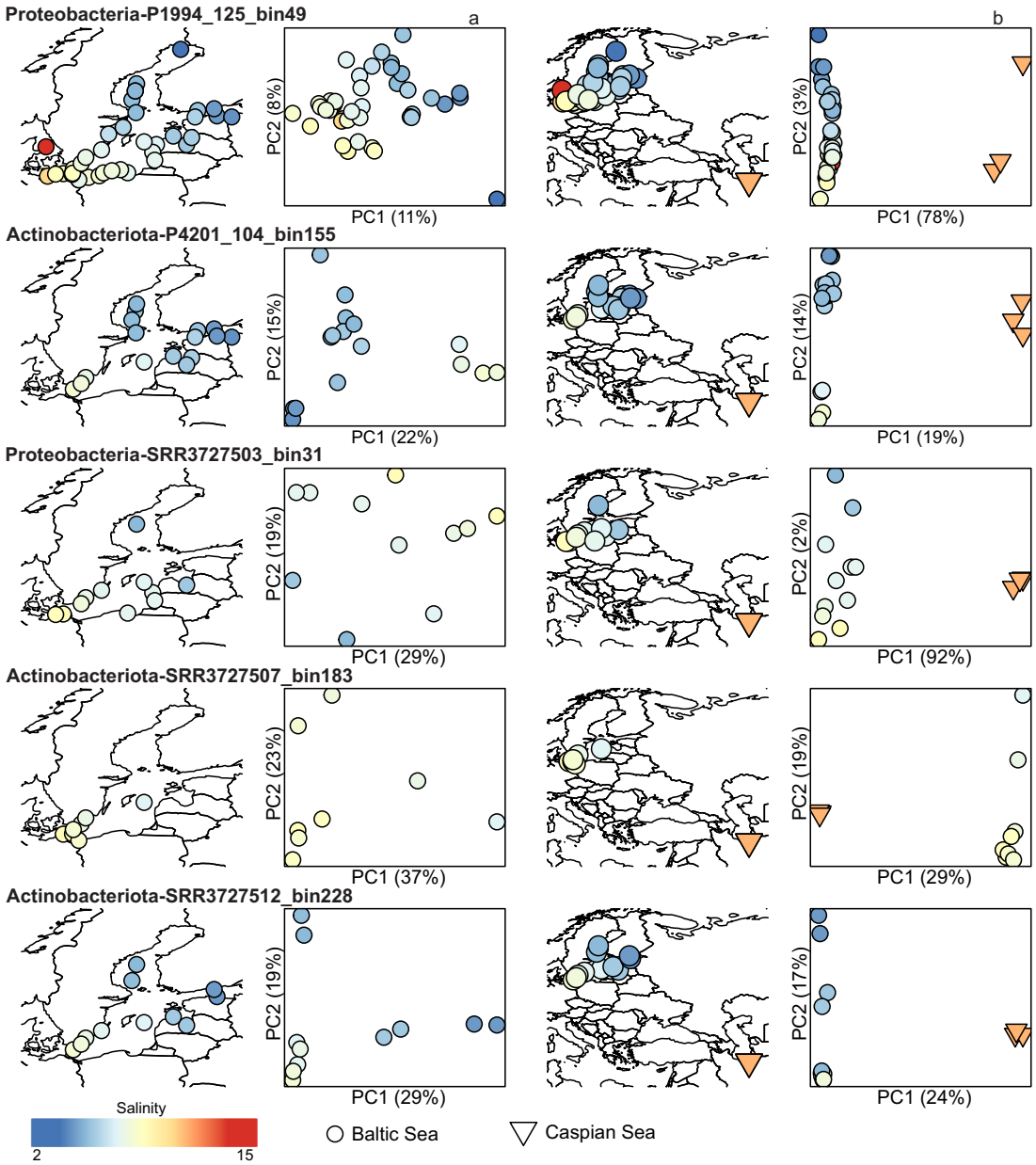


Fig. 2 | Population structure of brackish genome clusters. **a-b,** Geographic origin of each metagenome sample included for the MAG is shown to the left and a PCoA based on pairwise F_{ST} values to the right. Only Baltic Sea samples were included in **a**. whereas both Baltic Sea and Caspian Sea samples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

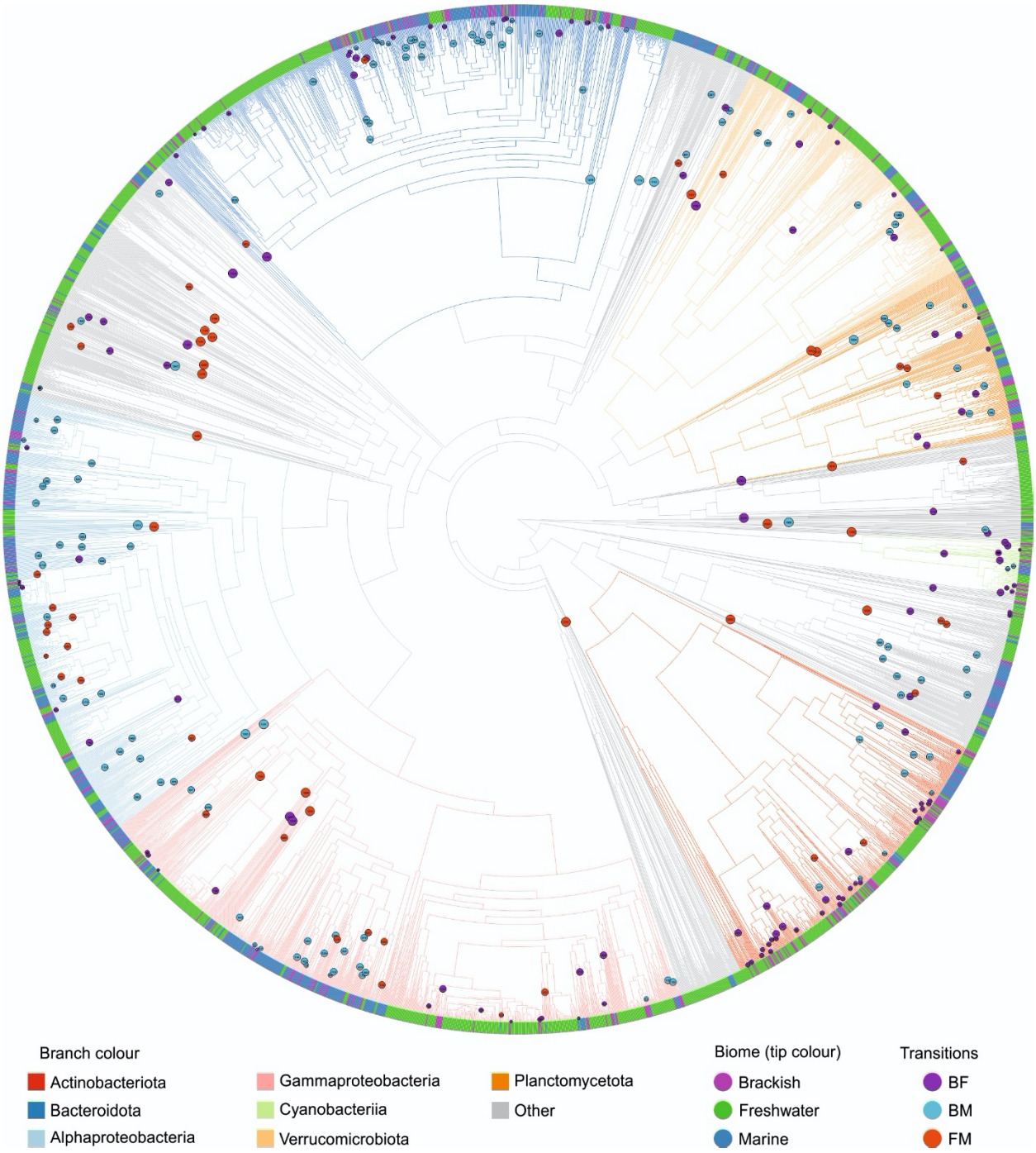


Fig. 3 | Reconstructed phylogeny of MAGs of the three aquatic biomes. Tree with branch lengths corresponding to estimated minimal times since divergence. Only tips present in the pruned tree are visualized. Identified transitions, i.e. nodes corresponding to most recent common ancestors of MSG pairs,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

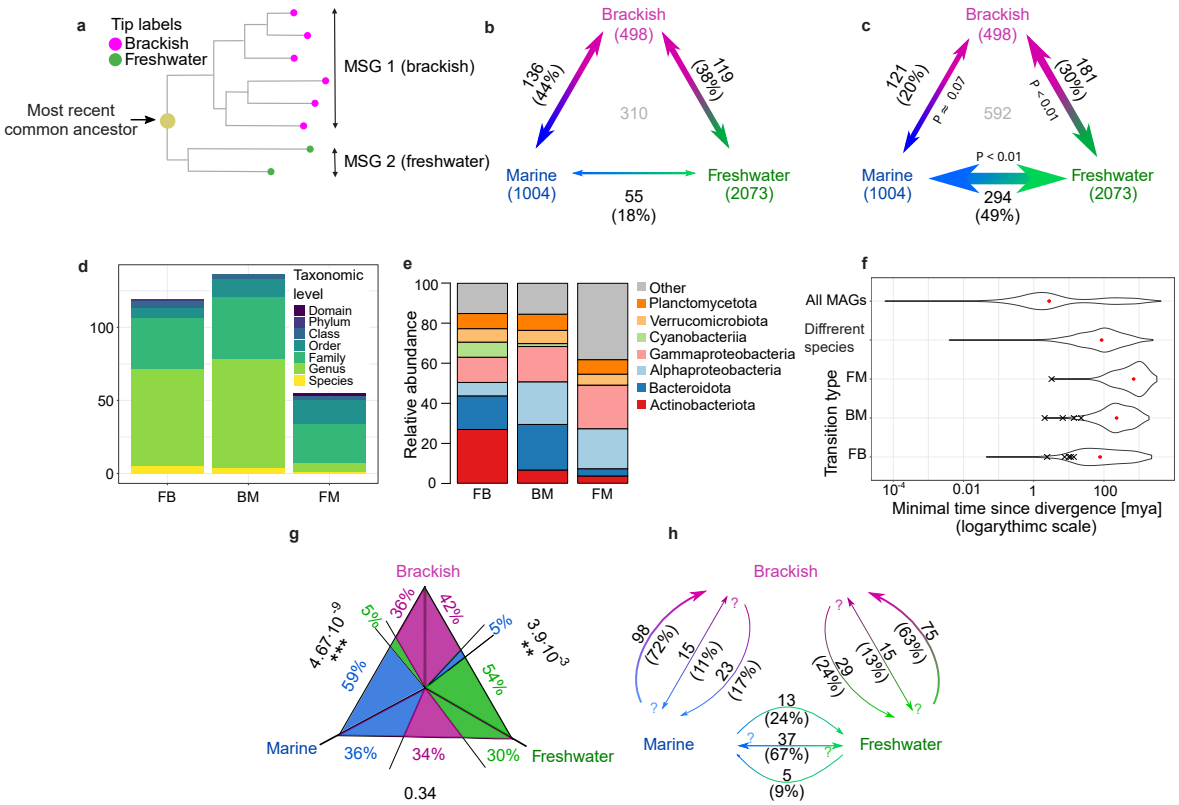


Fig. 4 | The most recent transitions between biomes. a. Illustration of an MSG pair corresponding to an FB transition **b**. Observed numbers of transitions between the biomes (their proportion in all transitions as a percentage). Numbers of MAGs included in the pruned tree are mentioned in brackets 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 is the lowest to which both MSGs pairs. If no annotation was available at a taxonomic level, a higher one was considered. Three FB pairs of MSGs belonged to the same >95% ANI cluster, but were not annotated to the same species by GTDB and species was considered as the lowest taxonomic level.**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 from pairwise Wilcoxon rank-sum test 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 transition type, right-hand side of the figure). **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. MAG: metagenome-assembled genome; MSG: monobiomic sister groups; MRCA: most recent common ancestor; BF: brackish ↔ freshwater; BM: brackish ↔ marine; FM: freshwater ↔ marine; mya: million years ago.

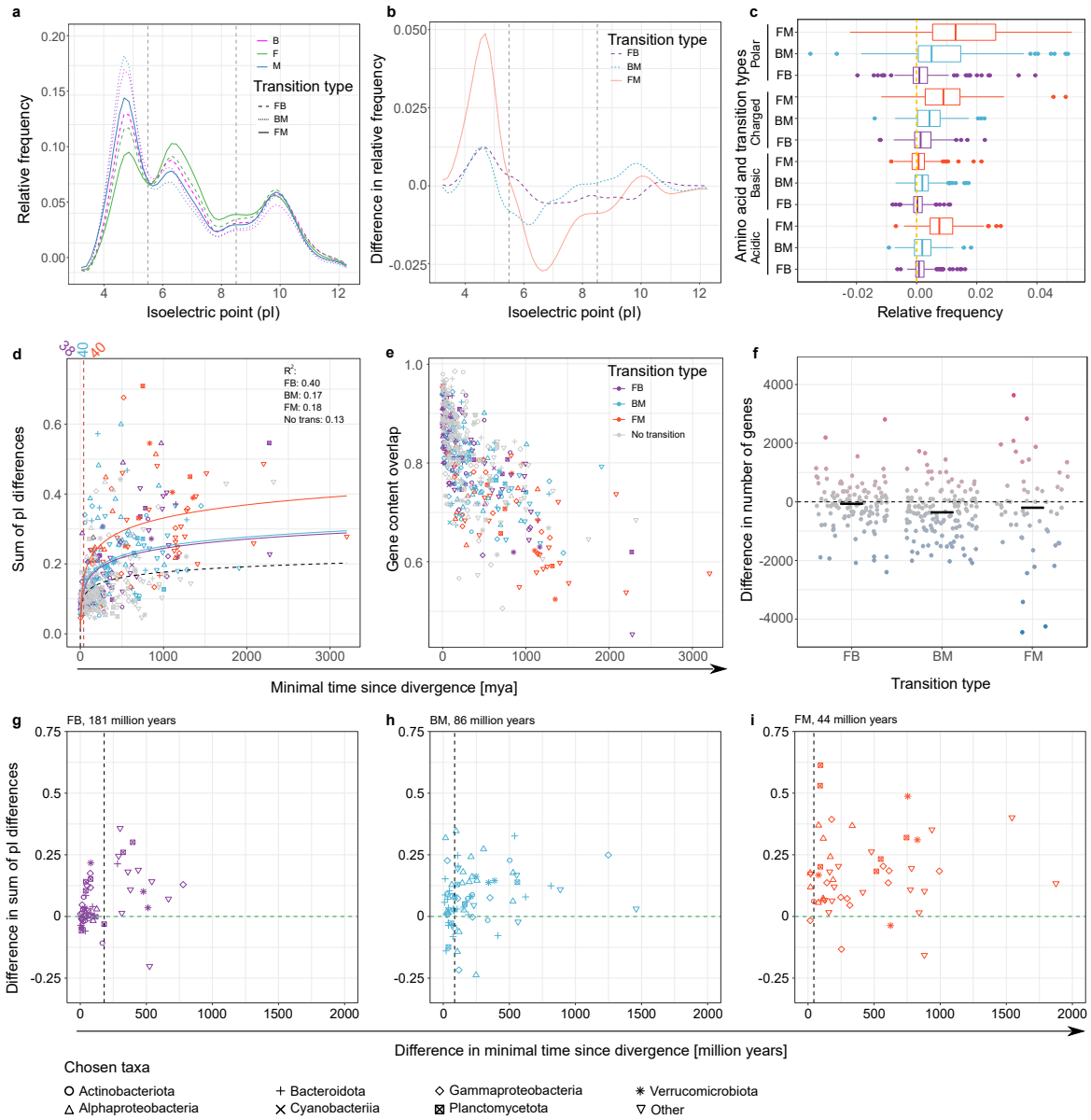


Fig. 5 | Changes in proteome properties and gene content with cross-biome transitions. a. Averaged distribution of pI values in predicted proteomes of MAGs. The distributions were averaged over all the MSGs from the same biome and the same transition type by taking the mean frequencies of proteins within 0.5 pH wide bins (from pI ∈ [3.0, 3.5) to pI ∈ [12.0, 12.5)). Gray vertical lines correspond to boundaries set to divide proteins based on their pI into three

groups: acidic (pI ∈ [3.0, 5.5)), neutral (pI ∈ [5.5, 8.5)) and basic (pI ∈ [8.5, 12.5)). **b.** Differences in pI distributions for each transition type (the difference in values between the pair of biomes for each transition type presented in **a**). **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 pI differences, as the sum of absolute values of differences over all used pI bins, 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 written above correspond to the lowest minimal time since transition for which predicted value - SE of the fit for a transition type is bigger than predicted value + SE for no transition events. **e.** Overlap in gene content vs minimal time since divergence for transitions and no transition, oldest divergence events within MSGs (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 biggest monophyletic groups within the MSGs formed after the transitions plotted against differences in minimal time since transition/divergence between the same events as in **d.** 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 (Wilcoxon rank-sum test p-value > 0.05). In **b**, **c** and **f**, the value differences are presented as value in the more saline - less saline biome. pI: isoelectric point; MSG: monobiomic sister groups; MRCA: most recent common ancestor; B: brackish; F: freshwater; M: marine; BF: brackish ↔ freshwater; BM: brackish ↔ marine; FM: freshwater ↔ marine; mya: million years ago; SE - standard error.

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, with its KO number and a descriptive name given on the right and a broader (KO level C) functional annotation marked to the left. To correct for labels inconsistent with the literature reference, the annotations of K01531[45] and K07301[94] were manually curated. The category "Enzymes with EC numbers" (represented by K00809) has been merged with "General function prediction only". The genes are marked as present in an MSG if they were found in >0.5 of the MAGs of the MSG. Only KOs found significantly (FDR adjusted $P < 0.1$) differentially present across transitions in a pairwise analysis are displayed. Descriptions of genes identified in more than one transition type were marked in bold. Groups of genes with at least one gene being in >50% of the cases being annotated also as the other(s) (Supplementary Data S4) were marked by adding #, $, ^ or * before the KO number.

Extended Data

Extended Data Table 1- Number of MAGs originated from each dataset

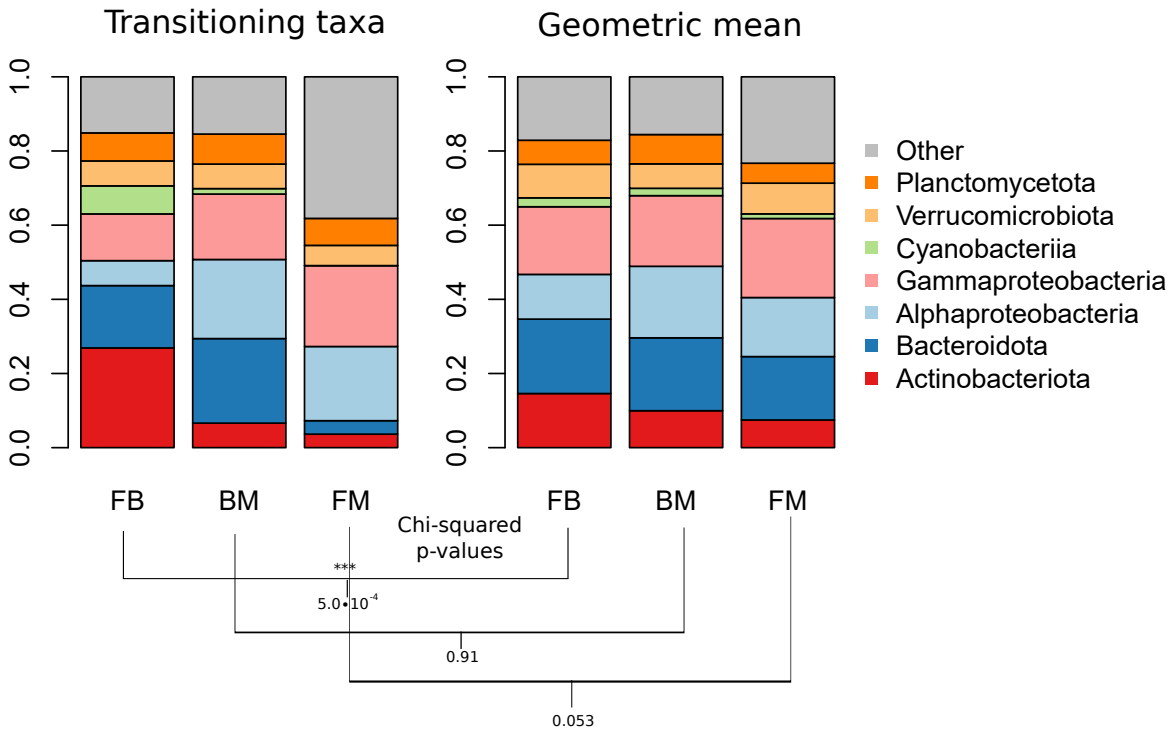
Dataset	#metagenomes	#Total MAGs	#filtered MAGs	#filtered Bacteria	#filtered Archaea	Reference
North American lakes (F)	141	193	84	84	0	Linz et al. (2018)
StratfreshDB (F)	267	8554	7654	7554	100	Buck et al. (2021)
Lake Baikal (F)	2	35	7	5	2	Cabello-Yeves et.al. (2018)
Baltic Sea (B)	82	1989	1989	1954	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)

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

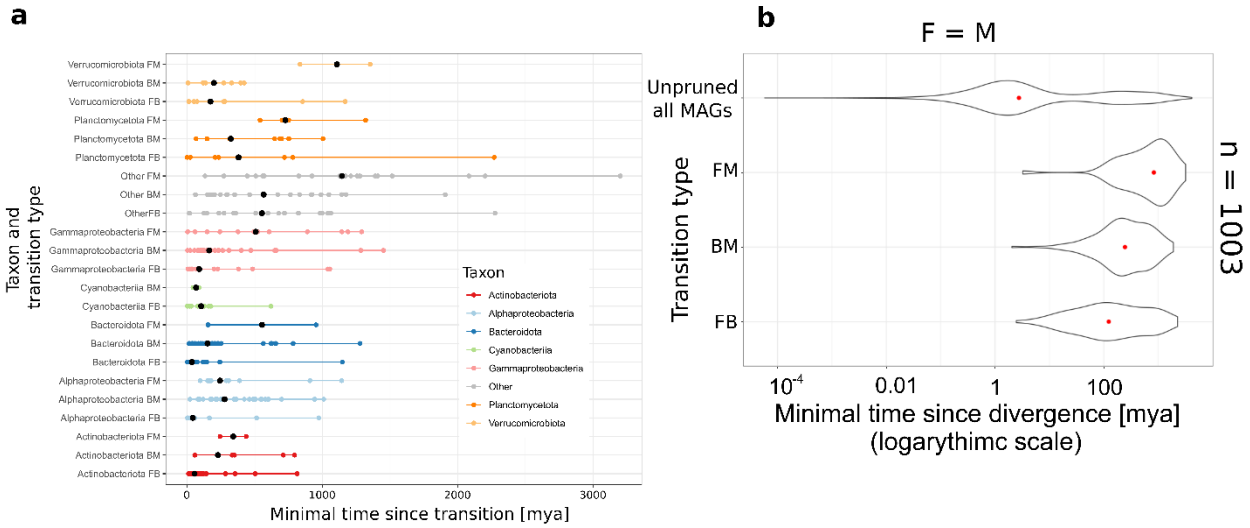
Extended Data Table 2- The number (proportion) of transitions of each kind for which a given taxonomic level is the lowest to which both MSGs belong. If no annotation was available at a taxonomic level, a higher one was considered. Pairs of MSGs belonged to the same >95% ANI cluster, but were not annotated to the same species by GTDB (3 FB cases), the lowest taxonomic level was considered to be species.

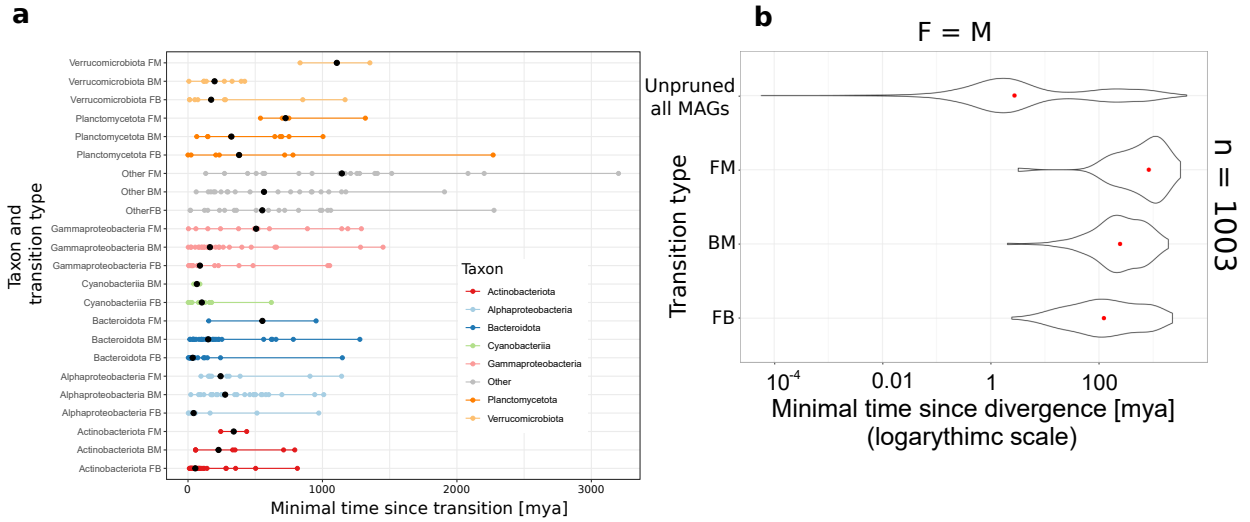
	FB	**BM**	**FM**
Domain	0 (0%)	0 (0%)	2 (3.6%)

Phylum	1 (0.84%)	0 (0%)	0 (0%)
Class	5 (4.2%)	3 (2.2%)	3 (5.5%)
Order	7 (5.9%)	12 (8.8%)	16 (29%)
Family	35 (29%)	43 (32%)	27 (49%)
Genus	66 (55%)	74 (54%)	6 (11%)
Species	5 (4.2%)	4 (2.9%)	1 (1.8%)

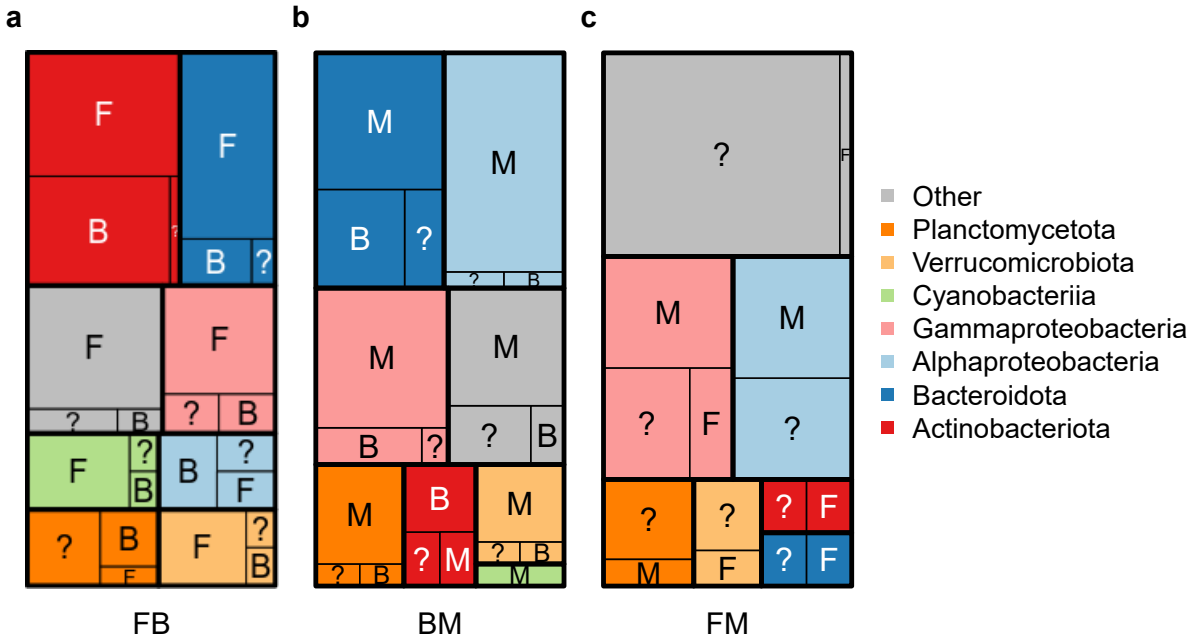


Extended Data Fig. 1 | Observed proportion of MSG pairs annotated to each of the chosen taxonomic groups vs expected numbers (geometric means of the proportion of lineages from each of taxonomic groups in respective biomes between which inferred transitions happened). *P*-values obtained from *X*-squared are given for each transition type below the bar charts.

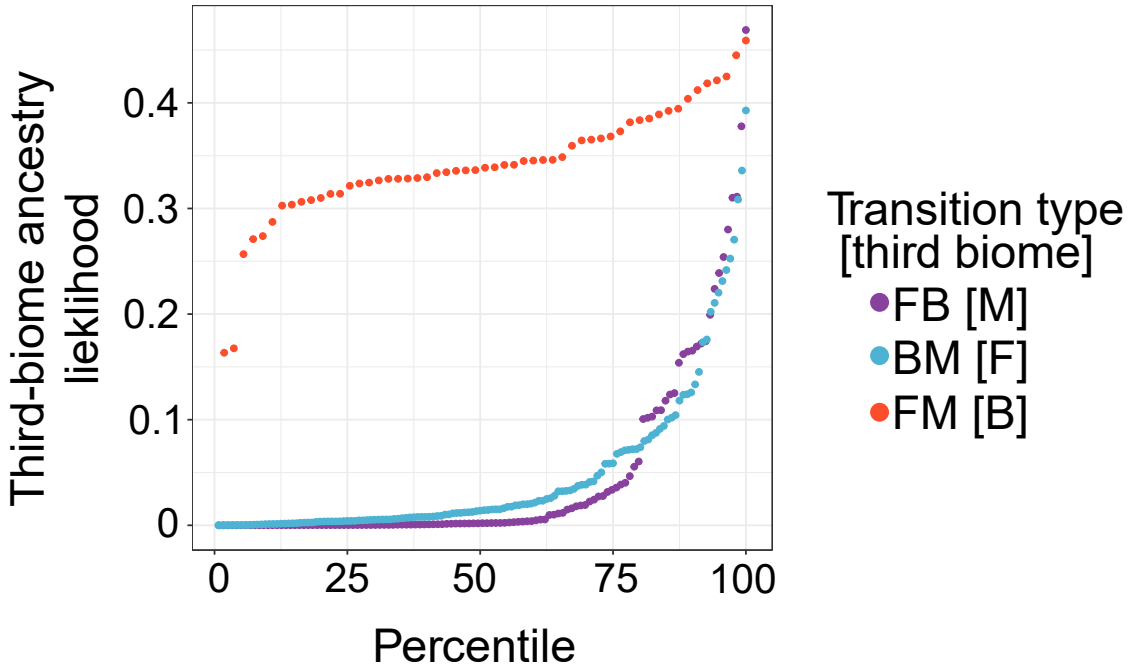




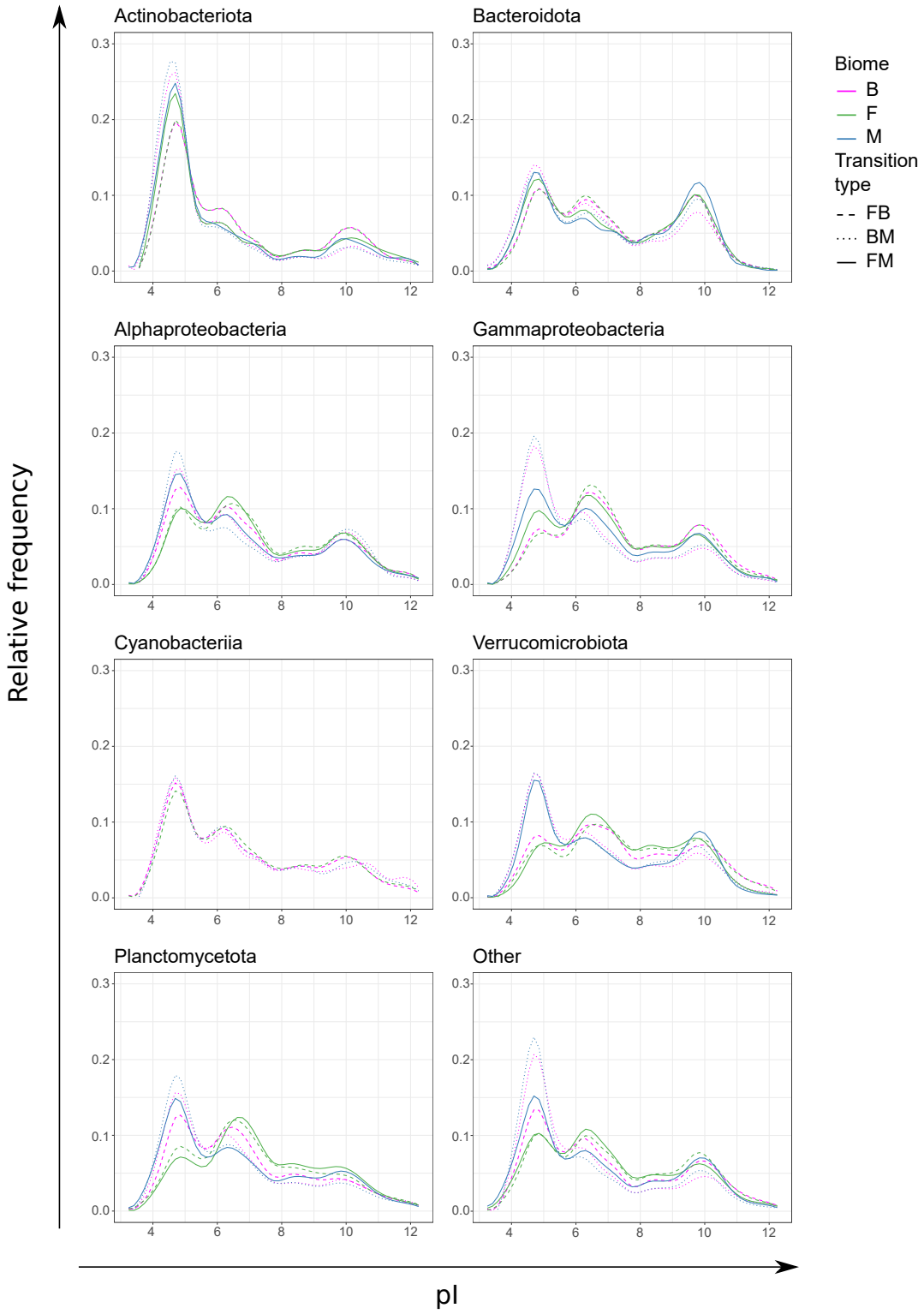
Extended Data Fig. 2 | Estimated times since the start of transitions a. grouped by taxa and **b.** based on the pruned tree further subsampled to the same number of freshwater and marine species. **a.** Dots correspond to estimated times since the start of transitions, and horizontal lines connect the most recent and the most ancient ones for each taxonomic group & transition type combination present in the data. **b.** The distributions of estimated minimal times since divergence for all nodes on the unpruned tree ("All MAGs") and for transitions of each type based on the subsampled tree.



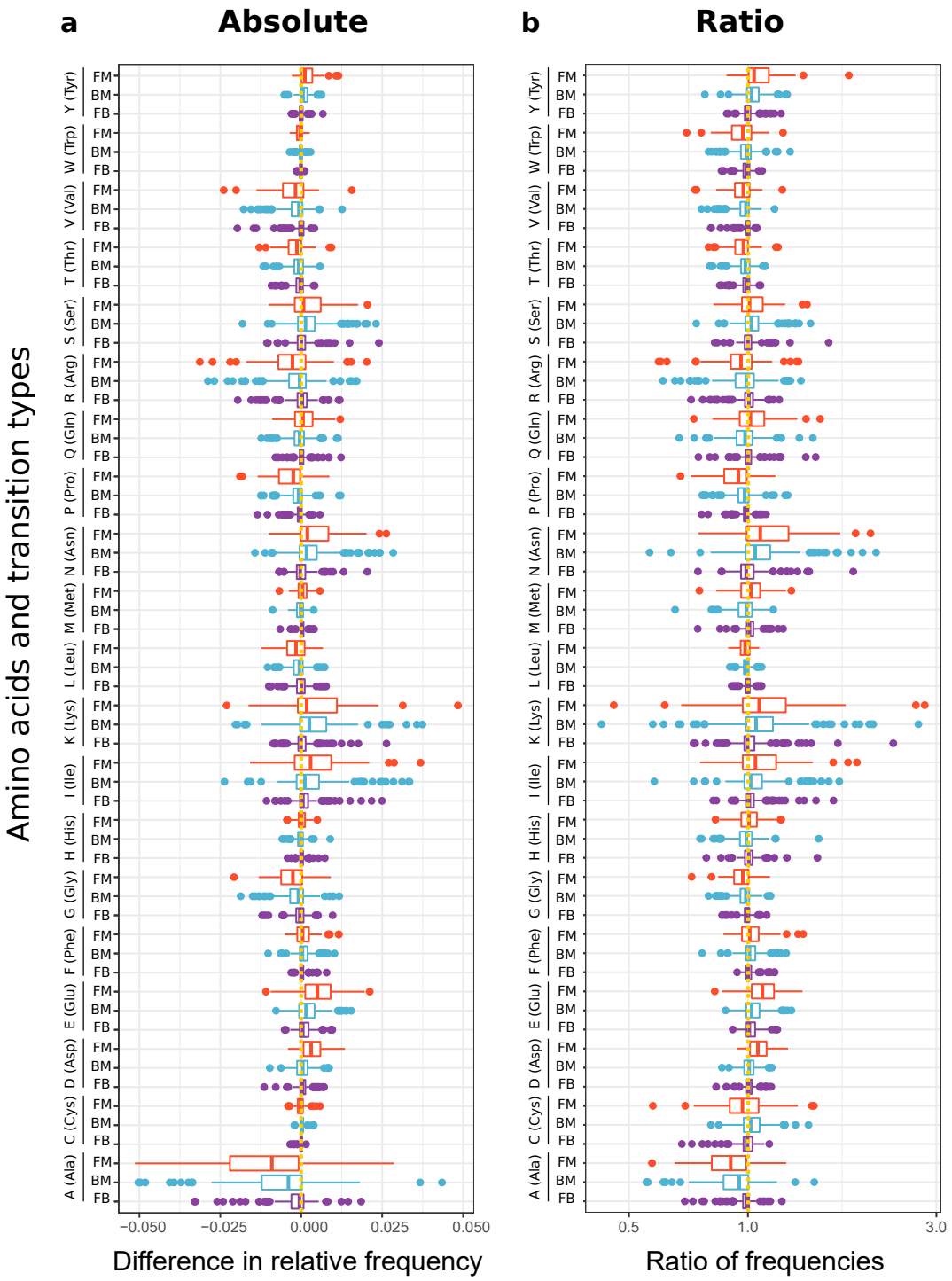
Extended Data Fig. 3 | Treemaps of dominant transition directions within taxa for **a.** FB, **b.** BM and **c.** FM transitions. The sizes of the boxes correspond to numbers of transitions for which the biome-ancestral state indicated by the letter (F - freshwater, B - brackish, M - marine) had higher likelihood than the two other ancestral biome-states taken together (*i.e.* likelihood >0.5). "?" indicates that the likelihood of all the ancestral states was <0.5.



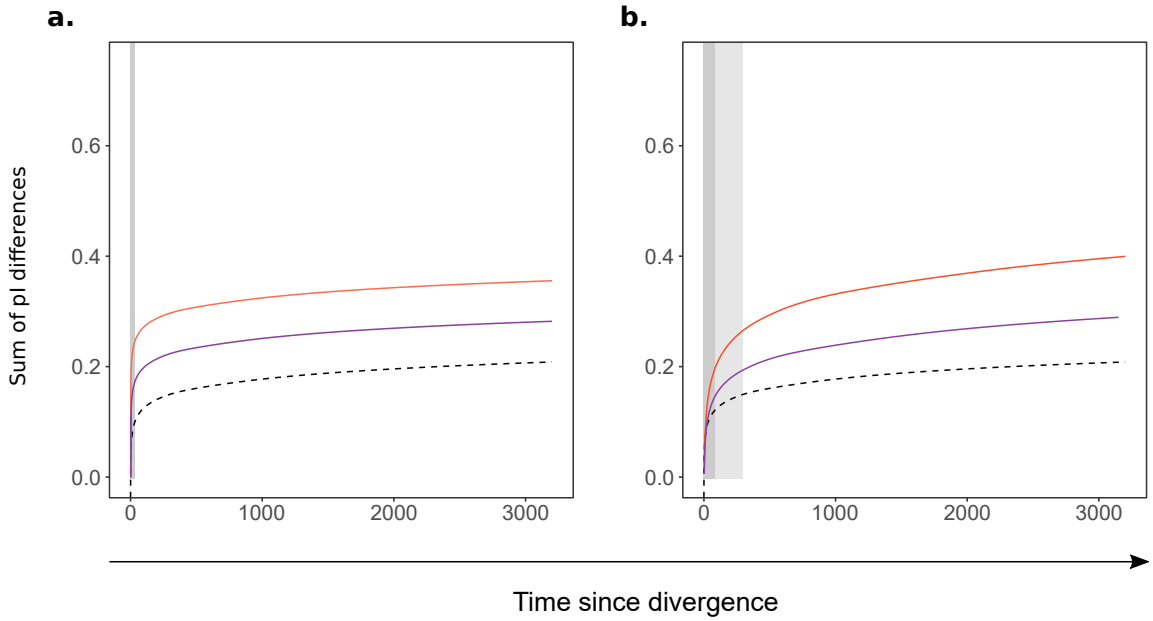
Extended Data Fig. 4 | The likelihood of MRCAs of MSG pairs originating from a different biome than any of its descendant MAGs (i.e. having a third-biome ancestral state). The transitions are ordered by increasing likelihood of the third-biome ancestral state.



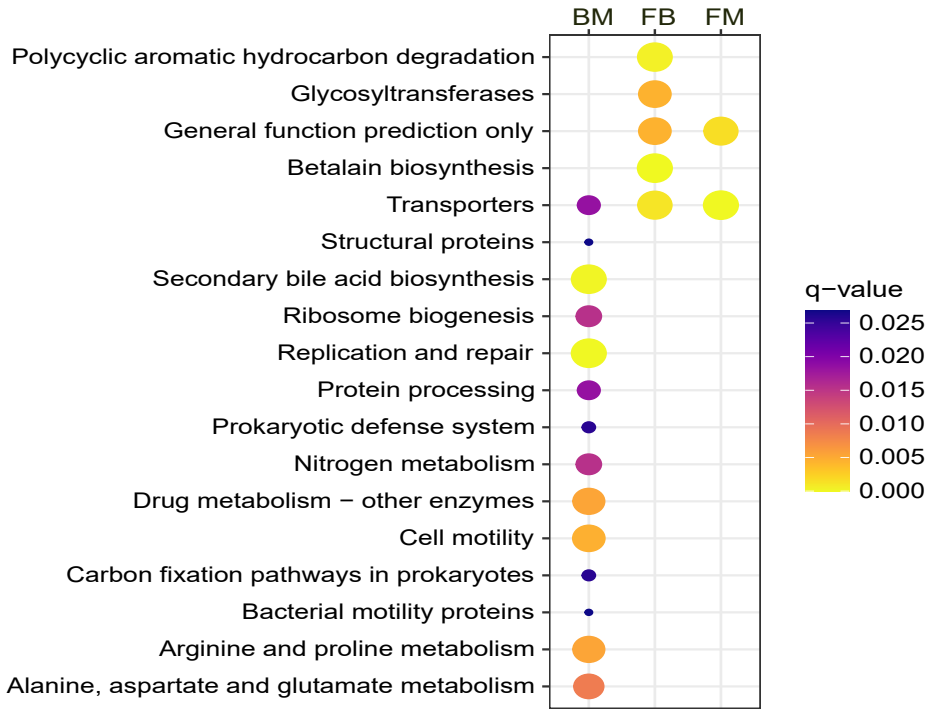
Extended Data Fig. 5 | Averaged distribution of isoelectric point (pI) values in predicted proteomes of MAGs. The distributions were averaged over all the MSGs from the same biome, transition type, and taxonomic group. The mean abundances of proteins within 0.5 pH wide bins (from pI ∈ [3.0, 3.5) to pI ∈ [12.0, 12.5)) were used to make the figure.



Extended Data Fig. 6 | Differences in frequencies of amino acids in the proteomes across the transitions. Presented as **a.** absolute differences in the relative frequencies or **b.** ratio of frequencies (logarithmic scale on x-axis).



Extended Data Fig. 7 | Models of proteome reorganization following cross-biome transitions. The curves illustrate hypothetical dynamics of change in pI distribution after divergence events (as in Fig. 5d) connected to cross-biome transitions (coloured lines) or within the same biome (dashed black line). **a.** Abrupt change in the distribution of pIs driven by biome-specific selection occurs in the early stage after transition, followed by random changes occurring in the diverging lineages regardless of selective pressure (model A) **b.** The transition alters the rate at which the distributions of pIs change, which is observable over a long time span as a combined effect of random changes and selective pressure (model B). Fig. 5d does not allow to correctly distinguish between the models, as taxonomic biases and differences in distributions of times since divergence affect the fitted logarithmic models. However, differences in pI distributions between transitions and no-transition divergence events within the same taxonomic groups (as in Fig. 5g-i) allows distinguishing which dynamics are more prevalent. Under model A, differences in pI distribution between transitions and divergence events occurring shortly after them should be significant, corresponding to abrupt changes after transitions (gray background in **a.**). Regardless of the size of change induced by the transition (orange vs purple line), it should be distinguishable from random effects after a similarly short time span. On the other hand, under model B, the increased rate of changes would lead to systematic differences being distinguishable from the noise only after a significant amount of time. The difference in time needed for differences in pI distributions to be distinguishable between transition and “no-transition” events would depend on how much the rate of change increases after transitions. Under a highly increased rate of change (orange line), only in the early stages (dark-gray background in **b.**), the differences are indistinguishable from the noise. If the rate increase is weaker, it takes substantially longer for the differences to reach the same level (light-gray background in **b.**). Results shown in Fig. 5g-i suggest that model B is the most common dynamic in the bacterial world.



Extended Data Fig. 8 | Results of over-representation analysis of the functional categories (KEGG Orthology level C) of differentially present KOs (i.e. gained/lost genes). The probability of obtaining the observed number of KOs annotated to a category among the differentially present KOs, as if they were drawn at random from all the KOs analyzed for a transition type, was assessed using a hypergeometric test.

679 ## References

680 1. Gould & Eldredge. Punctuated equilibria: an alternative to phyletic gradualism. *Models in*
681 *paleobiology*.
682 2. Louca, S. The rates of global bacterial and archaeal dispersal. *ISME J.* **16**, 159–167 (2022).
683 3. Puigbò, P., Lobkovsky, A. E., Kristensen, D. M., Wolf, Y. I. & Koonin, E. V. Genomes in
684 turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol.* **12**, 66
685 (2014).
686 4. Vos, M., Hesselman, M. C., Te Beek, T. A., van Passel, M. W. J. & Eyre-Walker, A. Rates
687 of Lateral Gene Transfer in Prokaryotes: High but Why? *Trends Microbiol.* **23**, 598–605
688 (2015).
689 5. Intergovernmental Panel on Climate Change. Coastal Systems and Low-Lying Areas. in
690 *Climate Change 2014 – Impacts, Adaptation and Vulnerability: Part A: Global and Sectoral*
691 *Aspects: Working Group II Contribution to the IPCC Fifth Assessment Report* 361–410
692 (Cambridge University Press, 2014).
693 6. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's
694 biogeochemical cycles. *Science* **320**, 1034–1039 (2008).
695 7. Ali, S. S., Yu, Y., Pfosser, M. & Wetschnig, W. Inferences of biogeographical histories
696 within subfamily Hyacinthoideae using S-DIVA and Bayesian binary MCMC analysis
697 implemented in RASP (Reconstruct Ancestral State in Phylogenies). *Ann. Bot.* **109**, 95–107
698 (2012).
699 8. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity.
700 *Nature* **551**, 457–463 (2017).
701 9. Logares, R. *et al.* Infrequent marine-freshwater transitions in the microbial world. *Trends*
702 *Microbiol.* **17**, 414–422 (2009).
703 10. Barberán, A. & Casamayor, E. O. Global phylogenetic community structure and β-diversity

704 patterns in surface bacterioplankton metacommunities. *Aquat. Microb. Ecol.* **59**, 1–10
705 (2010).
706 11. Walsh, D. A., Lafontaine, J. & Grossart, H.-P. On the Eco-Evolutionary Relationships of
707 Fresh and Salt Water Bacteria and the Role of Gene Transfer in Their Adaptation. in *Lateral*
708 *Gene Transfer in Evolution* (ed. Gophna, U.) 55–77 (Springer New York, 2013).
709 12. Hugerth, L. W. *et al.* Metagenome-assembled genomes uncover a global brackish
710 microbiome. *Genome Biol.* **16**, 279 (2015).
711 13. Delgado, L. F. & Andersson, A. F. Evaluating metagenomic assembly approaches for
712 biome-specific gene catalogues. *Microbiome* **10**, 72 (2022).
713 14. Cabello-Yeves, P. J. & Rodriguez-Valera, F. Marine-freshwater prokaryotic transitions
714 require extensive changes in the predicted proteome. *Microbiome* **7**, 117 (2019).
715 15. Dupont, C. L. *et al.* Functional tradeoffs underpin salinity-driven divergence in microbial
716 community composition. *PLoS One* **9**, e89549 (2014).
717 16. Eiler, A. *et al.* Productivity and salinity structuring of the microplankton revealed by
718 comparative freshwater metagenomics. *Environ. Microbiol.* **16**, 2682–2698 (2014).
719 17. Alneberg, J. *et al.* Ecosystem-wide metagenomic binning enables prediction of ecological
720 niches from genomes. *Commun Biol* **3**, 119 (2020).
721 18. Linz, A. M. *et al.* Freshwater carbon and nutrient cycles revealed through reconstructed
722 population genomes. *PeerJ* **6**, e6075 (2018).
723 19. Cabello-Yeves, P. J. *et al.* Genomes of Novel Microbial Lineages Assembled from the Sub-
724 Ice Waters of Lake Baikal. *Appl. Environ. Microbiol.* **84**, (2018).
725 20. Buck, M. *et al.* Comprehensive dataset of shotgun metagenomes from oxygen stratified
726 freshwater lakes and ponds. *Sci Data* **8**, 131 (2021).
727 21. Mehrshad, M., Amoozegar, M. A., Ghai, R., Shahzadeh Fazeli, S. A. & Rodriguez-Valera,
728 F. Genome Reconstruction from Metagenomic Data Sets Reveals Novel Microbes in the
729 Brackish Waters of the Caspian Sea. *Appl. Environ. Microbiol.* **82**, 1599–1612 (2016).

730 22. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft
731 metagenome-assembled genomes from the global oceans. *Sci Data* **5**, 170203 (2018).
732 23. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are
733 abundant in surface ocean metagenomes. *Nat Microbiol* **3**, 804–813 (2018).
734 24. Varghese, N. J. *et al.* Microbial species delineation using whole genome sequences.
735 *Nucleic Acids Res.* **43**, 6761–6771 (2015).
736 25. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat.*
737 *Biotechnol.* **38**, 1079–1086 (2020).
738 26. Sjöqvist, C., Delgado, L. F., Alneberg, J. & Andersson, A. F. Ecologically coherent
739 population structure of uncultivated bacterioplankton. *ISME J.* **15**, 3034–3049 (2021).
740 27. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify
741 genomes with the Genome Taxonomy Database. *Bioinformatics* (2019)
742 doi:10.1093/bioinformatics/btz848.
743 28. Battistuzzi, F. U., Tao, Q., Jones, L., Tamura, K. & Kumar, S. RelTime Relaxes the Strict
744 Molecular Clock throughout the Phylogeny. *Genome Biol. Evol.* **10**, 1631–1636 (2018).
745 29. Kim, H., Lee, S. & Jang, Y. Macroevolutionary patterns in the Aphidini aphids (Hemiptera:
746 Aphididae): diversification, host association, and biogeographic origins. *PLoS One* **6**,
747 e24749 (2011).
748 30. Moran, N. A., Munson, M. A., Baumann, P. & Ishikawa, H. A molecular clock in
749 endosymbiotic bacteria is calibrated using the insect hosts. *Proceedings of the Royal*
750 *Society of London. Series B: Biological Sciences* **253**, 167–171 (1993).
751 31. Thao, M. L. *et al.* Cospeciation of psyllids and their primary prokaryotic endosymbionts.
752 *Appl. Environ. Microbiol.* **66**, 2898–2905 (2000).
753 32. Cockerell, T. D. A. The Third Fossil Tsetse-Fly. *Nature* **98**, 70–70 (1916).
754 33. Lo, N., Bandi, C., Watanabe, H., Nalepa, C. & Beninati, T. Evidence for cocladogenesis
755 between diverse dictyopteran lineages and their intracellular endosymbionts. *Mol. Biol.*

Evol. **20**, 907–913 (2003).

34. Thorne, B. L., Grimaldi, D. A. & Krishna, K. Early Fossil History of the Termites. in *Termites: Evolution, Sociality, Symbioses, Ecology* (eds. Abe, T., Bignell, D. E. & Higashi, M.) 77–93 (Springer Netherlands, 2000).
35. Moran, N. A., Tran, P. & Gerardo, N. M. Symbiosis and insect diversification: an ancient symbiont of sap-feeding insects from the bacterial phylum Bacteroidetes. *Appl. Environ. Microbiol.* **71**, 8802–8810 (2005).
36. Shcherbakov, D. The 270 million year history of Auchenorrhyncha (Homoptera). https://www.academia.edu/download/33480266/Denisia__Biologiezentrum_Linz__Austria_Vol_0004-0029-0036.pdf.
37. Kuo, C.-H. & Ochman, H. Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. *Biol. Direct* **4**, 35 (2009).
38. Björck, S. A review of the history of the Baltic Sea, 13.0-8.0 ka BP. *Quat. Int.* **27**, 19–40 (1995).
39. Dumont, H. J. The Caspian Lake: History, biota, structure, and function. *Limnol. Oceanogr.* **43**, 44–52 (1998).
40. Morrison, K. L. & Weiss, G. A. Combinatorial alanine-scanning. *Curr. Opin. Chem. Biol.* **5**, 302–307 (2001).
41. Bossemeyer, D. *et al.* K+-transport Protein TrkA of Escherichia coli Is a Peripheral Membrane Protein That Requires other trk Gene Products for Attachment to the Cytoplasmic Membrane*. *J. Biol. Chem.* **264**, 16403–16410 (1989).
42. Holtmann, G., Bakker, E. P., Uozumi, N. & Bremer, E. KtrAB and KtrCD: two K+ uptake systems in Bacillus subtilis and their role in adaptation to hypertonicity. *J. Bacteriol.* **185**, 1289–1298 (2003).
43. Zhang, H. *et al.* Repeated evolutionary transitions of flavobacteria from marine to non-marine habitats. *Environ. Microbiol.* **21**, 648–666 (2019).

44. Levina, N. *et al.* Protection of Escherichia coli cells against extreme turgor by activation of MscS and MscL mechanosensitive channels: identification of genes required for MscS activity. *EMBO J.* **18**, 1730–1737 (1999).

45. Tao, T., Snavely, M. D., Farr, S. G. & Maguire, M. E. Magnesium transport in Salmonella typhimurium: mgtA encodes a P-type ATPase and is regulated by Mg2+ in a manner similar to that of the mgtB P-type ATPase. *J. Bacteriol.* **177**, 2654–2662 (1995).

46. Papp, K. M. & Maguire, M. E. The CorA Mg2+ transporter does not transport Fe2+. *J. Bacteriol.* **186**, 7653–7658 (2004).

47. Hsu, J.-L., Chen, H.-C., Peng, H.-L. & Chang, H.-Y. Characterization of the Histidine-containing Phosphotransfer Protein B-mediated Multistep Phosphorelay System in Pseudomonas aeruginosa PAO1*. *J. Biol. Chem.* **283**, 9933–9944 (2008).

48. Wösten, M. M. S. M., Wagenaar, J. A. & van Putten, J. P. M. The FlgS/FlgR Two-component Signal Transduction System Regulates the fla Regulon in Campylobacter jejuni*. *J. Biol. Chem.* **279**, 16214–16222 (2004).

49. Futamata, H. *et al.* Chemotactic response to amino acids of fluorescent pseudomonads isolated from spinach roots grown in soils with different salinity levels. *Soil Sci. Plant Nutr.* **44**, 1–7 (1998).

50. Larsen, M. H., Blackburn, N., Larsen, J. L. & Olsen, J. E. Influences of temperature, salinity and starvation on the motility and chemotactic response of Vibrio anguillarum. *Microbiology* **150**, 1283–1290 (2004).

51. Xu, X. *et al.* cheA, cheB, cheR, cheV, and cheY Are Involved in Regulating the Adhesion of Vibrio harveyi. *Front. Cell. Infect. Microbiol.* **10**, 591751 (2020).

52. Ramachandran, A., McLatchie, S. & Walsh, D. A. A Novel Freshwater to Marine Evolutionary Transition Revealed within Methylophilaceae Bacteria from the Arctic Ocean. *MBio* **12**, e0130621 (2021).

53. Snoeijs-Leijonmalm, P. & Andrén, E. Why is the Baltic Sea so special to live in? in

808 *Biological Oceanography of the Baltic Sea* (eds. Snoeijs-Leijonmalm, P., Schubert, H. &
809 Radziejewska, T.) 23–84 (Springer Netherlands, 2017).
810 54. Gleick, P. H. *Water in Crisis: A Guide to the World's Fresh Water Resources*. (Oxford
811 University Press, 1993).
812 55. Garrison, T. S. *Oceanography: An Invitation to Marine Science*. (Cengage Learning, 2012).
813 56. Cohan, F. M. Bacterial species and speciation. *Syst. Biol.* **50**, 513–524 (2001).
814 57. Louca, S. *et al.* Bacterial diversification through geological time. *Nat Ecol Evol* **2**, 1458–
815 1467 (2018).
816 58. Zaremba-Niedzwiedzka, K. *et al.* Single-cell genomics reveal low recombination
817 frequencies in freshwater bacteria of the SAR11 clade. *Genome Biol.* **14**, R130 (2013).
818 59. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98 (1973).
819 60. Kiraga, J. *et al.* The relationships between the isoelectric point and: length of proteins,
820 taxonomy and ecology of organisms. *BMC Genomics* **8**, 163 (2007).
821 61. Lanyi, J. K. Salt-dependent properties of proteins from extremely halophilic bacteria.
822 *Bacteriol. Rev.* **38**, 272–290 (1974).
823 62. Mevarech, M., Frolow, F. & Gloss, L. M. Halophilic enzymes: proteins with a grain of salt.
824 *Biophys. Chem.* **86**, 155–164 (2000).
825 63. Carstensen, J. & Duarte, C. M. Drivers of pH Variability in Coastal Ecosystems. *Environ.*
826 *Sci. Technol.* **53**, 4020–4029 (2019).
827 64. Slonczewski, J. L., Fujisawa, M., Dopson, M. & Krulwich, T. A. Cytoplasmic pH
828 Measurement and Homeostasis in Bacteria and Archaea. in *Advances in Microbial*
829 *Physiology* (ed. Poole, R. K.) vol. 55 1–317 (Academic Press, 2009).
830 65. Schneider, B. *et al.* Biogeochemical cycles. in *Biological Oceanography of the Baltic Sea*
831 (eds. Snoeijs-Leijonmalm, P., Schubert, H. & Radziejewska, T.) 87–122 (Springer
832 Netherlands, 2017).
833 66. Kirsch, M. L., Peters, P. D., Hanlon, D. W., Kirby, J. R. & Ordal, G. W. Chemotactic

834 methylesterase promotes adaptation to high concentrations of attractant in Bacillus subtilis.
835 *J. Biol. Chem.* **268**, 18610–18616 (1993).
836 67. Collard, F. *et al.* Molecular Identification of N-Acetylaspartylglutamate Synthase and β-
837 Citrylglutamate Synthase*. *J. Biol. Chem.* **285**, 29826–29833 (2010).
838 68. Li, H., Xu, H., Graham, D. E. & White, R. H. Glutathione synthetase homologs encode α-l-
839 glutamate ligases for methanogenic coenzyme F_{420} and tetrahydrosarcinapterin
840 biosyntheses. *Proceedings of the National Academy of Sciences* **100**, 9785–9790 (2003).
841 69. Kino Kuniki, Arai Toshinobu & Arimura Yasuhiro. Poly-α-Glutamic Acid Synthesis Using a
842 Novel Catalytic Activity of RimK from Escherichia coli K-12. *Appl. Environ. Microbiol.* **77**,
843 2019–2025 (2011).
844 70. Vigil-Stenman, T., Ininbergs, K., Bergman, B. & Ekman, M. High abundance and
845 expression of transposases in bacteria from the Baltic Sea. *ISME J.* **11**, 2611–2623 (2017).
846 71. Rankin, D. J., Rocha, E. P. C. & Brown, S. P. What traits are carried on mobile genetic
847 elements, and why? *Heredity* **106**, 1–10 (2011).
848 72. Koonin, E. V. & Wolf, Y. I. Genomics of bacteria and archaea: the emerging dynamic view
849 of the prokaryotic world. *Nucleic Acids Res.* **36**, 6688–6719 (2008).
850 73. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life.
851 *Nat. Rev. Genet.* **16**, 472–482 (2015).
852 74. Iranzo, J., Wolf, Y. I., Koonin, E. V. & Sela, I. Gene gain and loss push prokaryotes beyond
853 the homologous recombination barrier and accelerate genome sequence divergence. *Nat.*
854 *Commun.* **10**, 5376 (2019).
855 75. Hoetzinger, M. & Hahn, M. W. Genomic divergence and cohesion in a species of pelagic
856 freshwater bacteria. *BMC Genomics* **18**, 794 (2017).
857 76. Hanage, W. P., Spratt, B. G., Turner, K. M. E. & Fraser, C. Modelling bacterial speciation.
858 *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 2039–2044 (2006).
859 77. Marttinen, P., Croucher, N. J., Gutmann, M. U., Corander, J. & Hanage, W. P.

860 Recombination produces coherent bacterial species clusters in both core and accessory
861 genomes. *Microb Genom* **1**, e000038 (2015).
862 78. Whitaker, R. J., Grogan, D. W. & Taylor, J. W. Recombination shapes the natural
863 population structure of the hyperthermophilic archaeon Sulfolobus islandicus. *Mol. Biol.*
864 *Evol.* **22**, 2354–2361 (2005).
865 79. Sunagawa, S. *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev.*
866 *Microbiol.* **18**, 428–445 (2020).
867 80. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic
868 through eastern tropical Pacific. *PLoS Biol.* **5**, e77 (2007).
869 81. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM:
870 assessing the quality of microbial genomes recovered from isolates, single cells, and
871 metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
872 82. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High
873 throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries.
874 *Nat. Commun.* **9**, 5114 (2018).
875 83. Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between
876 domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477 (2019).
877 84. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
878 identification. *BMC Bioinformatics* **11**, 119 (2010).
879 85. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology
880 Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
881 86. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically
882 annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids*
883 *Res.* **47**, D309–D314 (2019).
884 87. Coleman, G. A. *et al.* A rooted phylogeny resolves early bacterial evolution. *Science* **372**,
885 (2021).

886 88. Pagel, M. Detecting correlated evolution on phylogenies: a general method for the
887 comparative analysis of discrete characters. *Proceedings of the Royal Society of London.*
888 *Series B: Biological Sciences* **255**, 37–45 (1994).
889 89. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open
890 Software Suite. *Trends Genet.* **16**, 276–277 (2000).
891 90. Lenth, R. V. Least-Squares Means: The R Package lsmeans. *J. Stat. Softw.* **69**, 1–33
892 (2016).
893 91. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1**, 80–83
894 (1945).
895 92. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful
896 approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
897 93. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference
898 resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–62 (2016).
899 94. Besserer, G. M., Nicoll, D. A., Abramson, J. & Philipson, K. D. Characterization and
900 Purification of a Na+/Ca2+ Exchanger from an Archaebacterium*. *J. Biol. Chem.* **287**,
901 8652–8659 (2012).

902