1 **Defining marine microbial biomes from environmental and dispersal filtered**
2 **metapopulations**
3
4 Markus V. Lindh[a][*]
5
6 [a]Department of Biology, Lund University, SE-22362 Lund, Sweden
7 *Corresponding author, *E-mail address*: markusvlindh@gmail.com
8
11
12 **Summary**

13 Energy and matter fluxes essential for all life[1] are modulated by spatial and temporal shifts in

14 microbial community structure resulting from environmental and dispersal filtering[2,3],

15 emphasizing the continued need to characterize microbial biogeography[4,5]. Yet, application of

16 metapopulation theory, traditionally used in general ecology for understanding shifts in

17 biogeographical patterns among macroorganisms, has not been tested extensively for defining

18 marine microbial populations filtered by environmental conditions and dispersal limitation at

19 global ocean scales. Here we show, from applying metapopulation theory on two major global

20 ocean datasets[6,7], that microbial populations exhibit core- and satellite distributions with

21 cosmopolitan compared to geographically restricted distributions of populations. We found

22 significant bimodal occupancy-frequency patterns (the different number of species occupying

23 different number of patches) at varying spatial scales, where shifts from bimodal to unimodal

24 patterns indicated environmental and dispersal filtering. Such bimodal occupancy-frequency

25 patterns were validated in Longhurst's classical biogeographical framework[8] and *in silico*

26 where observed bimodal patterns often aligned with specific biomes and provinces described

27 by Longhurst and where found to be non-random in randomized datasets and mock

28 communities. Taken together, our results show that application of metapopulation theory

29 provides a framework for determining distinct microbial biomes maintained by environmental

30 and dispersal filtering.

31 Discovery of biogeographical patterns in marine microbial assemblages[6,7,9-12] caused

32 considerable excitements because it indicated distribution patterns analogous to

33 macroecological patterns[13-17]. These observations have provided clues that marine microbes

34 are, at least to some extent, limited by the environment and their dispersal capacity rather than

35 being cosmopolitan.

36   Metapopulation theory is a key framework in general ecology and incorporates

37 population dispersal to empirically test occupancy-frequency distributions of different

38 organisms ranging from insects to plants[18]. Predictions of occupancy-frequency patterns have

39 been developed in e.g. the core- and satellite hypothesis (CSH) by Ilkka Hanski[19] where the

40 so-called rescue effect supports the survival of metapopulations, and forms a bimodal

41 occupancy-frequency pattern (See Fig. 1A). Previous results applying the CSH to marine

42 ecosystems suggest that this approach could potentially allow for precise definitions of

43 microbial biomes[20].

44   Here we collected data on bacterial and archaeal populations (estimated from

45 16S rRNA gene sequencing and metagenomic data) obtained from two recent major global

46 ocean datasets, Tara oceans[6] and Malaspina[7] (Fig. S1) to examine the shape of occupancy-

47 frequency distributions in global ocean data and aimed to define distinct microbial biomes.

48   For the Tara oceans dataset[6] we focused our analyses on surface seawater (≤ 5

49 m) and analyzed 63 stations in total (Fig. S1). Occupancy-frequency analysis for the whole

50 transect revealed a significant bimodal pattern, with the number of populations demonstrating

51 a monotonic decrease with increasing number of sites occupied followed by an increase in the

52 number of populations occupying all sites (Fig. 1B; Table 1; Mitchell-Old's and Shaw's test,

53 $p < 0.05$). In addition, populations affiliated with Alphaproteobacteria (represented mainly by

54 the SAR11 clade bacteria) also exhibited bimodality for the whole transect (Table 1; Mitchell-

55 Old's and Shaw's test, $p < 0.05$). In contrast, Euryarchaeota, Cyanobacteria,

2

56 Gammaproteobacteria and Bacteroidetes did not exhibit significant bimodal patterns (Table 1;

57 Mitchell-Old's and Shaw's test, p>0.05). Such results indicate that, among individual taxa,

58 except Alphaproteobacteria, surface ocean microbial communities can be limited by the

59 prevailing environmental conditions and their dispersal capacity.

60 For subsets in the whole Tara Oceans transect, here exemplified by the South

61 Pacific and North Atlantic, we found significant bimodal occupancy-frequency patterns for

62 the total community (Table 1; Mitchell-Old's and Shaw's test, p<0.05). We note that

63 Euryarchaeota populations only exhibited a bimodal pattern in the North Atlantic basin (Table

64 1; Mitchell-Old's and Shaw's test, p<0.05). Our results therefore indicate that the total

65 community and most individual taxa are bimodal within a coherent oceanic region.

66 Reciprocally, particular microbial groups are subjected to environmental and dispersal

67 filtering[21] at different spatial scales resulting in geographically constricted populations.

68 The Malaspina dataset[7] contained 30 stations collected from the deep-sea,

69 typically ≥4000 m deep (Fig. S1). This dataset also included size-fractionated samples

70 corresponding to free-living (>0.2 and <0.8 μm) and particle-attached (>0.8 and <20 μm)

71 microbial assemblages. Analysis of occupancy-frequency patterns among deep-sea microbial

72 assemblages revealed a significant bimodal pattern of the total community in the whole

73 transect for both free-living and particle-attached populations (Fig. 1C; Table 1). Thus,

74 although most deep-sea microbes have been suggested to be confined to a specific region[7],

75 our results indicate that a core deep-sea community are still distributed across the whole

76 Malaspina transect, suggesting a similar microbial biome without environmental or dispersal

77 filtering.

78 Variations in the shape of occupancy-frequency distributions in the Malaspina

79 transect were found for different size fractions, taxa, oceanic basins and taxa within different

80 basins (Table 1). For example, Thaumarchaeota populations displayed bimodality in the

3

81 South Atlantic in the particle-attached community (Table 1; Mitchell-Old's and Shaw's test,

82 p<0.05) but not in the free-living community (Table 1; Mitchell-Old's and Shaw's test,

83 p>0.05). In the South Atlantic, Gammaproteobacteria populations exhibited a bimodal pattern

84 in both the particle-attached and free-living community (Table 1; Mitchell-Old's and Shaw's

85 test, p<0.05). We note that samples collected within the same water mass, here exemplified

86 by "WMbSC 2: NADW−CDW−AABW" and "WMbSC 4: Purest AABW Ross", often

87 showed a bimodal occupancy-frequency pattern (Table 1; Mitchell-Old's and Shaw's test,

88 p<0.05). Overall these findings substantiate that particular taxa such as Thaumarchaeota can

89 be limited by environmental conditions and dispersal capacity and thus geographically

90 restricted.

91 We further aimed to validate our observed microbial biomes estimated from

92 bimodal occupancy-frequency patterns in the framework of Longhurst's ecological biomes

93 and provinces that define biogeographic distributions derived in large part from satellite-

94 based estimations of sea-surface chlorophyll *a* (Chl *a*) concentrations[8]. Longhurst defines four

95 primary oceanic biomes; the Westerlies, Trades, Polar and Coastal (Fig. S1). These biomes

96 differ in nutrient supply, light and seasonal variation in water column mixing. The biomes are

97 in turn subdivided into several provinces based on Chl *a* distributions (Fig. S1). Our analysis

98 revealed that samples collected as part of the Tara Oceans transect within the Coastal,

99 Westerlies and Trades biomes displayed a distinct bimodal occupancy-frequency pattern

100 (Table 1; Mitchell-Old's and Shaw's test, p<0.05). In addition, provinces such as the Indian

101 Monsoon Gyres Province (MONS) and South Atlantic Gyre Province (SATL), exhibited

102 significant bimodal patterns (Table 1; Mitchell-Old's and Shaw's test, p<0.05).

103 During the Malaspina transect examining the deep-sea microbes, only the

104 particle-attached microbial community exhibited bimodal occupancy-frequency in the SATL

105 (Table 1; Mitchell-Old's and Shaw's test, p<0.05). Gammaproteobacteria displayed

4

106   bimodality in the SATL both within the free-living and particle-attached community (Table 1;

107   Mitchell-Old's and Shaw's test, p<0.05). However, only particle-attached Alpha- and

108   Actinobacteria displayed bimodality in SATL (Table 1; Mitchell-Old's and Shaw's test,

109   p<0.05). Collectively, these findings emphasize that occupancy-frequency patterns among

110   microbial assemblages can likely be used to validate and significantly extend analyses of

111   oceanic divisions as determined by e.g. Chl *a* satellite data[8].

112   The difficulty in assessing and discerning sampling effects from true biological

113   effects for understanding metapopulation dynamics is well recognized in terrestrial studies of

114   macroorganisms[18]. Here we aimed to (i) examine the effect of sequencing depth on the

115   cumulative number of core and satellite populations, and (ii) elucidate if bimodal patterns

116   could arise from pure chance and hence be an artefactual effect in marine microbial

117   assemblages. The number of core and satellite populations in subsampled rarefied datasets

118   reached saturation around 25,000 sequence reads (Fig. S2). *In silico* tests with randomized

119   datasets and mock communities confirmed that bimodal patterns were non-random, as no

120   significant bimodal pattern was found in any randomized dataset or mock communities (Fig

121   2A-B; Mitchell-Old's and Shaw's test, p>0.05). Our extensive *in silico* analyses provided a

122   first clue that bimodality could be a biological pattern rather than an artefactual effect.

123   To test the possibility of allowing the metapopulation framework to pinpoint

124   microbial biomes without testing against a pre-defined oceanic region we analyzed 5 patches

125   at a time in sequence along the Tara oceans transect (Fig. 2C-D). This exercise highlighted

126   oceanic regions where the core populations (bimodal occupancy-frequency patterns; Mitchell-

127   Old's and Shaw's test, p<0.05) prevailed and the community was not limited by the

128   environment or dispersal capacity. Reciprocally, we could also define specific microbial

129   biomes along the transect (non-bimodal occupancy-frequency patterns; Mitchell-Old's and

130   Shaw's test, p>0.05). It was noteworthy that for all taxa combined, stations obtained in the

131   Southern Ocean limited the distribution of microbial assemblages (Fig. 2C).

132   For individual taxa in this unsupervised model, the occupancy-frequency pattern

133   varied substantially and different major microbial groups were subjected to different

134   environmental and dispersal filters. Euryarchaeota only had significant bimodal patterns in

135   patches sampled from the North Atlantic, Red Sea, Arabian Sea and the Indian Ocean.

136   Cyanobacteria exhibited bimodality in the Pacific Ocean but the pattern broke down upon

137   entry into the Atlantic Ocean (Fig. 2D). Notably, within Cyanobacteria, *Synechococcus* had a

138   geographical restriction between the North and South Pacific Ocean (see insert Fig. 2D).

139   *Synechococcus* were also limited in their distributions in the Indian Ocean by a single site.

140   Alphaproteobacteria had a similar metapopulation distribution limitation as the total

141   community and Cyanobacteria in the Southern Ocean. Yet, within Alphaproteobacteria,

142   SAR11 clade bacteria (see insert Fig. 2D) had a wider distribution, exemplified by a bimodal

143   pattern between sites obtained in the South Atlantic and Southern Ocean. Still, SAR11 were

144   limited upon entry into the Pacific Ocean. Gammaproteobacteria and Bacteroidetes displayed

145   a clear distinction between samples obtained in the Indian Ocean and South Atlantic where

146   the core community was constrained by the station outside Cape Town.

147   Taken together, we demonstrate that empirical tests of metapopulation dynamics

148   allows for biogeographical analyses of marine microbes to define microbial biomes. We note

149   that a sequence depth of 25,000 sequence reads is sufficient to capture most of the core- and

150   satellite populations and could be considered as lower limit for microbial biogeography

151   analyses at large spatial scales[23]. Thus, variations in microbial biogeography, in particular,

152   deviations from microbial biomes as defined by the CSH, can potentially be used in

153   monitoring environmental changes, and might therefore be valuable tools in marine

154   management.

155

156 **Methods**

157   Occupancy-frequency distributions were analyzed as described in[20]. In brief, an

158 equivalent to Tokeshi's test of bimodality was performed using Mitchell-Olds' and Shaw's

159 test[22] for the location of quadratic extremes. The global ocean transect data were downloaded

160 from the European Nucleotide Archive (ENA; accession number PRJEB7988) and the NCBI

161 Sequence Read Archive (SRA; accession number SRP031469), Tara Oceans and Malaspina

162 transects, respectively. For *in silico* tests we first tested the effect of sequencing depth and

163 subsampled the Tara oceans dataset to 100,000 sequence reads and rarefied this subsampled

164 dataset to 1000, 2000, 5000, 10,000, 20,000, 40,000, 60,000 and 80,000 sequence reads and

165 plotted the cumulative number of satellite compared to core populations. We further created

166 one randomized dataset by shuffling the presence/absence of OTUs using the "permatfull"

167 function in R by keeping the sample sums (100,000 sequence reads) static and performing

168 999 unrestricted permutations of the OTUs from the subsampled Tara oceans dataset and

169 picked one random permuted community as mock community and rarefied as above. Thirdly,

170 for the occupancy-frequency analyses we subsampled the Tara oceans and Malaspina dataset

171 to 40,590 and 25,000 sequence reads, respectively, well above the suggested lower limit for

172 diversity analyses of marine microbial assemblages[23] and the 25,000 sequence reads noted in

173 the test above (Fig. S2). Finally, we randomized each of the subsampled Tara oceans and

174 Malaspina dataset and specific oceanic regions (North Atlantic, South Pacific and South

175 Atlantic, Brazil basin, Tara oceans and Malaspina datasets respectively) and we used one

176 randomized dataset as a mock community and this artificial community were randomized as

177 described above. All statistical tests were performed in R 3.3.3[24], and using the package

178 "Vegan"[25].

References

179

180   1   Falkowski, P. G., Barber, R. T. & Smetacek, V. Biogeochemical Controls and

181   Feedbacks on Ocean Primary Production. *Science* **281**, 200-206 (1998).

182   2   Kirchman, D. L., Dittel, A. I., Malmstrom, R. R. & Cottrell, M. T.

183   Biogeography of major bacterial groups in the Delaware Estuary. *Limn*

184   *Oceanogr* **50**, 1697-1706 (2005).

185   3   Fuhrman, J. A. et al. Annually reoccurring bacterial communities are predictable

186   from ocean conditions. *Proc Natl Acad Sci USA* **103**, 13104-13109 (2006).

187   4   Martiny, J. B. H. et al. Microbial biogeography: putting microorganisms on the

188   map. *Nat Rev Microbiol* **4**, 102-112 (2006).

189   5   Hanson, C. A., Fuhrman, J. A., Horner-Devine, M. C. & Martiny, J. B. Beyond

190   biogeographic patterns: processes shaping the microbial landscape. *Nat Rev*

191   *Microbiol* **10**, 497-506 (2012).

192   6   Sunagawa, S. et al. Structure and function of the global ocean microbiome.

193   *Science* **348** (2015).

194   7   Salazar, G. et al. Global diversity and biogeography of deep-sea pelagic

195   prokaryotes. *ISME J* **10** (2016).

196   8   Longhurst, A. R. Ecological Geography of the Sea (Second Edition) 1-17

197   (Academic Press, 2007).

198   9   Finlay, B. J. Global Dispersal of Free-Living Microbial Eukaryote Species.

199   *Science* **296**, 1061-1063 (2002).

200   10   Fenchel, T. Biogeography for Bacteria. *Science* **301**, 925-926 (2003).

201   11   Pommier, T. et al. Global patterns of diversity and community structure in

202   marine bacterioplankton. *Mol Ecol* **16**, 867-880 (2007).

203 12 Barberan, A. & Casamayor, E. O. Global phylogenetic community structure and

204   beta-diversity patterns in surface bacterioplankton metacommunities. *Aquat*

205   *Microb Ecol* **59**, 1-10 (2010).

206 13 Horner-Devine, M. C. & Bohannan, B. J. M. Phylogenetic clustering and

207   overdispersion in bacterial communities *Ecology* **87**, S100-S108 (2006).

208 14 Ghiglione, J. F. et al. Pole-to-pole biogeography of surface and deep marine

209   bacterial communities. *Proc Natl Acad Sci USA* **109**, 17633-17638 (2012).

210 15 Amend, A. S. et al. Macroecological patterns of marine bacteria on a global

211   scale. *J. Biogeogr*. **40**, 800-811 (2013).

212 16 Sul, W. J., Oliver, T. A., Ducklow, H. W., Amaral-Zettler, L. A. & Sogin, M. L.

213   Marine bacteria exhibit a bipolar distribution. *Proc Natl Acad Sci USA* **110**,

214   2342-2347 (2013).

215 17 Zinger, L., Boetius, A. & Ramette, A. Bacterial taxa–area and distance–decay

216   relationships in marine environments. *Mol Ecol* **23**, 954-964 (2014).

217 18 McGeoch, M. A. & Gaston, K. J. Occupancy frequency distributions: patterns,

218   artefacts and mechanisms. *Biol Rev* **77**, 311-331 (2002).

219 19 Hanski, I. Dynamics of regional distribution - the core and satellite species

220   hypothesis. *Oikos* **38**, 210-221 (1982).

221 20 Lindh, M. V. et al. Metapopulation theory identifies biogeographical patterns

222   among core and satellite marine bacteria scaling from tens to thousands of

223   kilometers. *Environ Microbiol*, **19**, 1222–1236 (2016).

224 21 Mehranvar, L. & Jackson, D. A. History and taxonomy: their roles in the core-

225   satellite hypothesis. *Oecologia* **127**, 131-142 (2001).

226 22   Mitchell-Olds, T. & Shaw, R. G. Regression Analysis of Natural Selection:

227   Statistical Inference and Biological Interpretation. *Evolution* **41**, 1149-1161

228   (1987).

229 23   Lundin, D. et al. Which sequencing depth is sufficient to describe patterns in

230   bacterial alpha- and beta-diversity? *Env Microbiol Rep* **4**, 367-372 (2012).

231 24   R Core Development Team. A Language and Environment for Statistical

232   Computing. *https://cran.r-project.org/* (2014).

233 25   Oksanen, J. et al. vegan: Community Ecology Package. R package version 1.17-

234   5. *https://cran.r-project.org/web/packages/vegan/index.html* (2010).

235

236   **Figure and Table legends**

237   **Figure 1.** Conceptual drawing of the core satellite hypothesis (CSH; A), and Occupancy-

238   frequency distributions (the different number of populations occupying different number of

239   sites) of populations estimated from 16S rRNA gene amplicon and metagenomic data

240   obtained from the Tara oceans[6] (B), and Malaspina[7] (C) datasets. (A) was modified from[20]; *P*

241   is the fraction of occupied sites, *C* is colonization rate and *E* is extinction rate. The quadratic

242   colonization and extinction rates are calculated according to $dP/dt = CP(1 - P) - EP(1 - P)$[19].

243   The CSH predicts a bimodal pattern, and incorporates positive feedback mechanisms between

244   local abundance and regional occupancy[19]. High rates of colonization in the CSH protect a

245   population from extinction and are known as the rescue effect.

246

247   **Figure 2.** *In silico* tests of randomized sequence data and microbial biomes defined by shifts

248   from bimodal occupancy-frequency patterns to unimodal patterns. Randomizations of

249   observational data were performed for complete datasets and specific oceanic regions (A),

250   and in mock communities (B). The typical occupancy-frequency pattern of the randomized

251   datasets was characterized by most populations being detected at 2-4 sites with a monotonical

252   decrease in the number of populations occupying increasing number of sites. No randomized

253   community dataset exhibited a significant bimodal pattern. The randomized mock community

254   typically displayed prevalence of most populations occupying 2-4 sites but with no significant

255   bimodal pattern found. Microbial biomes (C-D) were defined by shifts from bimodal

256   occupancy-frequency patterns to unimodal patterns by testing the occupancy-frequency

257   distributions with combinations of sites within the Tara oceans transect. Red dashed lines

258   denote environmental and dispersal filtering of microbial assemblages resulting

259   geographically constricted populations. Color denotes significance level of Mitchell-Old's

260   and Shaw's test for bimodality.

11

261   **Supporting information**

262

263   **Supplementary figures and tables**

264

265   **Figure S1.** Schematic map of sampled stations in the Tara oceans[6] and Malaspina[7] datasets

266   included in the analyses and superimposed with Longhurst's biogeographical division of the

267   ocean into biomes and provinces[8]. Blue, Brown, Green and Yellow color denote Polar,

268   Coastal, Westerlies and Trades biomes, respectively.

269

270   **Figure S2.** *In silico* tests of the effects of sequencing depth on the cumulative number of core

271   compared to satellite populations for the rarefied total community obtained from Tara Oceans

272   (A), and one mock community (B) obtained from randomizing the Tara Oceans dataset and

273   rarefying as in (A).

274

Table 1. Prevalence of bimodal compared to unimodal occupancy-frequency patterns within the Tara Oceans and Malaspina transects for different oceanic basins and within Longhurst's biomes and provinces[8]. Asterisks denote level of significance (*** <0.001, ** <0.01, *<0.05).

		Taxa					
Dataset	**Region**	**ALL OTUs (n=35651)**	**Euryarchaeota (n=550)**	**Cyanobacteria (n=963)**	**Alphaproteobacteria (n=7151)**	**Gammaproteobacteria (n=7766)**	**Bacteroidetes (n=3380)**
	Full transect	BIMODAL**	OTHER	OTHER	BIMODAL***	OTHER	OTHER
	South Pacific	BIMODAL***	OTHER	BIMODAL*	BIMODAL***	BIMODAL***	BIMODAL***
	North Atlantic	BIMODAL***	BIMODAL***	BIMODAL***	BIMODAL***	BIMODAL**	BIMODAL**
	Coastal Biome (L)	BIMODAL*	OTHER	OTHER	BIMODAL***	BIMODAL*	OTHER
	Trades Biome (L)	BIMODAL***	OTHER	BIMODAL*	BIMODAL***	BIMODAL***	BIMODAL***
Tara Oceans	Westerlies (L)	BIMODAL***	OTHER	BIMODAL*	BIMODAL***	BIMODAL***	BIMODAL***
	MEDI (L)	BIMODAL***	UNIMODAL	BIMODAL*	BIMODAL***	BIMODAL***	BIMODAL***
	EAFR (L)	BIMODAL***	BIMODAL***	BIMODAL***	BIMODAL***	BIMODAL**	BIMODAL***
	SATL (L)	BIMODAL**	UNIMODAL	BIMODAL***	BIMODAL***	UNIMODAL	BIMODAL*
	SPSG (L)	BIMODAL***	OTHER	BIMODAL*	BIMODAL***	BIMODAL***	BIMODAL***
	NASW (L)	BIMODAL***	BIMODAL***	BIMODAL***	BIMODAL***	BIMODAL*	BIMODAL*

Table 1 continued.

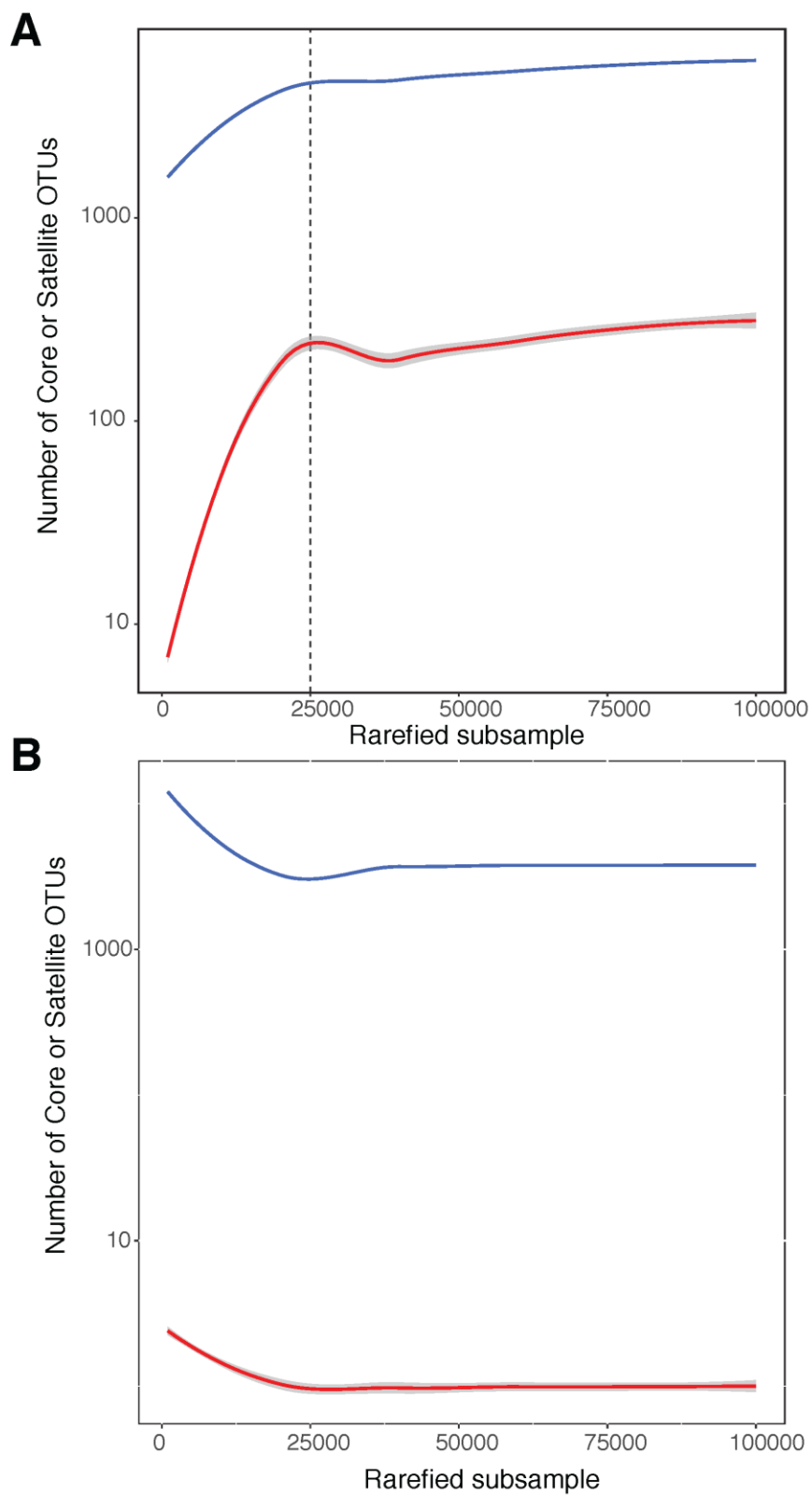|		Taxa					
Dataset	**Region**	**ALL OTUs (n=3902)**	**Thaumarchaeota (n=48)**	**Gammaproteobacteria (n=384)**	**Alphaproteobacteria (n=405)**	**Actinobacteria (n=143)**	**Deltaproteobacteria (n=433)**
	Full transect	BIMODAL*/ BIMODAL	OTHER/OTHER	OTHER/OTHER	OTHER/BIMODAL***	OTHER/OTHER	OTHER/OTHER
Malaspina	South Atlantic	BIMODAL**/BIMODAL*	OTHER/BIMODAL**	BIMODAL**/BIMODAL**	OTHER/BIMODAL**	UNIMODAL/UNIMODAL	BIMODAL*/UNIMODAL
0.2/0.8 µm	Indian Ocean	ND/BIMODAL*	ND/OTHER	ND/BIMODAL**	ND/BIMODAL**	ND/BIMODAL***	ND/UNIMODAL
	WMbSC 2: NADW−CDW−AABW	BIMODAL*/UNIMODAL	BIMODAL***/ND	BIMODAL**/ND	BIMODAL**/ND	OTHER/UNIMODAL	BIMODAL*/UNIMODAL
	WMbSC 4: Purest AABW Ross	ND/BIMODAL*	ND/BIMODAL***	ND/BIMODAL***	ND/BIMODAL***	ND/BIMODAL*	UNIMODAL/BIMODAL*
	SATL (L)	UNIMODAL/BIMODAL*	OTHER/OTHER	BIMODAL***/BIMODAL*	OTHER/BIMODAL***	OTHER/BIMODAL*	UNIMODAL/UNIMODAL
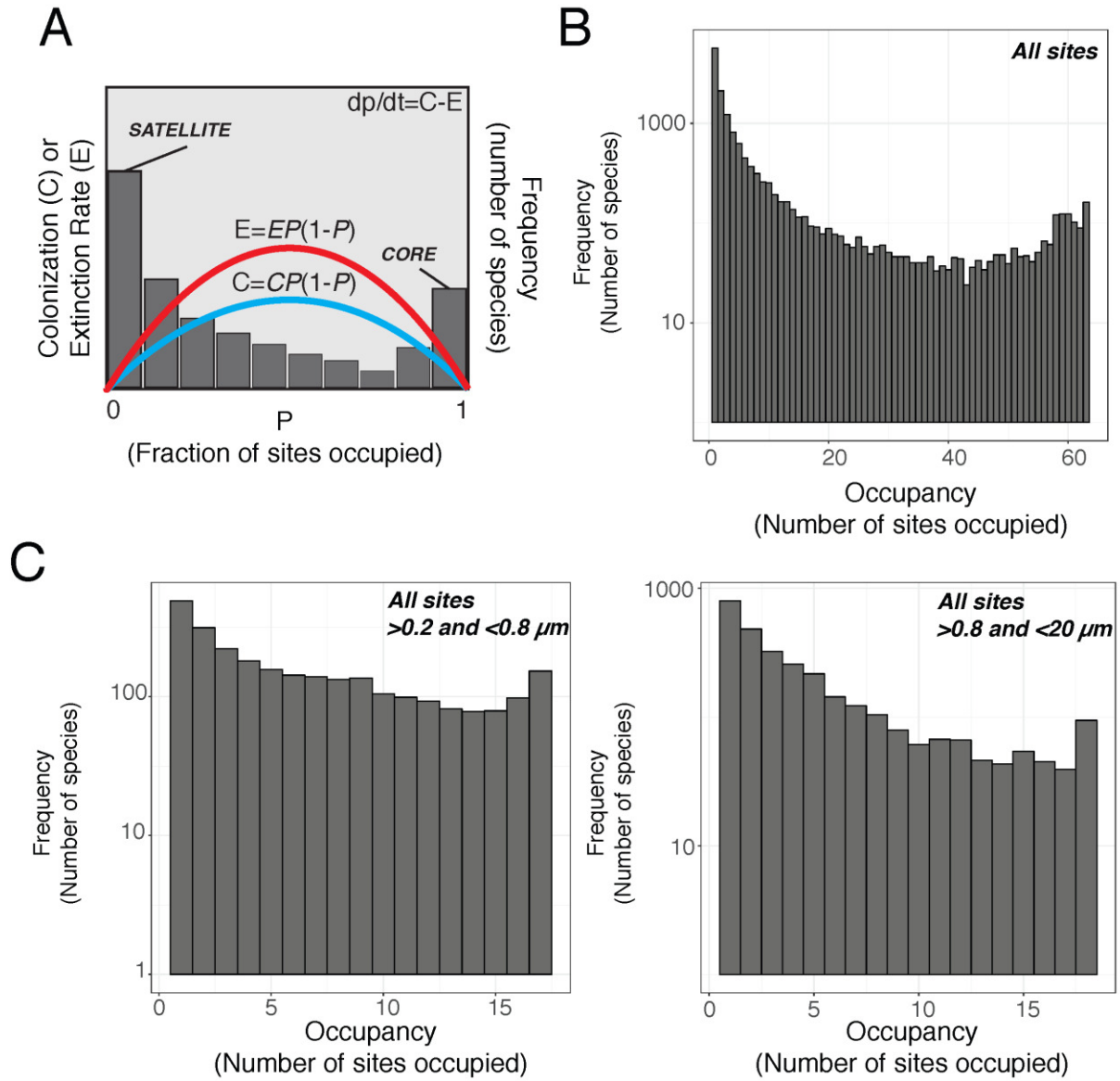
Figure 1
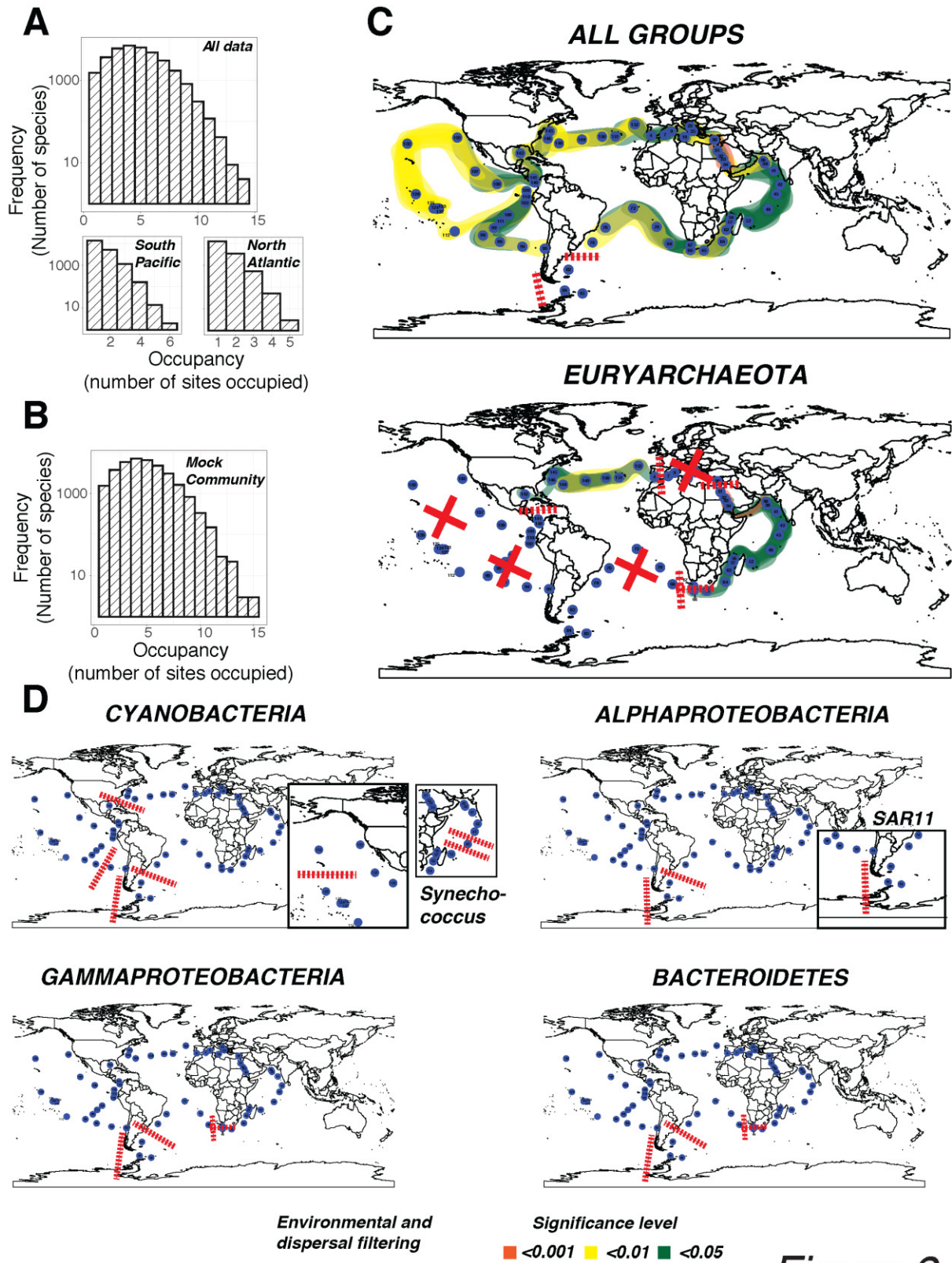
Figure 2
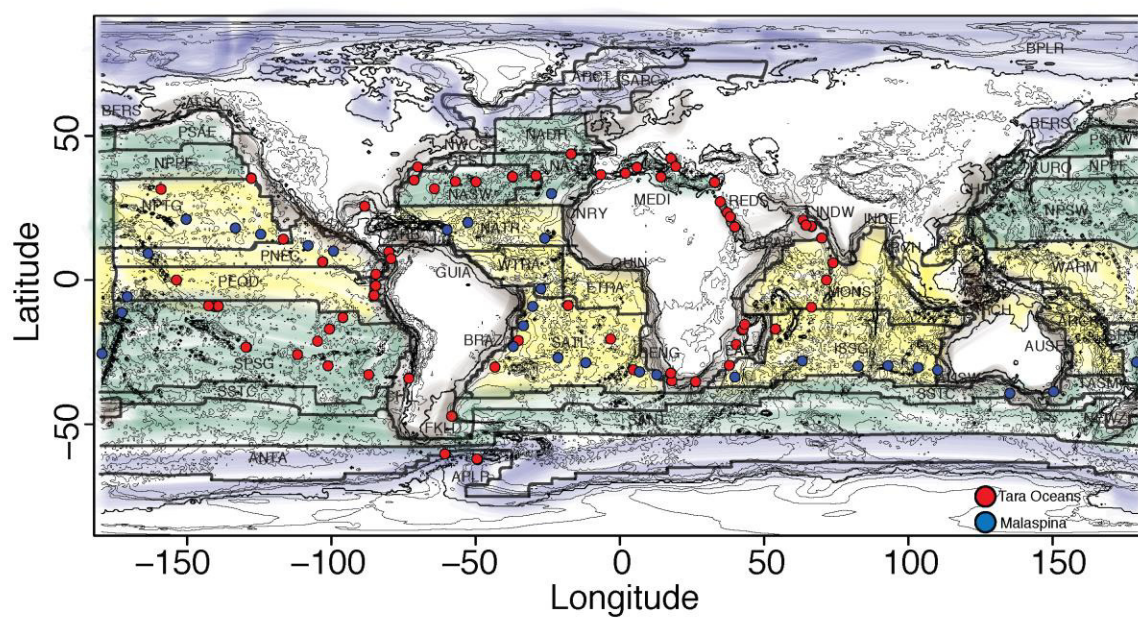
Supplementary Figure 1

Supplementary Figure 2