2

3   **GLOBAL DIVERSITY AND DISTRIBUTION OF AEROBIC ANOXYGENIC**

4   **PHOTOTROPHS IN THE TROPICAL AND SUBTROPICAL OCEANS**

5

6   Carlota R. Gazulla[1,2*], Adrià Auladell[2], Clara Ruiz-González[2], Pedro C. Junger[3], Marta Royo-

7   Llonch[2], Carlos M. Duarte[4], Josep M. Gasol[2,6], Olga Sánchez[1*], Isabel Ferrera[5*]

8   [1] Departament de Genètica i de Microbiologia. Universitat Autònoma de Barcelona,
9   08193 Bellaterra, Catalunya, Spain

10   [2] Departament de Biologia Marina i Oceanografia, Institut de Ciències del Mar, ICM-
11   CSIC, 08003-Barcelona, Catalunya, Spain

12   [3] Laboratory of Microbial Processes & Biodiversity (LMPB), Department of
13   Hydrobiology (DHB), Universidade Federal de São Carlos (UFSCar), São Carlos
14   13565-905, SP, Brazil

15   [4] Red Sea Research Center (RSRC) and Computational Bioscience Research Center
16   (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal,
17   Saudi Arabia

18   [5] Centro Oceanográfico de Málaga, Instituto Español de Oceanografía, IEO-CSIC,
19   29640 Fuengirola, Málaga, Spain

20   [6] Centre for Marine Ecosystems Research, School of Sciences, Edith Cowan University,
21   Joondalup WA Australia

22   *Corresponding authors: Carlota R. Gazulla: Carlota.Ruiz@uab.cat; Olga Sánchez:

23   Olga.Sanchez@uab.cat; Isabel Ferrera: isabel.ferrera@ieo.es

24

25

26   **Running title:** *Photoheterotrophs across the global ocean*

27   **Keywords:** AAP bacteria, *puf*M gene, ecological processes, selection, dispersion,

28   biogeography, Malaspina Expedition

29

ABSTRACT

The aerobic anoxygenic phototrophic (AAP) bacteria are common in most marine environments but their global diversity and biogeography remain poorly characterized. Here, we analyzed AAP communities across 113 globally-distributed surface ocean stations sampled during the Malaspina Expedition in the tropical and subtropical ocean. By means of amplicon sequencing of the *puf*M gene, the genetic marker for this functional group, we show that AAP communities along the surface ocean were mainly composed of members of the Halieaceae (Gammaproteobacteria), which were adapted to a large range of environmental conditions, and of different clades of the Alphaproteobacteria, that seemed to dominate under particular circumstances, such as in the oligotrophic gyres. AAP taxa were spatially structured within each of the studied oceans, with communities from adjacent stations sharing more taxonomic similarities. AAP communities were composed of a large pool of rare members and several habitat-specialists. When compared to the surface ocean prokaryotic and picoeukaryotic communities, it appears that AAP communities display an idiosyncratic global biogeographical pattern, dominated by selection processes and less influenced by dispersal limitation. Our study contributes to the understanding on how AAP communities are distributed in the horizontal dimension and the mechanisms underlying their distribution across the global surface ocean.

INTRODUCTION

The discovery of marine aerobic photoheterotrophs (i.e., aerobic anoxygenic phototrophic (AAP) bacteria and proteorhodopsin-containing bacteria) (Béjà *et al.*, 2000; Kolber *et al.*, 2000) challenged the classic view of bacterioplankton being composed of photoautotrophic microorganisms as primary producers and of chemoheterotrophs as consumers. Since then, many studies have investigated their abundance, diversity, and distribution in the ocean, and ultimately tried to understand their role in the marine ecosystem (DeLong and Béjà, 2010; Kirchman and Hanson, 2013; Koblížek, 2015; Pinhassi *et al.*, 2016). AAP bacteria are photoheterotrophs that use dissolved organic matter but that harvest solar energy using bacteriochlorophyll *a* (Bchla) to supplement their metabolism. In the marine environment, these organisms can typically constitute up to 10% of total prokaryotes (Schwalbach and Fuhrman, 2005; Sieracki *et al.*, 2006; Jiao *et al.*, 2007; Hojerová *et al.*, 2011) and are an active part of the community because they consist of large cells that display higher growth rates and receive higher grazing pressure than most bacteria (Sieracki *et al.*, 2006; Koblížek *et al.*, 2007; Ferrera *et al.*, 2011, 2017). It has thus been hypothesized that this functional group plays a remarkably important role in the processing of organic matter, and as a consequence, in the global carbon cycle (see review by Koblížek, 2015).

Phylogenetically, marine AAP bacteria belong mainly to the Alpha- and Gammaproteobacteria classes. The *puf*M gene, involved in the synthesis of Bchla, is commonly used to screen the diversity of AAPs in environmental samples and to describe their distribution patterns. The first studies showed AAP communities as being mainly affiliated to the alphaproteobacterial *Roseobacter*-like clade (Béjà *et al.*, 2002; Oz *et al.*, 2005) but the Global Ocean Sampling (GOS), based on metagenomic data, unveiled that an important fraction of marine AAP bacteria were associated to phylogroups without cultured representatives (Yutin *et al.*, 2007). The later study also showed that, while the *Roseobacter*-like AAPs were the most ubiquitous clade, unidentified uncultured groups dominated in open ocean areas, while Gammaproteobacteria dominated in coastal sites (Yutin *et al.*, 2007). Later investigations showed that

74 Gammaproteobacteria have in fact a widespread distribution and can constitute an important

75 fraction of AAP communities in diverse sites of contrasting trophic status (Mašín *et al.*, 2006;

76 Lehours *et al.*, 2010; Ferrera *et al.*, 2014; Lehours and Jeanthon, 2015; Auladell *et al.*, 2019). In

77 contrast, AAPs from the Betaproteobacteria clade are rarely prevalent in marine environments

78 and they seem to prefer low-salinity waters (Waidner and Kirchman, 2008; Cottrell and

79 Kirchman, 2009; Boeuf *et al.*, 2013).

80 Although most AAP diversity studies have been restricted to particular areas of the world's

81 ocean, a few studies have already compared communities across different oceanic regions. The

82 pioneering metagenomic study by Yutin *et al.*, (2007), which covered a transect between 45°N

83 in the Atlantic Ocean and 15°S in the Pacific Ocean, showed that the composition of AAP

84 communities varied between different biogeographical regions. By constructing clone libraries

85 in a limited number of samples (N=10) from the Pacific, Atlantic and Indian oceans, Jiao *et al.*,

86 (2007) reported diversity patterns linked to the trophic regime of the oceanic region. Later,

87 another study compared clone libraries of different seas encompassing a very large

88 environmental variability (Mediterranean Sea, North Pacific Ocean, Western Beaufort Sea,

89 Barents Sea and Norwegian Sea), and found that deterministic processes largely influenced the

90 structuring of AAP assemblages (Lehours *et al.*, 2018). This study further concluded that

91 diverse AAP lineages showed some habitat preference, suggesting the existence of a certain

92 degree of ecological cohesiveness for AAP clades, at least when comparing contrasting biomes.

93 Besides, a study applying high-throughput sequencing to coastal Australian waters concluded

94 that AAP communities exhibited niche partitioning whereas others shared their preferred niches

95 (Bibiloni-Isaksson *et al.*, 2016). Altogether, these results indicate that AAP assemblages –and

96 the taxa within them– display complex spatial patterns (Jiao *et al.*, 2007; Yutin *et al.*, 2007;

97 Lehours *et al.*, 2010; Jeanthon *et al.*, 2011; Boeuf *et al.*, 2013; Lehours and Jeanthon, 2015;

98 Bibiloni-Isaksson *et al.*, 2016), probably driven by environmental selection (Lehours *et al.*,

99 2018). Nevertheless, these conclusions are drawn from studies performed at different scales,

100   using various methodologies and biased towards particular –and often coastal– ocean regions,

101   so a coherent global assessment is still lacking.

102   The exploration of the worldwide distribution of marine microorganisms, and thus, the

103   definition of global biogeographical patterns, has become feasible in the last decade thanks to

104   contemporary global oceanographic circumnavigations like the Malaspina Circumnavigation

105   Expedition (Duarte 2015) or the *Tara* Oceans Expedition (Karsenti *et al.*, 2011), that used

106   standardized procedures in a large collection of samples, coupled with recent advances in

107   sequencing methodologies. Large scale surveys have also been key in the definition of the

108   underlying ecological mechanisms in bulk prokaryotic and small eukaryotic communities (de

109   Vargas *et al.*, 2015; Salazar *et al.*, 2015; Sunagawa *et al.*, 2015; Ruiz-González *et al.*, 2019;

110   Logares *et al.*, 2020; Obiol *et al.*, 2020). Data generated from large sequencing initiatives have

111   also been used to retrieve new diversity (Tully *et al.*, 2018; Nayfach *et al.*, 2020), including that

112   within the AAPs (from the *Tara* Oceans expedition, Graham *et al.*, 2018). Hence, a

113   comprehensive study defining the global ocean biogeography of AAP assemblages and the

114   mechanisms underlying their patterns is now feasible, but yet to be performed.

115   Here, we present a global assessment of AAP bacteria communities across the global tropical

116   and subtropical ocean based on the Malaspina Circumnavigation Expedition. In particular, we

117   studied the diversity and biogeography of AAP communities at a fine scale in the surface ocean

118   using amplicon sequence variants (ASVs) of the *puf*M gene. Our objectives were three-fold: 1)

119   to describe the diversity and biogeography of the surface AAP assemblages along the global

120   tropical and subtropical ocean, 2) to disentangle the factors driving global patterns of AAP

121   communities, and 3) to compare the trends observed in the AAP communities with those of the

122   broader surface ocean microbiota (i.e., whole prokaryotic and picoeukaryotic communities). For

123   this purpose, we analyzed the composition of AAP communities based on the dominance or

124   rarity of each individual taxa in an approach based on the spatial abundance distribution of each

125   ASV. Further, we estimated the role of different ecological processes shaping the structure of

126   AAP communities. Since AAP bacteria, as a whole, display ecological traits that differentiate

127   them from the rest of the bacterioplankton (i.e., photoheterotrophy, high growth rates, and

128   higher susceptibility to predation than other prokaryotes), we hypothesize that their ecological

129   patterns may deviate from those of the bulk communities.

130

131   **RESULTS AND DISCUSSION**

132   *Oceanographic context*

133   The 113 studied stations were representative of the tropical and subtropical regions of the three

134   major oceans, the Pacific, the Atlantic and the Indian Ocean (Table 1, Fig. S1). The cruise track

135   spanned across all five subtropical oceanic gyres, characterized by their oligotrophic conditions,

136   as well as over relatively more productive areas such as the Equatorial Pacific, the Caribbean

137   Sea, the Benguela coastal province or the South Subtropical convergence current, in the South

138   Australian Bight (Estrada *et al.*, 2016). Across this route, temperatures ranged between 15.8 and

139   29.3ºC (mean 24.5ºC), with the coldest waters found in the South Australian Bight and the

140   warmest temperatures in samples located along the Equatorial Pacific and Atlantic Oceans (Fig.

141   S2). Salinity ranged from 33.15 to 37.65 PSU, being the highest in stations from the Atlantic

142   Ocean and lowest in certain stations from the Indian and Pacific oceans. Chlorophyll *a* (Chla)

143   ranged between 0.034 (Station 38 in the South Atlantic) and 0.647 mg·m^{-3} (Station 45 in the

144   Benguela current coast) with a mean value of 0.155 mg·m^{-3}. Phosphate, nitrate, and silicate had

145   higher concentrations in the Equatorial Pacific, in the South African stations and in the South of

146   Australia (Fig. S2). Water mass properties and productivity regimes for the stations sampled in

147   the Malaspina Circumnavigation Expedition have been previously described in detail (e.g.

148   Estrada et al 2016, Teira et al 2019, Regaudie-de-Gioux et al 2019, and Villamaña et al 2019).

149

150   *Contrasting patterns of alpha diversity in distinct biogeographical provinces*

151   Our survey of the *puf*M gene allowed us to generate the largest dataset of amplicon sequence

152   variants of the *puf*M gene thus far available. Partial sequencing of this marker resulted in 1119

153   distinct ASVs that clustered into 229 OTUs (94% similarity). Rarefaction curves reached a

154   plateau for all samples (Fig. S3A), indicating that we obtained a fair representation of the

155   AAPs' surface ocean diversity for each individual sample, while the global sample-based

156   rarefaction curve (Fig. S3B) suggested that the number of ASVs would rise had more stations

157   been sampled. We observed a large variability in the richness estimates (Chao1 index) per

158   community (Fig. 1), which varied between 14 and 132 ASVs (mean 61) while the Shannon

159   diversity index ranged between 0.9 and 3.9 (mean 2.9). Overall, richness values were within the

160   same range than those previously reported from the Mediterranean Sea or Australian coastal

161   waters using similar methodologies (Bibiloni-Isaksson *et al.*, 2016; Auladell *et al.*, 2019).

162   Richness and diversity of AAP communities were highest in the North Atlantic (mean richness

163   81, mean Shannon diversity 3.1) compared to other regions (Tukey test, p<0.001, Fig. 1).

164   Taxonomic richness and diversity varied between and within some Longhurst provinces. In

165   general, AAP bacteria diversity was lower in eutrophic areas (correlation between Shannon and

166   Chla concentration, N=113, R=−0.33, p<0.001 and primary production, N=96, R=−0.38,

167   p<0.001), consistent with previous observations (Jiao *et al.*, 2007; Jeanthon *et al.*, 2011). In

168   contrast, AAP communities having higher richness values were associated with low

169   concentrations of phosphate (N=89, R=−0.48, p<0.0001) and nitrate (N=89, R=−0.34,

170   p=0.001), and correlated positively with temperature and salinity (N=113, R=0.24, p=0.011;

171   R=0.29 p=0.002, respectively, see Table S1). Temperature and salinity had been shown to

172   influence AAP bacterial richness at local scales (Lehours and Jeanthon, 2015; Bibiloni-Isaksson

173   *et al.*, 2016). Our results demonstrate that temperature, salinity, and trophic status govern

174   patterns of AAP bacterial alpha diversity at the global scale.

175   We also explored whether the patterns of AAP diversity were similar to the trends observed for

176   other picoplanktonic groups. To this end, we compared the Shannon diversity of prokaryotes

177   and picoeukaryotes (previously determined in the same sample set, Ruiz-González *et al.*, 2019;

178   Logares *et al.*, 2020) with the values obtained for AAP bacteria. We observed a significant

179   negative correlation between the Shannon diversity index of AAP communities and that of total

180   prokaryotes (N=104, R=−0.32, p=0.001, Fig. S4), while no significant correlation was found

181   with the picoeukaryotic community values. However, the low diversity values observed for

182   AAP bacteria in some eutrophic regions (PEQD, PNEC and SSTC provinces, cf. Table 1 for

183   complete names) were not observed in the whole prokaryotic dataset, suggesting that trophic

184   status may exert a strongest role in shaping the diversity of the AAP subcommunity than of the

185   bulk prokaryotic assemblage.

186   ***Spatially structured communities dominated by distinct taxonomic groups***

187   We classified all the ASVs into 7 broad taxonomic groups based on their placement in a

188   reference phylogenetic tree (Fig. S5). One group contained sequences assigned to the family

189   Halieaceae of the Gammaproteobacteria (here-after 'Gamma-Halieaceae' group), while the

190   'Betaproteobacteria' group included sequences from the Burkholderiales order. Members of the

191   Alphaproteobacteria were distinguished into four subgroups: 'Methylobacteriaceae' (sequences

192   from order Rhizobiales, family Methylobacteriaceae), 'Rhodobacterales' (order

193   Rhodobacterales), 'Sphingomondales' (order Sphingomonadales), and 'Alpha-Others', which

194   grouped other members of the Alphaproteobacteria that could not be further assigned. Finally,

195   sequences that did not belong to any of these groups were classified as 'Others'.

196   Most of the studied communities (75 out of 113 sampled stations, Fig. 2 and Fig. S6) were

197   dominated by Gamma-Halieaceae, followed by 24 stations dominated by Alpha-

198   Rhodobacterales. The overall dominance of these groups is in agreement with previous studies

199   from the Mediterranean Sea (Lehours *et al.*, 2010; Jeanthon *et al.*, 2011; Ferrera *et al.*, 2014;

200   Auladell *et al.*, 2019), the Baltic Sea (Mašín *et al.*, 2006), the Arctic Ocean (Lehours and

201   Jeanthon, 2015), and Australian waters (Bibiloni-Isaksson *et al.*, 2016).

202   The large dominance of gammaproteobacterial clades in marine AAP communities has been a

203   matter of debate and it has been argued that it could be due to possible primer biases in

204   amplicon-based studies (Lehours *et al.*, 2010; Ferrera *et al.*, 2014). In fact, PCR-based

205   approaches can suffer from amplification biases that could result in misrepresentation of the

206   relative abundances of various taxa as well as in low phylogenetic coverage. Nevertheless, a

207 recent comparison of AAP assemblages in the Mediterranean Sea using metagenomics and

208 *puf*M amplicon sequencing showed that, despite there were some discrepancies in the relative

209 abundance of certain taxa, Gammaproteobacteria were abundant in both the amplicon and

210 metagenomic datasets, which showed comparable patterns of diversity and community structure

211 (Auladell *et al.*, 2019). This study also showed that despite that the amplicon approach –

212 identical to the one used here– missed some phylogenetic groups, it allowed the identification of

213 other groups that were overlooked by metagenomics because they were present in low

214 abundances, as well as the retrieval of more variants, enabling the definition of distinct ecotypes

215 among very similar sequences (Auladell *et al.*, 2019). Metagenomics is often considered the

216 least biased approach for functional gene analysis, but it is limited in its capacity to retrieve the

217 least abundant members of the communities, and AAP taxa are generally present at relatively

218 low abundances in natural samples (often below 10%). Although technically possible, the cost

219 of conducting a high-resolution global ecological study based on a specific functional gene

220 using metagenomics would be prohibitive and unfeasible for most researchers because, among

221 other reasons, metagenomes retrieve less copies of specific marker genes for a given sequencing

222 investment. Given that the goal of this work was to establish the global ecological patterns of

223 AAP communities, and to understand how these are assembled at the fine-scale, we consider

224 that the *puf*M amplicon sequencing, despite not free of biases, was the most suitable approach to

225 address our questions.

226

227

228

229

230 Interestingly, Gamma-Halieaceae and Alpha-Rhodobacterales-dominated communities were not

231 randomly distributed but appeared to be spatially structured, with a marked succession of

232 samples dominated by either one or the other group (see Fig. 2A). Gamma-Halieaceae

233   contributed between 0.1% and 99.7% of total community *puf*M sequences (median or mean??).

234   In locations where they were not abundant, the contribution of Alpha-Rhodobacterales was

235   high, suggesting a replacement of the dominant taxonomic group across space (Fig. 2). Both

236   groups also showed high intragroup diversity across samples (see Fig. 2), yet we observed that

237   in some stations in the North Pacific region (Stations 114, 115, 118 and 119) this intragroup

238   diversity decreased, and one single sequence assigned to the Gamma-Halieaceae (ASV217)

239   represented abundances over 50% coinciding with a decrease in salinity. The relative

240   contribution of Alpha-Rhodobacterales increased toward ultraoligotrophic gyre waters,

241   characterized by low Chla concentrations (N=107, R=−0.42, p<0.001) and deeper chlorophyll

242   maxima (N=113, R=0.42, p<0.001). While the negative correlation between the contribution of

243   Gamma-Halieaceae and Alpha-Rhodobacterales was observed before (Ferrera *et al.*, 2014;

244   Bibiloni-Isaksson *et al.*, 2016; Auladell *et al.*, 2019), in those cases, higher relative abundances

245   of Alpha-Rhodobacterales were linked to higher concentrations of Chla and, in general, to

246   higher nutrient levels. While those studies were conducted in coastal stations, the Malaspina

247   Expedition occupied open-ocean stations, yet covered contrasting regions, from some relatively

248   eutrophic areas (such as the Equator, South African provinces or the South Australian Bight) to

249   the oligotrophic open ocean gyres. Just like seasonal ecotypes have been defined within the

250   Alpha-Rhodobacterales based on 16S rRNA gene sequencing (Mena *et al.*, 2020), one possible

251   explanation for the observed contrasting results is that closely related, but ecologically different

252   Alpha-Rhodobacterales could be divided into an ecotype with a preference for productive

253   regions such as coastal areas and an ecotype dominant in oligotrophic environments like the

254   oceanic gyres.

255   Representatives from the Alpha-Sphingomonadales and 'Other Alpha' were scarce across the

256   surface ocean with some localized exceptions (see Fig. 2B). The relative abundances of Alpha-

257   Sphingomonadales-like members correlated positively with prokaryotic heterotrophic

258   production (N=113, R=0.44; *p*<0.005), prokaryotic cell volume and total biomass (N=113,

259   R=0.52; *p*<0.001 and R=0.46; *p*<0.005 respectively). Interestingly, in stations where

260 Sphingomonadales dominated, this dominance was due to a single ASV (ASV512), which

261 contributed up to 50% of the total AAP community reads. This ASV is related to an uncultured

262 bacterial sequence (96% of identity in the *puf*M nucleotide sequence) detected in the Beaufort

263 Sea (Boeuf *et al.*, 2013) but does not resemble any cultured AAP bacteria. Thus, information on

264 the physiology of the organism behind this sequence is missing. In any case, its widespread

265 distribution from the Arctic to the tropical oceans and its ability to dominate communities under

266 different conditions are remarkable. Other ASVs that could only be assigned to the Alpha-

267 Proteobacteria level (and were thus grouped as "Alpha-Others") dominated communities in

268 some stations across the whole transect (Fig. 2), such as one in the Pacific Ocean (Station 93),

269 one in the Atlantic Ocean (Station 1, adjacent to the Strait of Gibraltar) and several stations in

270 the South Australian Bight (Stations 72, 72 and 75), coinciding with the South Subtropical

271 Convergence zone (SSTC Longhurst province). In these stations, two ASVs were dominant,

272 ASV860 in the Atlantic Ocean and ASV1102 in the Pacific and in the South of Australia.

273 Although we could not classify them further and they do not have close cultured representatives,

274 they are very similar to sequences from previous studies. In particular, ASV860 is very similar

275 (99.5% identity) to a sequence retrieved from the Atlantic Ocean (OTU SPIT34 in (Lehours and

276 Jeanthon, 2015), accession number KM654597) and ASV1102 is identical to an uncultured

277 bacterium found in the East coast of Tasmania (Bibiloni-Isaksson *et al.*, 2016). This ASV

278 appears to be associated to low water temperature (correlation with temperature, N=113, R

279 $=-0.40$, p<0.001) and higher concentrations of nitrate (N=89, NO_3^-, R=0.47, p<0.001). Finally,

280 Betaproteobacteria representatives were scarce along the dataset (only 11 sequences with very

281 low abundances) as expected, since this group is mostly absent in the marine environment

282 (Ferrera *et al.*, 2014; Bibiloni-Isaksson *et al.*, 2016; Lehours *et al.*, 2018; Auladell *et al.*, 2019).

283 The taxonomic composition hitherto described here is based on the relative abundances of

284 ASVs. In order to obtain data on the absolute abundance of AAP bacteria, we quantified them

285 by epifluorescence microscopy. Unfortunately, we were not able to quantify AAP abundance

286 along the entire transect, but only in a subset of 21 stations (samples for other stations were

287   either not available or of insufficient quality). Yet, the stations for which the abundance was

288   quantified, AAPs were uniformly distributed along the transect (except for the Indian Ocean for

289   which samples were not available) and should provide a good representation of the abundance

290   variation along the tropical and subtropical oceans. Abundances ranged between $5.52 \cdot 10^2$ and

291    $6.2 \cdot 10^4$ cells/mL and the percentage of AAP bacteria within the prokaryotic community varied

292   between 0.1 and 10% (Fig. S7). Although we estimated AAP abundances for a subgroup of

293   samples, their absolute and relative cell abundance are in line with the abundances reported in

294   previous studies using the same methodology (see data reviewed in Koblížek, 2015). We

295   observed higher AAP bacteria concentrations at lower latitudes (correlation between latitude

296   and %AAP, N=21, R=0.50, p=0.024, Fig. S7), and interestingly, we did not find any

297   relationship between the abundance of AAP bacteria and the taxonomic composition of the

298   AAP communities (see Fig. S7). This observation indicates that despite several communities

299   were dominated by different ASVs, there was not a single dominant taxonomic group associated

300   to the increase in absolute AAP bacterial abundances.

301   ***ASVs displaying bimodal and lognormal Spatial Abundance Distributions (SpADs) dominate***

302   ***AAP assemblages***

303   We explored the spatial patterns of AAPs and found that most of the individual taxa (64%) were

304   only found in one oceanic region, and these sequences represented only around 10% of the total

305   number of reads. On the contrary, very few sequences (30 ASVs) appeared in all sampled areas,

306   and they represented almost 50% of the total number of reads. Within this group of prevalent

307   sequences, we found representatives of all the taxonomic groups defined above (data not

308   shown), and thus dominance or rarity of individual sequences does not seem to be linked to

309   taxonomy. For this reason, to better understand the ecological behavior of AAP taxa we went

310   beyond their taxonomy affiliations by analyzing the Spatial Abundance Distribution (SpADs) of

311   the individual taxa, an approach that has proven as a useful tool for identifying groups of

312   bacteria sharing similar spatial patterns regardless of their identity (Niño-García *et al.*, 2016;

313   Ruiz-González *et al.*, 2020). In particular, the SpADs analysis classifies individual taxa into

different categories according to the shape of their abundance distribution (see Experimental Procedures). The different shapes can be interpreted as ecological traits because the abundance distribution of a given taxon will be the result of the combination of its physiological capacities, environmental tolerances or ability to persist under unfavorable conditions, but also of the external factors controlling its abundance. This approach has been previously used to explore the mechanisms behind the ubiquity or rarity of taxa within aquatic prokaryotic or picoeukaryotic communities (Niño-García *et al.*, 2016; Mangot *et al.*, 2018; Ruiz-González *et al.*, 2019; LaBrie *et al.*, 2021), but to our knowledge this is the first time that it is restrictively applied to a functional group.

We only found 2 ASVs displaying normal-like distributions presenting high abundances and broad environmental tolerances (Fig. S8A,C); the bimodal category (N=15 ASVs) included ASVs with lower average abundances and occurrence, likely representing less generalist taxa whose presence is restricted to specific areas, while lognormal (N=228) and logistic (N= 872) distributions, which represented the majority of cases, were characteristic of globally rare and endemic AAPs (Fig. S8). AAP assemblages in the surface ocean were dominated by bimodal and lognormal ASVs (Fig. 3C), mostly associated to Gamma-Halieaceae, Alpha-Rhodobacterales and Alpha-Sphingomonadales groups (Fig. S8), and only few communities were dominated by either normal-like or logistic taxa. The two normal-like ASVs were Sphingomonadales-like (Fig. S8), suggesting large environmental tolerances for this category, regardless of its relatively low contribution in most stations (Fig. 2B).

Communities dominated by logistic ASVs in our study appeared spatially clustered and coincided with productive regions such as the Benguela coastal province in the South Atlantic, the Caribbean Sea, the Equatorial Pacific and the station nearest to the Strait of Gibraltar (Fig. 3C). In fact, the relative abundances of logistic ASVs showed a significant positive correlation with the mean chlorophyll *a* concentration across stations (N=107, R=0.43, p<0.0001), pointing to local selection of globally-rare opportunistic AAP bacteria in nutrient-rich areas, as shown for prokaryotic and picoeukaryotic bloomers (Ruiz-González *et al.*, 2019; Logares *et al.*, 2020).

341   Yet, the overall distribution of SpADs in our study differs from that reported by Ruiz-González

342   *et al.* (2019) for the whole prokaryotic communities from the same Malaspina Expedition

343   surface samples. Whereas bimodal and lognormal ASVs were prevalent in AAP communities,

344   bulk prokaryotic assemblages were dominated by a few cosmopolitan normal-like OTUs

345   (operational taxonomic units). For the bulk community, bimodal and logistic OTUs increased in

346   stations with anomalies in temperature and productivity with respect to the average values. This

347   different distribution suggests that AAP bacteria are less homogeneously distributed than the

348   bulk bacterioplankton (or at least than their dominant members) and that changes in the

349   environment have a large effect on the AAP communities, promoting larger compositional shifts

350   across environmental gradients and the increase of habitat specialists within this functional

351   group.

352   ***Environmental setting drives marked differences in community structure among oceanic***

353   ***regions***

354   We further analyzed the AAP community structure along the Malaspina transect using Bray-

355   Curtis dissimilarity metrics. The overall Bray-Curtis dissimilarity (mean 0.85 ± 0.15) was

356   significantly higher than that described for prokaryotic and picoeukaryotic assemblages in the

357   same transect (prokaryotes mean = 0.61 ± 0.19; picoeukaryotes mean = 0.74 ± 0.08, Logares *et*

358   *al.*, 2020), meaning that changes in the species composition and abundance distributions across

359   AAP communities are larger than across bulk microbial groups. The higher beta diversity

360   observed is consistent with these results, showing that AAP communities are mainly composed

361   of habitat specialists (ASVs with a bimodal distribution) and rare taxa (lognormal distribution),

362   while bulk prokaryotic communities were dominated by few abundant and ubiquitous species

363   (Ruiz-González *et al.*, 2019).

364   Moreover, we explored which abiotic or biotic variables influenced AAP community structure

365   across the global ocean through PERMANOVA (p<0.001), and temperature, salinity and Chla

366   emerged as the most important variables (Table S2). When we pulled all samples together in a

367   distance-based redundancy analysis (dbRDA, Fig. S9), the first two axis explained only 16% of

368   the variation, and there was no obvious clustering of communities based on region or province,

369   even though simple analyses of variance showed statistical differences (p=0.001, Table S3).

370   Thus, we further analyzed the samples for each ocean separately (Fig. 4). Higher percentages of

371   variation were explained by the two first axis (28.9%, Pacific Ocean, 29.1%, Atlantic Ocean and

372   40.2% Indian Ocean), and the main variables associated were temperature (with 1st axis) and

373   salinity (with 2nd axis) (Fig. 4 and Fig. S10). Stations from the same Longhurst province

374   clustered together in most cases and, in general, we observed that communities from adjacent

375   locations where more similar to each other than communities from distant stations, pointing to

376   gradual changes in community structure along areas of the surface ocean (Fig. 4). Previous

377   studies restricted to specific areas of different ocean basins observed that the composition of

378   AAP bacteria varied with the trophic conditions (Jiao *et al.*, 2007; Yutin *et al.*, 2007), while

379   studies from the Arctic Sea showed that the hydrological context of the water masses were also

380   relevant (Boeuf *et al.*, 2013; Lehours and Jeanthon, 2015). Our results indicate that temperature,

381   salinity and the general environmental context (as defined by the Longhurst provinces) largely

382   structure AAP surface communities in the global tropical and subtropical ocean.

383   ***Community dissimilarity increases with increasing geographic distance***

384   To visualize the turnover of AAP communities along the Malaspina track we plotted taxonomic

385   community dissimilarities versus geographic distance (Fig. 3A) –considering only pairwise

386   comparisons within the same ocean– which unveiled a strong pattern of biogeography, that is, a

387   remarkable increase of community dissimilarity with increasing distance within each ocean. To

388   further explore community turnover at a fine scale, we explored the sequential changes of beta

389   diversity across the whole sampling transect and found 17 stations displaying Bray-Curtis (BC)

390   dissimilarity values > 0.75 which can be interpreted as sites of abrupt changes in community

391   structure (Fig. 3B). In general, the pattern of sequential beta diversity followed the changes

392   shown through the SpAD analysis (see Fig. 3B and C). Stations showing the highest

393   dissimilarity values (BC > 0.9) were located in the South African Atlantic Coast (BENG and

394   EARF) and the Costa Rica Dome (PNEC), where some sequences –belonging to logistic ASVs–

395 presented remarkably high relative abundances, associated to an increase in Chla concentration

396 (see above). Other sites (BC values 0.75−0.9) were in the borders of several Longhurst

397 provinces, such as the South Subtropical Convergence (SSTC), the Pacific Equatorial

398 Divergence (PEQD), the North Atlantic tropical gyre (NATR) or provinces in the South

399 Atlantic (SATL, BENG and EARF) (see Fig. 3B). The partition of the surface ocean into

400 biogeographical provinces was proposed by Longhurst (1998) based on changes in

401 environmental oceanic variables and their annual dynamics. This subdivision has been

402 extensively used in several studies analyzing the surface ocean microbiota and has been proven

403 to explain their biogeographic structure (see for example Friedline *et al.*, 2012; Frank *et al.*,

404 2016; Milici *et al.*, 2016; Logares *et al.*, 2020; Ruiz-González *et al.*, 2020). We indeed observed

405 that different Longhurst provinces harbored distinct AAP communities but it should also be

406 considered that the borders between these provinces are dynamic and change seasonally

407 (Reygondeau *et al.*, 2013). For example, during the boreal summer, the Northwest Atlantic

408 subtropical gyral (NASW, not included in this sampling) and North Atlantic tropical gyre

409 (NATR) provinces tend to become mixed and an infiltration from the NASW province into the

410 NATR province has clearly been observed (see Fig. 4 in Reygondeau *et al.*, 2013). In this same

411 area (NATR province) and timing (during June and July) we observed two samples (Stations

412 133 and 135) that differed largely from the rest, as seen by their different taxonomic

413 composition (see Leg 7 in Fig. 2) and high BC sequential dissimilarities (Fig. 3B). This

414 difference could not be attributed to any measured environmental variable. Although this is

415 speculative, the infiltration of water from a different province or some other physical

416 oceanographic feature (Baltar *et al*., 2010, 2016, Bagnaro *et al*., 2020), could explain the abrupt

417 changes seen in the North Atlantic in our study. Overall, we observed that AAP communities

418 displayed strong biogeographic patterns, with large dissimilarities across the surface ocean

419 which surpassed in magnitude those described for the bulk surface ocean microbiota.

420 ***Selection has a prominent role in structuring AAP communities***

421 Our analyses of AAP community turnover clearly showed a biogeography pattern across the

422 surface ocean. The different patterns of diversity and species composition across spatial scales

423 result from the combination of different ecological processes, such as selection, dispersal, or

424 drift (Vellend, 2010). Changes in microbial species composition across space could be related to

425 selection processes driven by changes in environmental variables (Fig. 4). Nevertheless, we

426 observed that environmental conditions at adjacent stations were generally comparable, so these

427 changes could also arise from dispersal limitation imposed by physical oceanographic features

428 (Baltar *et al.*, 2010, 2016, Bagnaro *et al.*, 2020). In fact, previous studies have shown that

429 oceanic features such as boundaries between different ocean regions can act as strong barriers

430 and delimit the distribution of microbes in the ocean (Baltar *et al.*, 2016, Raes *et al.* 2018).

431 Whether the pattern observed is the result of environmental selection and/or dispersal limitation

432 cannot be determined based on our previous analysis (see also Hanson *et al.*, 2012). Thus, to

433 further investigate the ecological processes shaping AAP communities across the global surface

434 ocean, we applied the approach proposed by Stegen *et al.* (2013), which quantitatively estimates

435 the influence of selection, dispersal and drift based on the phylogenetic turnover of

436 communities. Since this method relies solely on the phylogeny of the *puf*M gene and on null

437 models (randomization), it avoids the problem of unmeasured environmental variables that can

438 potentially be associated with selection or dispersal (Stegen *et al.*, 2013). The influence of

439 selection was estimated using the β-nearest taxon index (βNTI), which is the difference between

440 the observed phylogenetic turnover for a given pair of communities and the null model after 999

441 randomizations (see Experimental Procedures). The values of βNTI were calculated for the all

442 the pairwise community comparisons possible in the dataset. We found that ~23% of the

443 pairwise comparisons had values of βNTI < −2, which implies that there is a shorter

444 phylogenetic distance within these pairs of communities, than expected by chance (Stegen *et al.*,

445 2012). Lower turnover between communities is expected when environmental conditions are

446 very similar and there is a –homogeneous– selection of closely related taxa in these

447 communities. Likewise, ~27% of the pairwise comparisons had βNTI > 2, which is associated

448 with a greater phylogenetic distance than the expected under a null model and can be interpreted

449   as different environmental conditions –heterogeneously– selecting distantly related taxa (Stegen

450   *et al.*, 2012). Overall, ~50% of the observed turnover could be explained by selection, with

451   homogeneous and heterogeneous selection being almost equally important at a global scale (Fig.

452   5). Within samples located in the same Longhurst province, homogeneous selection had an

453   important role, as the main ecological process in most provinces (see Fig. S11). In turn,

454   heterogeneous selection had a modest role within Longhurst provinces, and only operated in

455   some provinces. Based on βNTI values of comparisons between adjacent stations,

456   heterogeneous selection was high in areas were logistic ASVs dominated (data not shown),

457   pointing towards the selection of rare taxa in productive areas. These results are in line with

458   previous studies that already pointed to selection as a major ecological process structuring AAP

459   communities in both spatial (Lehours *et al.*, 2018) and temporal studies (Auladell *et al.*, 2019).

460

461   For the remaining pairwise comparisons, the value of Bray–Curtis-based Raup–Crick (RC_{bray})

462   characterized the magnitude of deviation between the observed BC and the null BC. RC_{bray}

463   distribution varied between -1 and 1, and only values $|RC_{bray}| > 0.95$ were considered as

464   significant departures from drift (see Experimental Procedures and Fig. 5).

465   Dispersal limitation explained ~20% of the community turnover ($RC_{bray} > 0.95$) while

466   homogenizing dispersal was observed only for 18 pairwise comparisons (0.7%). The limited

467   role of dispersal limitation structuring AAP communities could be expected, since distant

468   microbial communities are known to be connected on a global scale, under what has been

469   described as the "Microbial Conveyor Belt" (Mestre and Höfer, 2021). Finally, almost ~30% of

470   the community turnover was explained by drift (stochastic processes), as the differences

471   between the null and the observed beta diversity were not significant. Stochastic processes are

472   difficult to predict and to distinguish from other ecological processes (Zhou and Ning, 2017),

473   however, they play an important role in microbial community assembly (Evans *et al*,. 2017,

474   Graham and Stegen, 2017) and their importance increases under high selection and low

475 dispersal (Fodelianakis *et al.*, 2020), as it happens in AAP communities across the surface

476 ocean.

477 Remarkably, the observed pattern is different from that reported for whole prokaryotic

478 communities along the same transect (Logares *et al.*, 2020), which appeared to be structured to a

479 similar extent by both selection and dispersal (representing each process ~25% of the

480 community turnover). In contrast, dispersal limitation had a much more important role in

481 structuring picoeukaryotic communities (~65%), likely due to their larger cell sizes and lower

482 abundances (Logares *et al.*, 2020). The relatively higher importance of selection mechanisms in

483 AAP communities suggests that AAP bacteria are more affected by small changes in the

484 environmental conditions than the prokaryotic community as a whole. As we have shown

485 above, the community turnover measured as Bray-Curtis dissimilarity is higher in this

486 functional group than in the bulk picoplankton, pointing to higher changes in the composition

487 and structure of AAP communities over short distances. Besides, while prokaryotic assemblages

488 are dominated by few cosmopolitan and very abundant taxa, AAP assemblages are mainly

489 composed by taxa classified as rare or habitat specialists, with more restricted environmental

490 tolerances.

491 **Concluding remarks**

492 In this study we described the global diversity and community structure patterns of marine AAP

493 bacteria in the tropical and subtropical oceans. Alpha diversity varied across biogeographical

494 provinces mainly related to temperature, salinity and trophic status and showed remarkably low

495 values in the more productive Longhurst provinces. AAP communities along the surface ocean

496 were mainly composed of members of the Halieaceae (Gammaproteobacteria), which were

497 adapted to a large range of environmental conditions, and by different clades of the

498 Alphaproteobacteria, that seemed to dominate under particular circumstances, such as in the

499 oligotrophic gyres. These taxa were not randomly distributed but appeared to be spatially

500 structured, with a marked succession of samples dominated by either one or the other class.

501 Communities from adjacent stations shared more taxonomic similarities, that is, community

502 dissimilarity increased with increasing distance, which resulted in a remarkable biogeographical

503 pattern. However, this pattern was to a large extent the result of –homogeneous and

504 heterogeneous– selection of individual taxa, while dispersal and drift had less of a role in

505 shaping the structure of AAP bacterial communities. While the seasonal patterns of AAPs have

506 been shown to be notably comparable to those of the bulk bacterioplankton, at a large-scale,

507 AAP communities seem to have their own spatial patterns that do not mimic those of the bulk

508 picoplankton. Of the measured environmental variables, temperature, salinity and Chla were

509 found to influence AAP community structure. Small changes in environmental conditions

510 translated into significant changes in AAP communities, and therefore, several habitat

511 specialists and many rare species dominated their communities. The photoheterotrophic

512 metabolism, high growth rates and high predation pressure on AAP bacteria, among other

513 attributable traits to this functional group, could explain the stronger role of selection in this

514 group compared to the bulk surface ocean microbiota. Overall, our results represent the most

515 comprehensive study investigating the global biogeography of AAP communities and shows

516 how different ecological processes explain these patterns.

517

518 **EXPERIMENTAL PROCEDURES**

519 **Sample collection**

520 The Malaspina 2010 Expedition took place between December 2010 and July 2011 (Duarte,

521 2015). Samples were collected in 113 stations across the tropical and subtropical waters of the

522 Pacific, Atlantic and Indian oceans. At each station, about 12 L of surface seawater (3 m depth)

523 were collected with a large (30 L) oceanographic bottle. Simultaneously, a CTD profiler was

524 used to profile temperature, salinity, conductivity, fluorescence and dissolved oxygen. Seawater

525 was prefiltered through a 20 μm nylon mesh and a 3 μm filter onto a 0.2 μm Millipore

526 polycarbonate filter. Samples were conserved at −80ºC until further processing. Samples for

527 enumerating AAP cells were pre-filtered through a 200 μm mesh and filtered onto 0.2 μm

528 polycarbonate filters. Cells were enumerated by infra-red epifluorescence microscopy in 21

529   stations as described in (Ferrera *et al.*, 2014). The environmental biotic and abiotic parameters

530   used in this study were determined as reported in Estrada *et al.*, (2016) and Ruiz-González *et*

531   *al.*, (2019).

DNA extraction, *puf*M amplification, sequencing and ASV generation

533   DNA was extracted from the 0.2 µm filter using the phenol-chloroform protocol as described in

534   (Massana *et al.*, 1997). Partial amplification of the *puf*M gene (~245 bp fragments) was done in

535   50 µl reactions using primers pufM forward (5′-TACGGSAACCTGTWCTAC-3′, (Béjà *et al.*,

536   2002)) and puf_WAW reverse (5′-AYNGCRAACCACCANGCCCA- 3′, (Yutin *et al.*, 2005))

537   as described in Auladell *et al.* (2019). DNA was sequenced in an Illumina MiSeq sequencer

538   (2×250 bp, Research and Testing Laboratory; http://rtlgenomics.com/). After sequencing, we

539   used cutadapt v1.16 (Martin, 2013) to remove primers and DADA2 v1.10 (Callahan *et al.*,

540   2016) to infer amplicon sequence variants (ASV) with the following parameters: maxEE =

541   c(2,6) and trunclen = c(210,150). After filtering chimeras and spurious sequences, we kept 82%

542   of the initial number of reads (mean 24173, min. 4503, max. 79968). To be able to compare our

543   data with previous studies that used OTUs (Operational Taxonomic Units), we clustered the

544   ASVs with UCLUST v10.0 (Edgar, 2010) at 94% similarity, the threshold usually employed for

545   the *puf*M gene (Zeng *et al.*, 2007).

Phylogenetic classification

547   We used phylogenetic placement for predicting the taxonomic assignment of the *puf*M gene

548   short sequences. Due to the lack of comprehensive public databases for AAP bacteria, we built a

549   custom made *puf*M database retrieving more than 750 sequences longer than 600 bp from the

550   Genome Taxonomy Database (GTDB) and GenBank, as well as from metagenomic datasets

551   from the Tara *Oceans* Expedition (Sunagawa *et al.*, 2015), the Malaspina Expedition

552   (unpublished), the Global Ocean Survey (GOS) (Yutin *et al.*, 2007; Cuadrat *et al.*, 2016), and

553   the Blanes Bay Microbial Observatory (Auladell *et al.*, 2019). Alignment was done using the

554   *Decipher* R package (Wright, 2016) and MAFFT v.7 (Katoh and Standley, 2013). After a

555   manual curation using AliView v1.26 (Larsson, 2014), we kept 673 sequences. A phylogenetic

556   tree was constructed using RAxML v8.2 (Stamakis 2014) (GTRGAMMA model, 100

557   bootstraps), and visualised using iTOL (Letunic and Bork, 2011), see Fig. S5. Finally, to infer

558   the phylogeny of the amplicon sequence variants, we applied the Evolutionary Placement

559   Algorithm v0.3.5 (Barbera *et al.*, 2019).

560   **Data analyses**

561   All statistical analyses were performed using R v3.6.3 (R Core Team 2020). The ASV table was

562   rarefied down to 4500 reads per sample using the *vegan* package. Alpha diversity was estimated

563   using Chao1 and Shannon diversity indices (Chao and Lee, 1992), with the *phyloseq* package.

564   Post-Hoc Tukey tests were employed to see if there were statistically significant differences

565   between the diversity of different regions. To test whether diversity was influenced by

566   environmental conditions, we performed Pearson correlations between a selection of

567   environmental variables and the diversity indices. We also compared the diversity of AAP

568   bacteria with those of the bulk prokaryotic communities using the 16S rRNA data presented in

569   Ruiz-González *et al.* (2019) and the picoeukaryotic community data presented in Logares *et al.*

570   (2020), both from the same samples taken during the Malaspina Expedition. Community

571   composition was analyzed and described using the *phyloseq* package in R.

572   In order to explore the spatial patterns of individual AAP bacteria across space, we analyzed the

573   abundance distribution of each *puf*M sequence across all samples. We used the rarefied table of

574   counts ($\log_{10}(x + 1)$ transformed) to select the statistical distribution that best fitted the spatial

575   abundance distribution (SpAD) of each ASV, as described in Niño-García *et al.* (2016) and

576   Ruiz-González *et al.* (2019). We could classify all ASVs into four SpAD categories: "normal-

577   like" ASVs showed a normal statistical distribution, which has previously been associated with

578   globally abundant and widespread taxa, and which might represent habitat generalists (Niño-

579   García et al. 2016, Ruiz-González et al. 2019). The distribution of "Bimodal" ASVs is

580   characterized by two density peaks, with the first one commonly corresponding to zero cases,

581   and could be considered less generalists because they are detected in certain regions only and

582   their average abundances are also lower. Finally, ASVs classified as "logistic" and "lognormal"

583 present a distribution with a zero-abundance mode, and they have been shown to comprise

584 mostly rare sequences (for more details on the analysis see Niño-García *et al*. 2016 and Ruiz-

585 González *et al*. 2019). For each category, we calculated the mean abundance and occurrence of

586 ASVs. We also estimated the individual environmental breath as the range of temperature,

587 salinity, Chla, and dissolved oxygen concentration in which each of the ASVs within the

588 different categories were detected.

589 The exploration of the main environmental drivers explaining the structure of AAP

590 communities was done using a Bray-Curtis dissimilarity matrix, built with the *vegdist()* function

591 from the *vegan* package and visualized in a distance-base redundancy analysis (dbRDA), with a

592 previous selection of significant environmental variables (PERMANOVA p <0.01). Permutation

593 tests (*adonis()* function from *vegan* package) were employed to examine community differences

594 among the six oceanic regions (South Pacific, North Pacific, North Atlantic, South Atlantic,

595 Indian and South Australian Bight) and Longhurst oceanographic provinces (Longhurst, 1998).

596 We used Mantel tests (1000 permutations) to compare the changes in the structure of AAP

597 communities between stations with differences in temperature, salinity and Chla. Additionally,

598 we performed partial mantel tests to compare the community structures of AAP, prokaryotes

599 and picoeukaryotes, removing the effect of temperature, salinity and Chla. The Bray-Curtis

600 dissimilarity matrix was also used to analyze the spatial community structure turnover, and to

601 explore sequential changes along the Malaspina transect, by comparing each sample with the

602 one sampled immediately before.

603 Finally, to quantify the relative importance of selection, dispersal and drift as processes

604 structuring the communities of AAP bacteria, we followed the framework developed by Stegen

605 *et al.* (2013). This approach assumes that there is a phylogenetic signal (Cavender-Bares *et al*.

606 2009) in the ASVs optimal habitat conditions (i.e., the habitat preferences of closely related taxa

607 are more similar than the preferences of distantly related taxa). To confirm this assumption, we

608 firstly compared ASVs niche distances (using temperature, salinity and Chla) and ASVs

609 phylogenetic distances using a Mantel correlogram test. We detected phylogenetic signal in the

610   *puf*M gene marker over relatively short phylogenetic distances (Fig. S12), as previously shown

611   with other marker genes (e.g.: Stegen *et al.*, 2013; Dini-Andreote *et al.*, 2015; Huber *et al.*,

612   2020; Logares *et al.*, 2020).

613   Then, to analyze the influence of selection we calculated the β-mean nearest taxon distance

614   (βMNT) metric, which quantifies the mean phylogenetic distances between two communities,

615   and compared them to a random expectation (999 randomizations). The difference between the

616   observed phylogenetic turnover (or βMNT) and the values obtained with the null model are

617   denoted as β-nearest taxon index (βNTI). Absolute βNTI values above 2 ($|βNTI| > 2$) indicate

618   that coexisting taxa are more closely related than expected by chance, thus pointing to the action

619   of selection. Afterwards, to differentiate whether drift or dispersal were the main structuring

620   processes, we calculated the Raup-Crick metric (Chase *et al.*, 2011) using Bray-Curtis

621   dissimilarities (RC_{bray}) (Chase *et al.*, 2011; Stegen *et al.*, 2013). RC_{bray} compares the measured

622   beta diversity to the beta diversity obtained by the null model (999 randomizations) that would

623   be obtained under random community assembly (drift). RC_{bray} values between -0.95 and $+0.95$

624   point to a community assembly governed by drift. On the contrary, RC_{bray} values ≥ 0.95 or $\leq$

625   0.95 indicate that community turnover is driven by dispersal limitation or homogenizing

626   dispersal, respectively (Stegen *et al.*, 2013). For this analysis, raw ASV sequences were aligned

627   with AliView v1.26 (Larsson, 2014), aligned sequences were visually curated with Seaview

628   (Gouy et al. 2010) and the phylogenic tree was constructed using FastTree v2.1.9 (Price et al.

629   2009). The βMNTD and βNTI metrics were calculated using the R package *Picante* (Kembel *et*

630   *al.*, 2010) and the RC_{bray} was calculated with the *raup_crick_abundance* function following

631   Stegen *et al.*, (2013). These analyses were performed in R v3.6.3 (R Core Team 2020) and

632   codes are available in Github *(*https://gitlab.com/crgazulla/malaspina_aaps*)*. Sequence data have

633   been deposited in the NCBI Sequence Read Archive (SRA) under BioProject ID

634   PRJNA736051.

635   **Acknowledgements**

649

References

651 Auladell, A., Sánchez, P., Sánchez, O., Gasol, J.M., and Ferrera, I. (2019) Long-term
652   seasonal and interannual variability of marine aerobic anoxygenic
653   photoheterotrophic bacteria. *ISME Journal* **13**: 1975–1987.

654 Bagnaro, A., Baltar, F., Brownstein, G. Lee, W.G., Morales, S.E., Pritchard, D.W., and
655   Hepburn, C.D. (2020) Reducing the arbitrary: fuzzy detection of microbial
656   ecotones and ecosystems – focus on the pelagic environment. *Environmental*
657   *Microbiome* 15, 16 https://doi.org/10.1186/s40793-020-00363-w

658 Baltar, F., Arístegui, J., Gasol, J. Lekunberri, I., and Herndl, G.J. (2010) Mesoscale
659   eddies: hotspots of prokaryotic activity and differential community structure in the
660   ocean. *ISME J* 4, 975–988. https://doi.org/10.1038/ismej.2010.33

661 Baltar, F., Currie, K., Stuck, E., Roosa, S., and Morales, S.E. (2016) Oceanic fronts:
662   Transition zones for bacterioplankton community composition. *Environmental*
663   *Microbiology Reports* **8**: 132–138.

664 Barbera, P., Kozlov, A.M., Czech, L., Morel, B., Darriba, D., Flouri, T., and Stamatakis,
665   A. (2019) EPA-ng: Massively Parallel Evolutionary Placement of Genetic
666   Sequences. *Systematic Biology* **68**: 365–369.

667 Béjà, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P., et al. (2000)
668   Bacterial Rhodopsin : Evidence for a New Type of Phototrophy in the Sea. *Science*
669   **289**: 1902–1906.

Béjà, O., Suzuki, M.T., Heidelberg, J.F., Nelson, W.C., Preston, C.M., Hamada, T., et al. (2002) Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415**: 630–633.

Bibiloni-Isaksson, J., Seymour, J.R., Ingleton, T., van de Kamp, J., Bodrossy, L., and Brown, M.V. (2016) Spatial and temporal variability of aerobic anoxygenic photoheterotrophic bacteria along the east coast of Australia. *Environmental Microbiology* **18**: 4485–4500.

Boeuf, D., Cottrell, M.T., Kirchman, D.L., Lebaron, P., and Jeanthon, C. (2013) Summer community structure of aerobic anoxygenic phototrophic bacteria in the western Arctic Ocean.

Callahan, B.J., Mcmurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. **13**, 581-583. doi: 10.1038/nmeth.3869.

Cavender-Bares J, Kozak K.H, Fine P.V.A, Kembel S.W. (2009). The merging of community ecology and phylogenetic biology. Ecol Lett 12: 693–715.

Chao, A. and Lee, S.M. (1992) Estimating the number of classes via sample coverage. *Journal of the American Statistical Association* **87**: 210–217.

Chase, J.M., Kraft, N.J.B., Smith, K.G., Vellend, M., and Inouye, B.D. (2011) Using null models to disentangle variation in community dissimilarity from variation in α-diversity. Ecosphere 2(2), 1-11 doi: https://doi.org/10.1890/ES10-00117.1

Cottrell, M.T. and Kirchman, D.L. (2009) Photoheterotrophic microbes in the arctic ocean in summer and winter. *Applied and Environmental Microbiology* **75**: 4958–4966.

Cuadrat, R., Ferrera, I., Grossart, H., and Dávila, A. (2016) Picoplankton Bloom in Global South? A High Fraction of Aerobic Anoxygenic Phototrophic Bacteria in Metagenomes from a Coastal Bay. *Omi A J Integr Biol*. **20**: 76–87.

DeLong, E.F. and Béjà, O. (2010) The light-driven proton pump proteorhodopsin enhances bacterial survival during tough times. *PLoS Biology* **8**: 1–5.

Dini-Andreote, F., Stegen, J.C., van Elsas J.D., Salles, J.F. (2015) Disentangling mechanisms that mediate the balance between stochastic and deterministic processes in microbial succession. *PNAS* **112**: E1326-E1332, doi: https://doi.org/10.1073/pnas.1414261112

Duarte, C.M. (2015) Seafaring in the 21st century: The Malaspina 2010 circumnavigation expedition. *Limnology and Oceanography Bulletin* **24**: 11–14.

Edgar R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. **26**:2460–1.

Estrada, M., Delgado, M., Blasco, D., Latasa, M., Cabello, A.M., Benítez-Barrios, V., et al. (2016) Phytoplankton across tropical and subtropical regions of the Atlantic, Indian and Pacific oceans. *PLoS ONE* **11**: 1–29.

Evans, S., Martiny, J. and Allison, S. (2017) Effects of dispersal and selection on stochastic assembly in microbial communities. *ISME* **11**: 176–185. https://doi.org/10.1038/ismej.2016.96

Ferrera, I., Borrego, C.M., Salazar, G., and Gasol, J.M. (2014) Marked seasonality of aerobic anoxygenic phototrophic bacteria in the coastal NW Mediterranean Sea as

714   revealed by cell abundance, pigment concentration and pyrosequencing of *pufM*
715   gene. *Environmental Microbiology* **16**: 2953–2965.
716 Ferrera, I., Gasol, J.M., Sebastián, M., Hojerova, E., and Koblížek, M. (2011)
717   Comparison of Growth Rates of Aerobic Anoxygenic Phototrophic Bacteria and
718   Other Bacterioplankton Groups in Coastal Mediterranean Waters *Appl Environ*
719   *Microbiol.* **77**: 7451–7458.
720 Ferrera, I., Sánchez, O., Kolářová, E., Koblížek, M., and Gasol, J.M. (2017) Light
721   enhances the growth rates of natural populations of aerobic anoxygenic
722   phototrophic bacteria. *ISME J* **11**: 2391–2393.
723 Fodelianakis, S., Valenzuela-Cuevas, A., Barozzi, A. *et al.* (2021) Direct quantification
724   of ecological drift at the population level in synthetic bacterial communities. *ISME*
725   *J* **15**, 55–66. https://doi.org/10.1038/s41396-020-00754-4
726 Frank, A.H., Garcia, J.A.L., Herndl, G.J., and Reinthaler, T. (2016) Connectivity
727   between surface and deep waters determines prokaryotic diversity in the North
728   Atlantic Deep Water. *Environmental microbiology* **18**: 2052–2063.
729 Friedline, C.J., Franklin, R.B., McCallister, S.L., and Rivera, M.C. (2012) Bacterial
730   assemblages of the eastern Atlantic Ocean reveal both vertical and latitudinal
731   biogeographic signatures. *Biogeosciences* **9**: 2177–2193.
732 Graham, E.D., Heidelberg, J.F., and Tully, B.J. (2018) Potential for primary
733   productivity in a globally-distributed bacterial phototroph. *ISME J* **12**: 1861–1866.
734 Graham, Emily B.; Stegen, James C. (2017). "Dispersal-Based Microbial Community
735   Assembly   Decreases   Biogeochemical   Function"   *Processes*   **5**:   65.
736   https://doi.org/10.3390/pr5040065
737 Hanson, C.A., Fuhrman, J.A., Horner-Devine, M.C., and Martiny, J.B.H. (2012)
738   Beyond biogeographic patterns: Processes shaping the microbial landscape. *Nature*
739   *Reviews Microbiology* **10**: 497–506.
740 Hojerová, E., Mašín, M., Brunet, C., Ferrera, I., Gasol, J.M., and Koblížek, M. (2011)
741   Distribution and growth of aerobic anoxygenic phototrophs in the Mediterranean
742   Sea. *Environmental Microbiology* **13**: 2717–2725.
743 Huber, P., Metz, S., Unrein, F. *et al.* (2020) Environmental heterogeneity determines the
744   ecological processes that govern bacterial metacommunity assembly in a
745   floodplain river system. *ISME J* **14**, 2951–2966. https://doi.org/10.1038/s41396-
746   020-0723-2
747 Jeanthon, C., Boeuf, D., Dahan, O., le Gall, F., Garczarek, L., Bendif, E.M., and
748   Lehours, A.C. (2011) Diversity of cultivated and metabolically active aerobic
749   anoxygenic phototrophic bacteria along an oligotrophic gradient in the
750   Mediterranean Sea. *Biogeosciences* **8**:1955–1970.
751 Jiao, N., Zhang, Y., Zeng, Y., Hong, N., Liu, R., Chen, F., and Wang, P. (2007) Distinct
752   distribution pattern of abundance and diversity of aerobic anoxygenic phototrophic
753   bacteria in the global ocean. *Environ Microbiol* **9**: 3091–3099.
754 Karsenti, E., Acinas, S.G., Bork, P., Bowler, C., de Vargas, C., Raes, J., *et al.* (2011) A
755   holistic approach to marine Eco-systems biology. *PLoS Biology* **9**: 7–11.

Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

Kembel, S.W., Cowan, P.D., Helmus, M.R., Cornwell, W.K., Morlon, H., Ackerly, D.D., *et al*. (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**: 1463–1464.

Kirchman, D.L. and Hanson, T.E. (2013) Bioenergetics of photoheterotrophic bacteria in the oceans. *Environmental Microbiology Reports* **5**: 188–199.

Koblížek, M. (2015) Ecology of aerobic anoxygenic phototrophs in aquatic environments. *FEMS Microbiology Reviews* **39**: 854–870.

Koblížek, M., Mašín, M., Ras, J., and Poulton, A.J. (2007) Rapid growth rates of aerobic anoxygenic phototrophs in the ocean. *Environ Microbiol* **9**: 2401–2406.

Kolber, Z.S., van Dover, C.L., Niederman, R.A., and Falkowski, P.G. (2000) Bacterial photosynthesis in surface waters of the open ocean. *Nature* **407**: 177–179.

LaBrie, R., Bélanger, S., Benner, R., and Maranger, R. (2021) Spatial abundance distribution of prokaryotes is associated with dissolved organic matter composition and ecosystem function. *Limnology and Oceanography* **66**: 575–587.

Larsson, A. (2014) AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**: 3276–3278.

Lehours, A. and Jeanthon, C. (2015) The hydrological context determines the beta-diversity of aerobic anoxygenic phototrophic bacteria in European Arctic seas but does not favor endemism. *Front Microbiol* **6**: 638.

Lehours, A.C., Cottrell, M.T., Dahan, O., Kirchman, D.L., and Jeanthon, C. (2010) Summer distribution and diversity of aerobic anoxygenic phototrophic bacteria in the Mediterranean Sea in relation to environmental variables. *FEMS Microbiology Ecology* **74**: 397–409.

Lehours, A.C., Enault, F., Boeuf, D., and Jeanthon, C. (2018) Biogeographic patterns of aerobic anoxygenic phototrophic bacteria reveal an ecological consistency of phylogenetic clades in different oceanic biomes. *Scientific Reports* **8**: 1–10.

Letunic, I. and Bork, P. (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* doi: 10.1093/nar/gkz239

Logares, R., Deutschmann, I.M., Junger, P.C., Giner, C.R., Krabberød, A.K., Schmidt, T.S.B., et al. (2020) Disentangling the mechanisms shaping the surface ocean microbiota. *Microbiome* **8**: 1–17.

Longhurst, A. (1998). Ecological geography of the sea. San Diego, CA: Academic Press.

Mangot, J.F., Forn, I., Obiol, A., and Massana, R. (2018) Constant abundances of ubiquitous uncultured protists in the open sea assessed by automated microscopy. *Environmental Microbiology* **20**: 3876–3889.

Martin, M. (2013) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17:10.

Mašín, M., Zdun, A., Stoń-Egiert, J., Nausch, M., Labrenz, M., Moulisová, V., and Koblížek, M. (2006) Seasonal changes and diversity of aerobic anoxygenic phototrophs in the Baltic Sea. *Aquatic Microbial Ecology* **45**: 247–254.

800 Massana, R., Murray, A.E., and Preston, C.M. (1997) Vertical Distribution and
801   Phylogenetic Characterization of Marine Planktonic. *Microbiology* **63**: 50–56.
802 Mena, C., Reglero, P., Balbín, R., Martín, M., Santiago, R., and Sintes, E. (2020)
803   Seasonal Niche Partitioning of Surface Temperate Open Ocean Prokaryotic
804   Communities. *Frontiers in Microbiology* **11** 28;11:1749. doi:
805   10.3389/fmicb.2020.01749.
806 Mestre, M.; Höfer, J. (2021) The Microbial Conveyor Belt: Connecting the Globe
807   through Dispersion and Dormancy. *Trends in Microbiology* **29**: 482-492.
808   https://doi.org/10.1016/j.tim.2020.10.007.
809 Milici, M., Tomasch, J., Wos-Oxley, M.L., Decelle, J., Jáuregui, R., Wang, H., et al.
810   (2016) Bacterioplankton biogeography of the Atlantic ocean: A case study of the
811   distance-decay relationship. *Frontiers in Microbiology* **7**: 1–15.
812 Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., *et al*. (2020)
813   A genomic catalog of Earth's microbiomes. *Nature Biotechnology* **39**: 499–509.
814   doi: 10.1038/s41587-020-0718-6.
815 Niño-García, J.P., Ruiz-González, C., and del Giorgio, P.A. (2016) Landscape-scale
816   spatial abundance distributions discriminate core from random components of
817   boreal lake bacterioplankton. *Ecology Letters* **19**:1506–1515.
818 Obiol, A., Giner, C.R., Sánchez, P., Duarte, C.M., Acinas, S.G., and Massana, R. (2020)
819   A metagenomic assessment of microbial eukaryotic diversity in the global ocean.
820   *Molecular Ecology Resources* **20**: 718–731.
821 Oz, A., Sabehi, G., Koblízek, M., Massana, R., and Béjà, O. (2005) Roseobacter-like
822   bacteria in Red and Mediterranean Sea aerobic anoxygenic photosynthetic
823   populations. *Applied and Environmental Microbiology* **71**: 344–353.
824 Pinhassi, J., DeLong, E.F., Béjà, O., González, J.M., and Pedrós-Alió, C. (2016) Marine
825   Bacterial and Archaeal Ion-Pumping Rhodopsins: Genetic Diversity, Physiology,
826   and Ecology. *Microbiology and Molecular Biology Reviews* **80**: 929–954.
827 Price MN, Dehal PS, Arkin AP (2010) FastTree 2 – Approximately Maximum-
828   Likelihood Trees for Large Alignments. *PLoS ONE* **5**: e9490.
829   https://doi.org/10.1371/journal.pone.0009490
830 R Core Team (2014) R: a language and environment for statistical computing.
831 Raes, E.; Bodrossy, L., van de Kamp, J.; Bissett, A.; Ostrowski, M.; Brown, M.V. *et al*.
832   (2018) Oceanographic boundaries constrain microbial diversity gradients in the
833   South Pacific Ocean. *PNAS* **115**: E8266-E8275; doi: 10.1073/pnas.1719335115
834 Regaudie-de-Gioux A., Huete-Ortega M., Sobrino C., López-Sandoval D. C., González
835   N., Fernández-Carrera A., Vidal M., Marañón E., Cermeño P., Latasa, M., Agustí
836   S., and Duarte, C.M. (2019). Multi-model remote sensing assessment of primary
837   production in the subtropical gyres. *J. Marine Systems* **196**:97–106
838   https://doi.org/10.1016/j.jmarsys.2019.03.007
839 Reygondeau, G., Longhurst, A., Martinez, E., Beaugrand, G., Antoine, D., and Maury,
840   O. (2013) Dynamic biogeochemical provinces in the global ocean. *Global
841   Biogeochemical Cycles* **27**: 1046–1058.
842 Ruiz-González, C., Logares, R., Sebastián, M., Mestre, M., Rodríguez-Martínez, R.,
843   Galí, M., *et al*. (2019) Higher contribution of globally rare bacterial taxa reflects

environmental transitions across the surface ocean. *Molecular Ecology* **28**: 1930–1945.

Ruiz-González, C., Mestre, M., Estrada, M., Sebastián, M., Salazar, G., Agustí, S., *et al.* (2020) Major imprint of surface plankton on deep ocean prokaryotic structure and activity. *Molecular Ecology* **29**: 1820–1838.

Salazar, G., Cornejo-Castillo, F.M., Borrull, E., Díez-Vives, C., Lara, E., Vaqué, D., *et al.* (2015) Particle-association lifestyle is a phylogenetically conserved trait in bathypelagic prokaryotes. *Molecular Ecology* **24**: 5692–5706.

Schwalbach, M.S. and Fuhrman, J.A. (2005) Wide-ranging abundances of aerobic anoxygenic phototrophic bacteria in the world ocean revealed by epifluorescence microscopy and quantitative PCR. *Limnology and Oceanography* **50**: 620–628.

Sieracki, M.E., Gilg, I.C., Thier, E.C., Poulton, N.J., and Goericke, R. (2006) Distribution of planktonic aerobic anoxygenic photoheterotrophic bacteria in the northwest Atlantic. *Limnology and Oceanography* **51**: 38–46.

Stamatakis A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30:1312–3.

Stegen, J.C., Lin, X., Fredrickson, J.K., Chen, X., Kennedy, D.W., Murray, C.J., et al. (2013) Quantifying community assembly processes and identifying features that impose them. *ISME Journal* **7**: 2069–2079.

Stegen, J.C., Lin, X., Konopka, A.E., and Fredrickson, J.K. (2012) Stochastic and deterministic assembly processes in subsurface microbial communities. *ISME Journal* **6**: 1653–1664.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., et al. (2015) Structure and function of the global ocean microbiome. *Science* **348**: 1–10.

Teira, E., Logares, R., Gutiérrez-Barral, A., Ferrera, I., Varela, M.M., Morán, X.A.G., and Gasol, J.M. (2019). Impact of grazing, resource availability and light on prokaryotic growth and diversity in the oligotrophic surface global ocean. *Environmental microbiology* **21**: 1482-1496.

Tully, B.J., Graham, E.D., and Heidelberg, J.F. (2018) The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data* **5**: 1–8.

de Vargas, C, Audie, S., Henry, N., Decelle, J., Mahé, F., Logares, R., *et al.* (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1261605–1/11.

Vellend M. (2016) The theory of ecological communities. 1st ed. Woodstock, UK: Princeton University Press.

Villamaña, M., Marañón, E., Cermeño, P., Estrada, M., Fernández-Castro, B., Figueiras, F.G., Latasa, M., Otero-Ferrer, J.L., Reguera, B. and Mouriño-Carballido, B. (2019). The role of mixing in controlling resource availability and phytoplankton community composition. *Progress in Oceanography*, **178**, p.102181.

Waidner, L.A. and Kirchman, D.L. (2008) Diversity and distribution of ecotypes of the aerobic anoxygenic phototrophy gene *pufM* in the Delaware estuary. *Applied and Environmental Microbiology* **74**: 4012–4021.

Wright, E.S. (2016) Using DECIPHER v2.0 to analyze big biological sequence data in R. *R Journal* **8**: 352–359.

Yutin, N., Suzuki, M.T., and Béjà, O. (2005) Novel primers reveal wider diversity among marine aerobic anoxygenic phototrophs. *Applied and Environmental Microbiology* **71**: 8958–8962.

Yutin, N., Suzuki, M.T., Teeling, H., Weber, M., Venter, J.C., Rusch, D.B., and Béjà, O. (2007) Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes. *Environmental Microbiology* **9**: 1464–1475.

Zeng, Y.H., Chen, X.H., and Jiao, N.Z. (2007) Genetic diversity assessment of anoxygenic photosynthetic bacteria by distance-based grouping analysis of *pufM* sequences. *Letters in Applied Microbiology* **45**: 639–645.

Zhou J, Ning D. 2017. Stochastic community assembly: does it matter in microbial ecology? *Microbiol Mol Biol Rev* **81**: e00002-17. https://doi.org/10.1128/MMBR.00002-17.

936


937
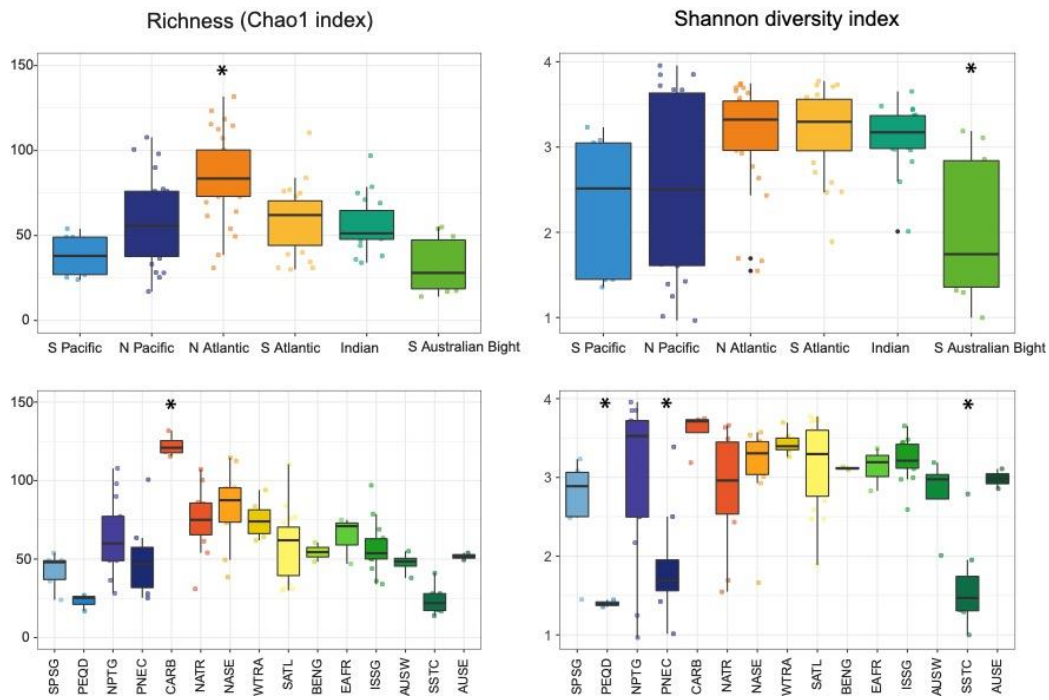
Figure 1. AAP alpha diversity measured as richness (Chao1 index) and Shannon diversity index
within each oceanic region (top panels) and each Longhurst province (bottom panels) sampled
during the Malaspina Expedition. The complete names of the Longhurst provinces are listed in
Table 1 and Fig S1. *Asterisks indicate regions or provinces that are statistically different from
the others, after a post-hoc Tukey test ($p<0.001$).

943
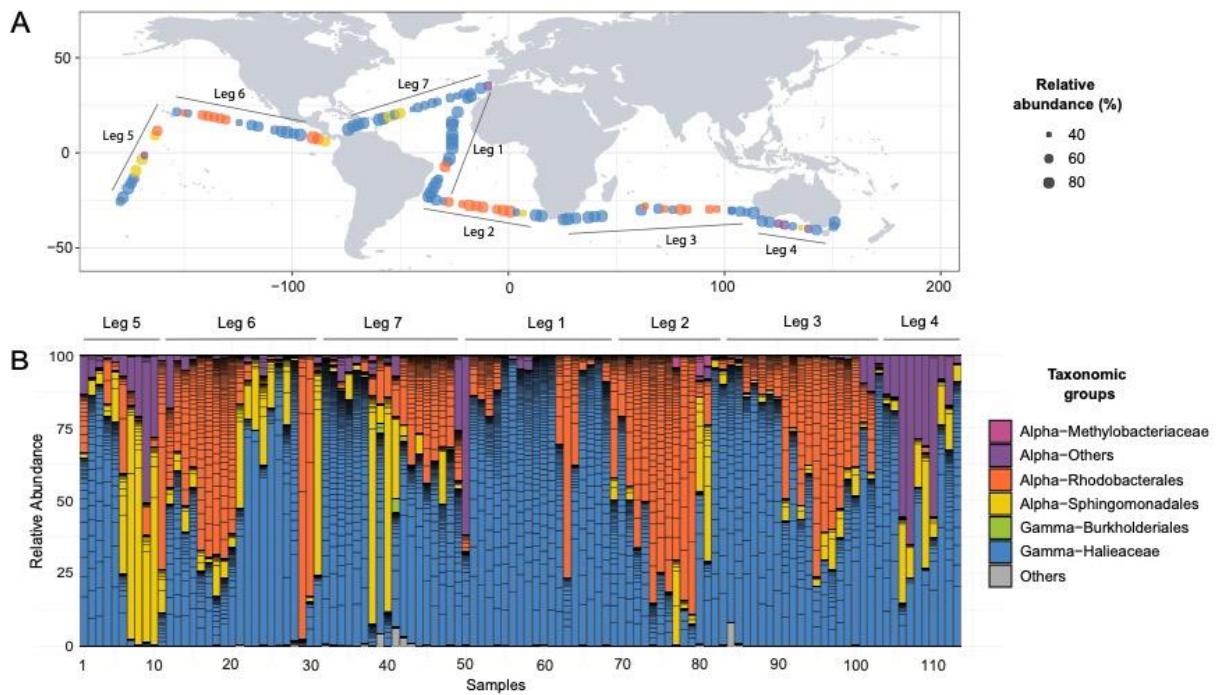
Figure 2. A) Dominant AAP taxonomic groups across the global tropical and subtropical
surface ocean. Each station is colour-coded by the most abundant taxonomic group in the
sample (see taxonomy legend in panel B), and the size of the dot is proportional to the relative

947 abundance of the dominant taxon. The Malaspina Expedition legs are indicated to help visualize
948 the cruise track. B) Community composition at each station, expressed as the relative
949 contribution of each *pufM* sequence colour-coded by its taxonomic affiliation. Samples are
950 ordered following the cruise path as in panel A.
951
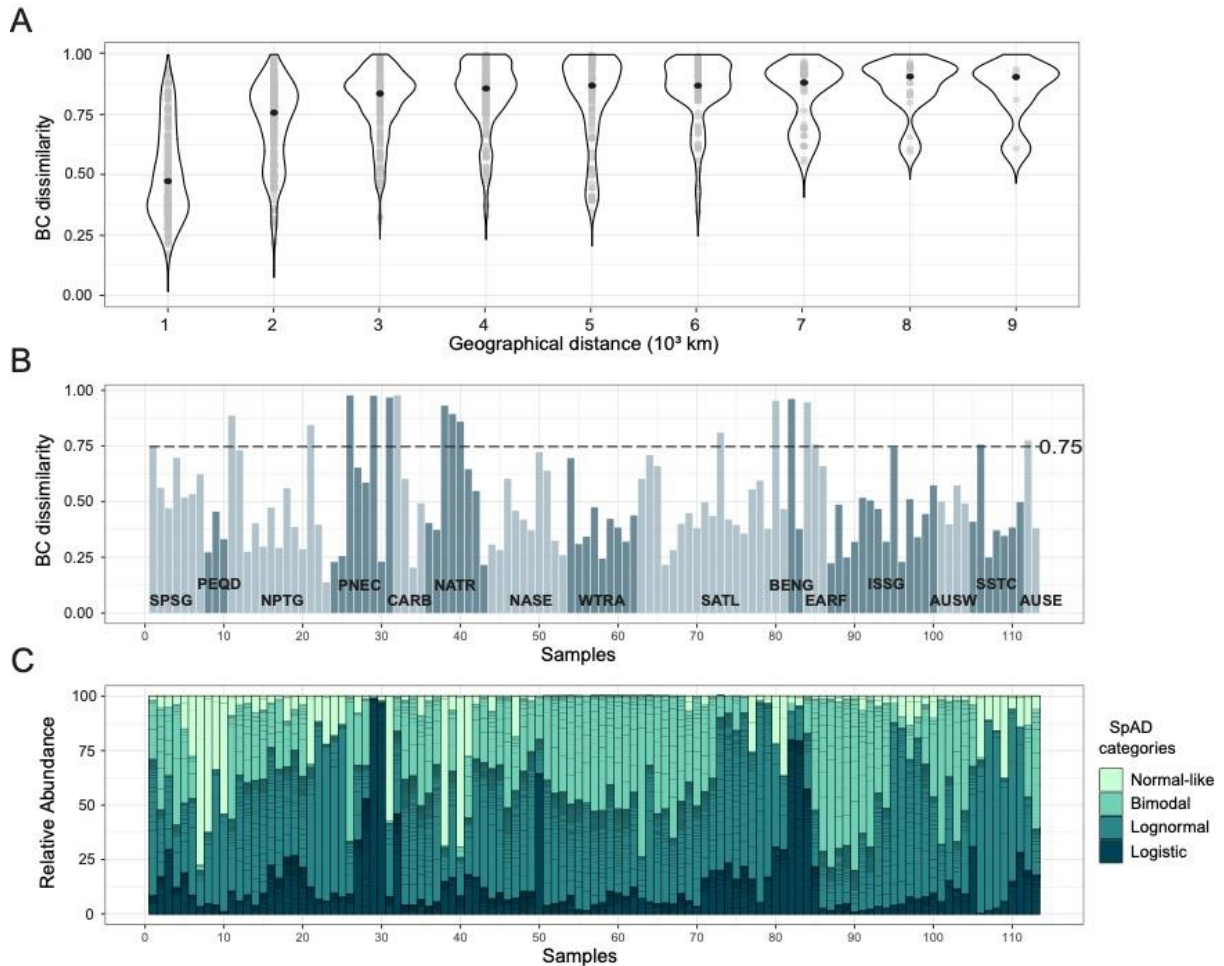
952

953 **Figure 3.** A) Changes in community dissimilarity between AAP assemblages, measured as
954 Bray-Curtis (BC) dissimilarity with regard to the geographical distance among samples. All
955 comparisons are represented by grey dots while black dots indicate the median value of
956 dissimilarity at each distance. We only considered pairwise comparisons between samples
957 located in the same ocean. B) Sequential change in community composition across space
958 (sequential beta diversity). Bars represent BC dissimilarity between each community and the
959 one sampled immediately before (e.g., first bar represents BC dissimilarity between stations 113
960 and 1, second bar represents BC dissimilarity between stations 1 and 2, and so on, up to stations
961 112 and 113). Samples are ordered following the cruise path as in Figure 2 for comparison.
962 Alternating light and dark colour represent a change in Longhurst provinces along the transect
963 and the provinces are indicated according to Longhurst (1998) abbreviations. C) Relative
964 contribution in each community of the four Spatial Abundance Distribution (SpAD) categories
965 of ASVs throughout the surface ocean, displayed using the same sample order as in panel B.
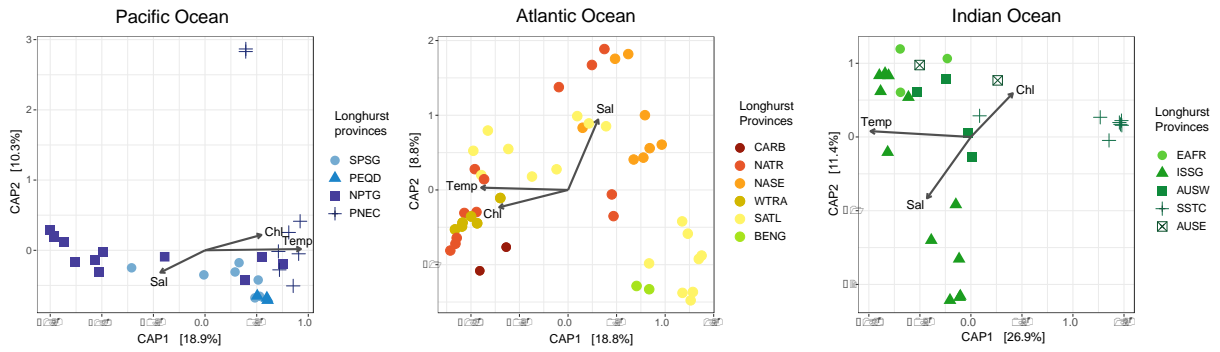
966

967 **Figure 4.** Distance-based redundancy analysis (dbRDA) performed separately for the Pacific,
968 Atlantic and Indian Ocean stations. Samples are color-coded according to the Longhurst
969 provinces to which they belong. Temperature, salinity, and chlorophyll *a* were the three
970 variables that explained the largest fraction of community variance, and they are represented by
971 arrows, where "Temp" is temperature, "Sal" is salinity, and "Chl" is chlorophyll *a*
972 concentration.
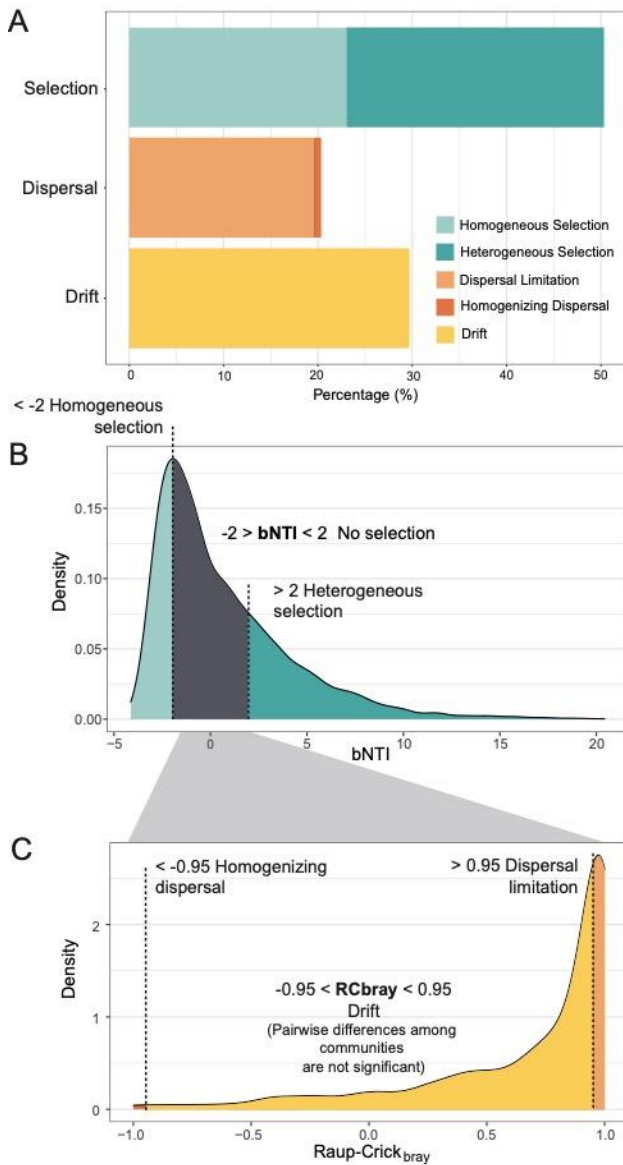
973

974 **Figure 5.** A) Percentage of the AAP bacterial community turnover associated to each ecological
975 process in the tropical and subtropical surface ocean. B) Distribution of βNTI estimates for the
976 total number of comparisons between all samples in the dataset. Absolute values of βNTI above
977 2 are considered as significant departures from random phylogenetic turnover and are associated
978 to homogeneous and heterogeneous selection (blue areas). The grey area represents the fraction
979 of nonsignificant βNTI values. To disentangle whether drift or dispersal are the main ecological
980 processes shaping the turnover between these communities, Bray–Curtis-based Raup–Crick
981 (RC$_{bray}$) was calculated. C) Distribution of RC$_{bray}$ for the pairwise community comparisons that
982 are not structured by selection. RC$_{bray}$ values between −0.95 and +0.95 point to a community
983 assembly governed by drift (yellow area). On the contrary, RC$_{bray}$ values > +0.95 or < −0.95
984 indicate that community turnover is driven by dispersal limitation or homogenizing dispersal
985 respectively (orange areas).

986 **TABLES**
987
988 **Table 1.** Provinces covered during the 2010 Malaspina Expedition and values (average
989 ± standard deviation, minimum to maximum) of temperature, salinity and Chlorophyll *a*
990 concentration measured in each province. Names and abbreviations according to
991 Longhurst (1998). N = number of stations visited in each Longhurst province.
992

Provinces	Province abbreviations	N	Temperature (ºC)	Salinity (PSU)	Chlorophyll *a* (mg·m⁻³)
East Australian Coastal	AUSE	2	21.39 ± 0.35 (21.14 to 21.64)	35.52 ± 0.08 (35.47 to 35.58)	0.34 ± 0.02 (0.32 to 0.36)
Australia-Indonesia Coastal	AUSW	4	23.05 ± 1.41 (21.36 to 24.8)	35.48 ± 0.09 (35.34 to 35.54)	0.13 ± 0.03 (0.1 to 0.16)
Benguela Current Coastal	BENG	2	20.55 ± 0.16 (20.44 to 20.66)	35.52 ± 0.06 (35.48 to 35.56)	0.14 ± 0.11 (0.06 to 0.22)
Caribbean	CARB	4	28.73 ± 0.29 (28.38 to 29.09)	35.6 ± 0.08 (35.54 to 35.71)	0.14 ± 0.04 (0.09 to 0.19)
East Africa Coastal	EAFR	3	23.94 ± 1.82 (22.56 to 26)	35.41 ± 0.12 (35.31 to 35.54)	0.3 ± 0.3 (0.09 to 0.65)
Indian South Subtropical Gyre	ISSG	14	23.57 ± 1.36 (21.74 to 25.92)	35.65 ± 0.25 (35.23 to 36.14)	0.07 ± 0.03 (0.04 to 0.14)
Northeast Atlantic Subtropical Gyre	NASE	10	21.35 ± 1.88 (18.45 to 24.31)	37.03 ± 0.39 (36.43 to 37.65)	0.1 ± 0.07 (0.04 to 0.25)
North Atlantic Tropical Gyre	NATR	11	26.82 ± 1.32 (24.83 to 28.85)	36.68 ± 0.67 (35.53 to 37.57)	0.14 ± 0.1 (0.05 to 0.31)
North Pacific Tropical Gyre	NPTG	13	23.86 ± 1.49 (21.65 to 26.35)	34.66 ± 0.2 (34.2 to 34.94)	0.17 ± 0.09 (0.08 to 0.44)
Pacific Equatorial Divergence	PEQD	3	27.5 ± 0.62 (26.89 to 28.13)	35.21 ± 0.31 (34.85 to 35.39)	0.24 ± 0.05 (0.19 to 0.29)
North Pacific Equatorial Countercurrent	PNEC	8	28.37 ± 0.58 (27.61 to 29.28)	33.84 ± 0.4 (33.15 to 34.28)	0.34 ± 0.11 (0.18 to 0.52)
South Atlantic Gyral province	SATL	19	24.7 ± 2.26 (20.9 to 27.33)	36.49 ± 0.45 (35.79 to 37.25)	0.07 ± 0.03 (0.03 to 0.12)
South Pacific Subtropical Gyre	SPSG	7	28 ± 2 (23.99 to 29.31)	35.04 ± 0.41 (34.43 to 35.59)	0.11 ± 0.05 (0.06 to 0.18)
South Subtropical Convergence Province	SSTC	7	17.29 ± 1.35 (15.75 to 19.55)	35.3 ± 0.2 (34.99 to 35.61)	0.25 ± 0.14 (0.1 to 0.52)
Western Tropical Atlantic Province	WTRA	6	27.6 ± 0.29 (27.27 to 28.05)	35.77 ± 0.38 (35.42 to 36.33)	0.23 ± 0.11 (0.11 to 0.44)

993