

Meta-analysis of 139 extant *Tara* ocean metagenomes to unveil the relationship between taxonomy and functionality in prokaryotes inhabiting aquatic ecosystems

Robert Starke

*The total microbiome functionality of bacteria was recently predicted to be 35.5 ±0.2 million of KEGG functions. Logically, due to the limitation in space and resource availability of the local community, local functionality will only comprise a small subset of the total functionality but the relationship between taxonomy and functionality is still uncertain. Here, I used a meta-analysis of 139 extant *Tara* ocean seawater samples from 68 locations across to globe with information on prokaryotic taxonomy on species level from 16S metabarcoding and functionality of prokaryotes on eggNOG gene family level from metagenomes to unveil the relationship between taxonomy and functionality, and to predict the global distribution of functionality. Functional richness showed a statistically significant increase with increasing species richness ($P < 0.0001$, $R^2 = 0.64$) and increasing species diversity ($P < 0.0001$, $R^2 = 0.26$) while functional diversity was similar across the different waters, ranging from 2.96 to 3.22. Globally, the highest functional richness was found in the Northern Pacific Ocean and in the North Atlantic Ocean, and decreased at extreme latitudes. Taken together, I unveil the relationship between taxonomy and functionality, and predict the global distribution of functional richness in prokaryotes inhabiting aquatic ecosystems, implying more pronounced effects in terrestrial ecosystems due to larger differences in environmental parameters especially for functional diversity.*

1 Ecosystem functioning is mediated by biochemical transformations performed by a
2 community of microbes from every domain of life ¹. Prokaryotes play key roles in
3 biogeochemical processes such as carbon and nutrient cycling ² and provide the basis for the
4 genetic diversity due to their biomass with 10⁴ to 10⁶ cells per milliliter combined with high
5 turnover rates and environmental complexity ³. The visible result of genetic diversity are
6 functions, which can be statistically inferred based upon homology to experimentally
7 characterized genes and proteins in specific organisms to find orthologs in other organisms
8 present in a given microbiome. This so-called ortholog annotation, among others, can be
9 performed in eggNOG ⁴ that comprises 721,801 orthologous groups encompassing a total of

10 4,396,591 genes and covers all three domains of life (more information about the database can
11 be obtained under <http://eggnogdb.embl.de/#/app/home>). However, the bottleneck of
12 describing microbiome functions is the low number of fully sequenced and annotated genomes
13 as they are mostly limited to those that have undergone isolation and extensive
14 characterization. Problematically, the vast majority of organisms were not yet studied^{5,6} and
15 the annotation is based on the similarity to the genomes of the very few studied model
16 organisms. Recently, the total functionality in bacteria were estimated to be 35.5 ± 0.2 million
17 functions⁷ but the relationship between taxonomy and functionality at the local scale and the
18 global distribution of functionality is still uncertain. Here, I used a meta-analysis of 139 extant
19 *Tara* ocean seawater samples using 16S metabarcoding for the taxonomic profile of bacteria
20 combined with metagenome sequencing and eggNOG affiliation for the functional profile of
21 prokaryotes. I aimed to estimate the number of prokaryotic microbiome functions and its
22 Shannon diversity in 20L seawater by identifying the model that best fitted their relationship to
23 species richness and species diversity, and to predict the global distribution of functional
24 richness and functional diversity. I hypothesize that (i) both richness and diversity of local
25 functionality will increase with increasing richness and diversity of prokaryotic species due to
26 the addition of rare functions and (ii) that the functional diversity is similar across different
27 seawater ecosystems as the environments are similar.

28 In the 139 *Tara* ocean seawater samples enriched in prokaryotes, functionality ranged
29 from 12,328 eggNOG gene families in the Southern Oceans (-61.969° latitude & -49.502°
30 longitude) to 25,238 in the South Pacific Ocean (-8.973° latitude & -139.239° longitude) with an
31 average of $19,523 \pm 2,682$ functions. The relationship between taxonomy and functionality
32 showed statistically significant (P-value <0.05) correlations but the coefficient of determination
33 depended on the specific comparison (**Figure 1 & Table 1**). The linear relationship of increasing
34 functional richness with increasing taxonomic richness showed the statistically best correlation
35 with a low P-value in combination with a high coefficient of determination (**Figure 1a**),
36 consistent with my first hypothesis. The addition of new species is likely to add new rare
37 functions⁷ to the total functional richness which is why an increasing number of species will
38 result in an increasing number of functions. However, this number is limited by space and

39 resource availability of the surrounding environment and its inhabiting microbial community. A
40 maximum of 25,238 functions were carried by 6,254 species but it is likely to assume that
41 seawater samples carry around the average at $19,523 \pm 2,682$ functions in $4,034 \pm 992$ species.
42 Otherwise, the nature of the correlations between taxonomic richness and functional diversity
43 (**Figure 1b**), taxonomic diversity and functional richness (**Figure 1c**) and taxonomic diversity and
44 functional diversity (**Figure 1d**) were all quadratic, implying a local minimum or maximum for
45 each function. Indeed, functional diversity showed a maxima at 3.12 ± 0.01 (with 3.11 and 3.13
46 as 2.5% confidence intervals) with a richness of 5,809 species but a minimum at a functional
47 diversity of 3.08 ± 0.01 (3.07-3.09) with a species diversity of 6.4. Functional richness showed a
48 minimum at $17,441 \pm 446$ (16,568-18,317) functions with a species diversity of 6.1. Noteworthy,
49 the increase in functional richness with decreasing species diversity is driven by three samples
50 from the Indian Ocean and their exclusion results in a statistically significant linear and positive
51 relationship ($P < 0.0001$, $R^2 = 0.27$) which is why I would argue increasing species diversity is
52 increasing functional richness similarly to species richness. Otherwise, functional diversity
53 showed opposing trends for species richness (local maximum) and species diversity (local
54 minimum). Again, the relationship of species diversity is driven by the three samples from the
55 Indian Ocean but also two samples from the Southern Ocean, making it more likely to be a
56 reasonable trend as it was found in different waters across the globe. However, functional
57 diversity ranges only from 2.96 to 3.22 with an average of 3.09 ± 0.05 across the 139 seawater
58 samples from different locations where functional richness ranged from 12,328 to 25,238
59 functions. In comparison, species diversity ranges from 2.48 to 6.97 with an average of 4.03
60 ± 0.99 and a species richness from 2,484 to 6,974. A three-fold larger range in functional
61 richness but a magnitude smaller range in functional diversity suggests, in my opinion, that
62 functional diversity is similar or at least comparable in all the different waters, in line with my
63 second hypothesis. The highest functional diversity reflects both a fit and a healthy community
64 that is able to perform a wide spectrum of possible transformations given by the space and the
65 resource availability of the surrounding environment without overproportioned abundance of
66 singular functions, which would cause a decrease in functional diversity - as seen in the
67 taxonomic data. Environmentally, similar functional diversity across the different waters makes

68 sense as similar processes are performed and the environmental variables such as temperature
69 ⁸, salinity ⁹, oxygen availability ¹⁰ and dissolved inorganic nutrients ¹¹ are similar among the
70 sampled regions. Otherwise, samples from more diverse regions such as the Arctic Ocean or
71 terrestrial ecosystems with a wider range of values of different environmental variables will
72 cause more pronounced differences in functional diversity.

73 Globally, functional richness was highest in Northern Pacific Ocean near the American
74 coast and in the North Atlantic Ocean, consistent with statistically significant (P-value <0.05)
75 higher averages of these waters compared to the regions with low functional richness such as
76 the Indian Ocean, the Mediterranean Sea, the Red Sea and the Southern Ocean (**Figure 2**).
77 Overall, the model comprising second-degree polynomial terms increased in significance
78 (adjusted $R^2 = 0.34$, P-value = $1.105e^{-6}$) when environmental variables were considered (adjusted
79 $R^2 = 0.64$, P-value = $7.132e^{-8}$) but nitrate concentration showed the most significant individual
80 effect among the tested environmental variables revealed by the lowest AIC (**Table 2**) and the
81 highest increase in significance (adjusted $R^2 = 0.56$, P-value = $1.059e^{-9}$). The correlation between
82 nitrate concentration and functional richness was significant and positively linear (Adjusted R^2
83 = 0.15 , P-value = $1.342e^{-5}$), similar to the significant and positive first-degree polynomial
84 contribution to the best fitting model (P-value = 0.000319) to infer high functional richness with
85 high nitrate concentrations. An increase in functionality with changing conditions from aerobic
86 near the surface and anaerobic with increasing depths aligns well with an increasing number of
87 transformation processes and related enzymes involved in microbial respiration. On the one
88 hand, aerobic breathing comprises only one reaction that oxidizes a carbon source to water and
89 carbon dioxide performed by mono- and dioxygenases. Otherwise, the marine nitrogen cycle
90 includes nitrogen fixation by bacteria, nitrate reduction of ammonia production/reduction by
91 phytoplankton in the euphotic zone, followed by sinking/mixing of ammonia and its nitrification
92 to nitrate in the 'dark ocean' from where denitrification to nitrogen or vertical mixing with the
93 euphotic zone takes place ¹². To my surprise as it is contrary to the positive relationship
94 between nitrate concentration and functional richness, low functional richness was found in
95 extreme negative latitudes even though the nitrate concentrations were reported to be highest
96 in these regions with sea-surface concentrations up to $30 \text{ mmol N per m}^3$ near Antarctica ^{11,13}.

97 However, none of these regions of presumably low functionality have actually been sampled
98 and the lower predictions are likely due to the nature of second-degree polynomial functions
99 that forced the maximum of functionality where the samples were taken and result in a
100 minimum towards the extremes. In favor of the lower functional richness near Antarctica are
101 three samples from the Southern Ocean, which align well with the prediction. Admittedly,
102 despite the high coefficient of determination and the significant P-value of the model, most
103 sampling points do not show a close match to the local prediction of functional richness which
104 is why more samples are necessary for a more precise prediction of functionality, especially in
105 the less sampled regions with low functionality such as the Arctic Ocean.

106 Altogether, I quantify the relationship between taxonomy and functionality in
107 prokaryotes inhabiting different waters locally and predicted the global distribution of
108 functional richness as functional diversity showed only marginal differences. Noteworthy, the
109 coverage of aquatic ecosystems of the data was admittedly low despite the massive effort of
110 sampling eight oceans over three years but the sampling of more oceans will be beneficial for
111 further predictions. Lastly, due to the grid cell based approach, only half of the bacteria-
112 enriched seawater samples were actually taken into account by the model which is why further
113 expeditions must consider sampling with more spatial separation. Moving forward, this
114 relationship must be examined for terrestrial ecosystems as those generally comprise larger
115 differences in resource availability and environmental variables, potentially resulting in larger
116 differences in functional diversity, as well as for other domains as those govern key roles in
117 terrestrial ecosystem functioning.

118 **Materials and Methods**

119 *Data collection and correlation between taxonomy and diversity*

120 The publicly available data used to describe the structure and function of the global
121 ocean microbiome¹⁴ was downloaded from [http://ocean-](http://ocean-microbiome.embl.de/companion.html)
122 [microbiome.embl.de/companion.html](http://ocean-microbiome.embl.de/companion.html). 139 samples enriched in bacteria comprised the
123 taxonomic profile as annotated 16S OTU count table and the functional profile of prokaryotes
124 as eggNOG gene families annotated to the eggNOG version 3 database⁴ from the metagenome;

125 both derived from extracted DNA. The richness was determined as the number of different
126 eggNOG gene families or species. The diversity was determined as Shannon diversity H
127 according to Equation 1 where p_i is the relative abundance of the eggNOG gene families or
128 prokaryotic species.

129 Eq. 1: $H = -\sum_{i=1}^k p_i \log(p_i)$

130 The estimates on functional richness and functional diversity were modelled to species
131 richness as a linear, a logarithmic and a quadratic function using non-linear least squares in the
132 R package *nlme*¹⁵. The best fitted model was chosen based on the lowest Akaike's An
133 Information Criterion (AIC)¹⁶ with a penalty per parameter set to k equals two. The P-value of
134 their correlation was determined with the function *rcorr* from the R package *Hmisc*¹⁷ using the
135 Spearman's rank correlation. The pseudo coefficient of determination (R^2) of the non-linear
136 models were estimated with the function *Rsq* in the R package *soilphysics*¹⁸.

137 *Global diversity of functional richness*

138 To explore the geographic patterns of functional richness in prokaryotes inhabiting
139 aquatic ecosystems, I assigned the samples to 1x1 degree grid cells covering the globe. Grid-
140 based rather than locality-based analyses can be used to standardize the geographic scale of
141 the analysis, which facilitates cross-region comparisons and limits false presences in the data¹⁹.
142 The grid-based approach is broadly favored in biogeographic analyses for its suitability for
143 large-scale comparisons²⁰. In cells containing multiple samples, the sample with the highest
144 number of eggNOG families was selected, resulting in a total number of 74 samples. I used non-
145 parametric smoothing to investigate the changes in functionality (number of eggNOG families)
146 with latitude and longitude in second-degree polynomial terms added to the single or all
147 second-degree polynomial terms of six environmental variables (depth, generation time, nitrate
148 concentration, oxygen concentration, phosphate concentration and temperature). Nitrate
149 concentration combined with latitude and longitude showed the best fit of the data, which was
150 closest to the significance of the model with all environmental variables (**Table 2**). Then, each
151 combination of first- and second-degree polynomial terms for the three variables was modelled
152 and evaluated. The best fitting model used second-degree polynomial terms for latitude and

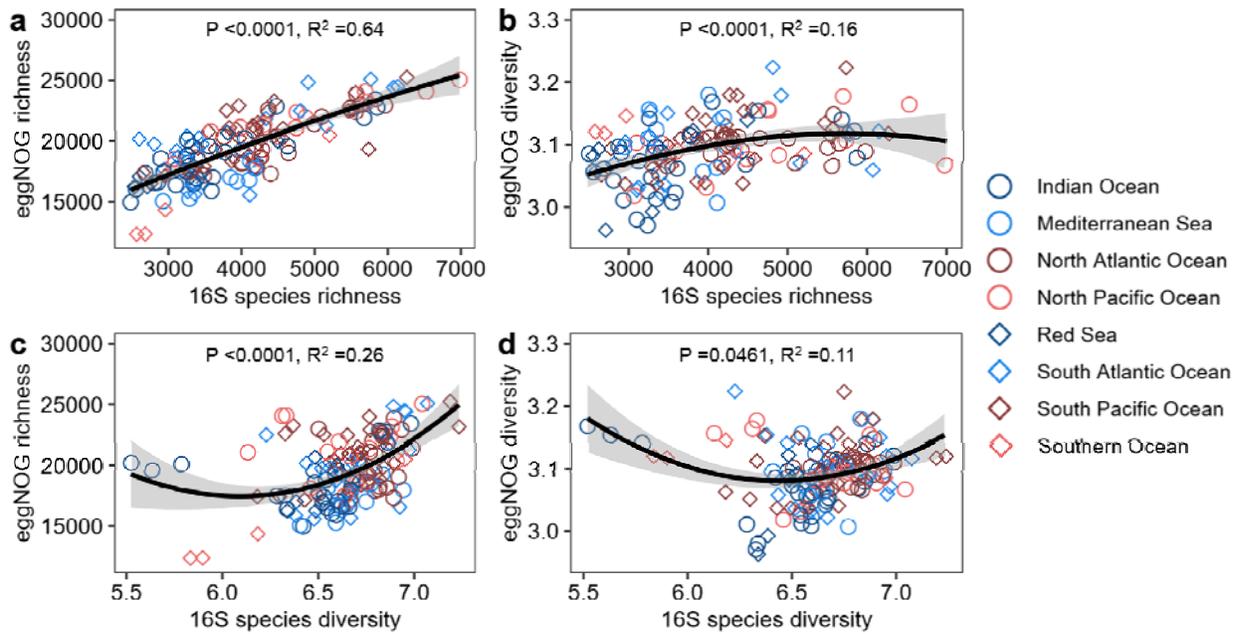
153 longitude combined with a first-degree polynomial term for nitrate concentration and was used
154 to predict functional richness in a 5x5 degree grid from -180 to 190 degrees longitude, -90 to 90
155 degrees latitude and -5 to 45 $\mu\text{mol/L}$ nitrate using the function *dpred* from the R package *iqspr*
156 ²¹. Admittedly, it is questionable that negative nitrate concentration exist but the data was
157 taken as it is available online and since it was present in 28 of 139 samples, their exclusion or
158 further data manipulation could potentially change the data structure. However, it could be the
159 reason for the very low functional richness with extreme latitudes.

160 **References**

- 161 1. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms:
162 proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci.* (1990).
163 doi:10.1073/pnas.87.12.4576
- 164 2. Falkowski, P. G., Barber, R. T. & Smetacek, V. Biogeochemical controls and feedbacks on
165 ocean primary production. *Science* (1998). doi:10.1126/science.281.5374.200
- 166 3. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc.*
167 *Natl. Acad. Sci. U. S. A.* (1998).
- 168 4. Powell, S. *et al.* eggNOG v3.0: Orthologous groups covering 1133 organisms at 41
169 different taxonomic ranges. *Nucleic Acids Res.* (2012). doi:10.1093/nar/gkr1060
- 170 5. Pham, V. H. T. & Kim, J. Cultivation of unculturable soil bacteria. *Trends in Biotechnology*
171 (2012). doi:10.1016/j.tibtech.2012.05.007
- 172 6. Martiny, A. C. High proportions of bacteria are culturable across major biomes. *ISME J.*
173 (2019).
- 174 7. Starke, R., Capek, P., Morais, D., Callister, S. J. & Jehmlich, N. The total microbiome
175 functions in bacteria and fungi. *J. Proteomics* 103623 (2019).
- 176 8. Locarnini, R. A. *et al.* *World Ocean Atlas 2013. Vol. 1: Temperature.* S. Levitus, Ed.; A.
177 *Mishonov, Technical Ed.; NOAA Atlas NESDIS* (2013). doi:10.1182/blood-2011-06-357442
- 178 9. Zweng, M. M. *et al.* *World Ocean Atlas 2013, Volume 2: Salinity.* NOAA Atlas NESDIS

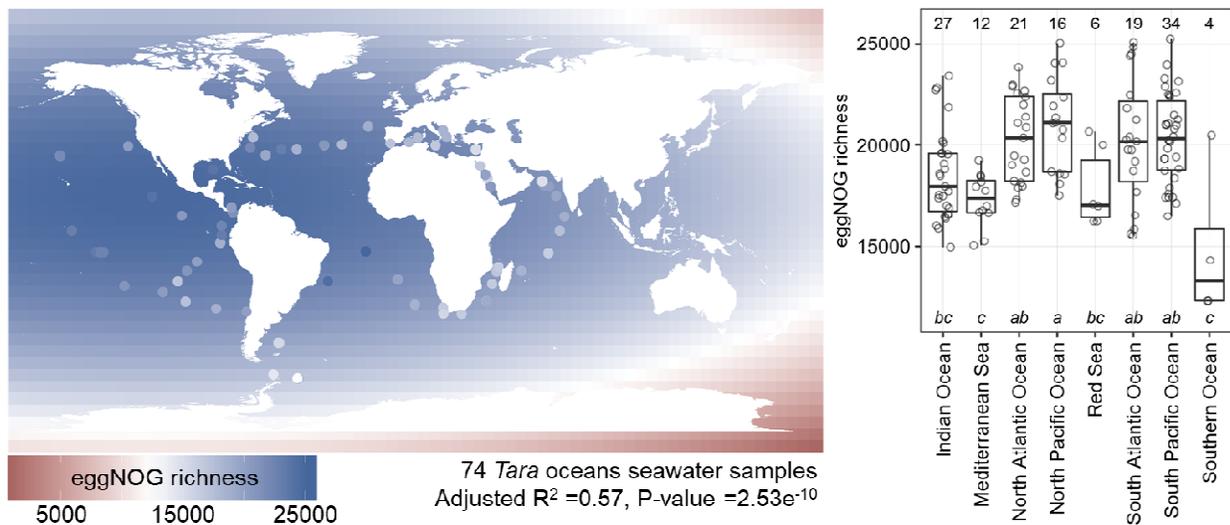
- 179 (2013). doi:10.1182/blood-2011-06-357442
- 180 10. Garcia, H. E. *et al.* World Ocean Atlas 2013. Volume 3: dissolved oxygen, apparent oxygen
181 utilization, and oxygen saturation. *NOAA Atlas NESDIS 75* (2013).
- 182 11. Garcia, H. E. *et al.* *World Ocean Atlas 2013, Volume 4*: Dissolved Inorganic Nutrients
183 (phosphate, nitrate, silicate). *NOAA Atlas NESDIS 76* (2013).
- 184 12. Miller, C. *Biological Oceanography*. (Blackwell Publishing, 2008).
- 185 13. Garcia, H. E. *et al.* *World Ocean Atlas 2009, Volume 4: Nutrients (phosphate, nitrate, and*
186 *silicate)*. *NOAA World Ocean Atlas* (2010).
- 187 14. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* (80-
188). (2015). doi:10.1126/science.1261359
- 189 15. Pinheiro, J., Bates, D., DebRoy, S. & Sarkar, D. nlme: Linear and Nonlinear Mixed Effects
190 Models. *R Dev. Core Team* (2007). doi:Doi 10.1038/Ncb1288
- 191 16. Bertrand, P. V., Sakamoto, Y., Ishiguro, M. & Kitagawa, G. Akaike Information Criterion
192 Statistics. *J. R. Stat. Soc. Ser. A (Statistics Soc.* (2006). doi:10.2307/2983028
- 193 17. Harrell, F. E. & Dupont, C. Package ‘Hmisc’: Harrell Miscellaneous. *R Top. Doc.* (2016).
- 194 18. da Silva, A. R. & de Lima, R. P. Soilphysics: An R package to determine soil
195 preconsolidation pressure. *Comput. Geosci.* (2015). doi:10.1016/j.cageo.2015.08.008
- 196 19. Hurlbert, A. H. & Jetz, W. Species richness, hotspots, and the scale dependence of range
197 maps in ecology and conservation. *Proc. Natl. Acad. Sci. U. S. A.* (2007).
198 doi:10.1073/pnas.0704469104
- 199 20. Větrovský, T. *et al.* A meta-analysis of global fungal distribution reveals climate-driven
200 patterns. *Nat. Commun.* (2019).
- 201 21. Ikebata, H., Hongo, K., Isomura, T., Maezono, R. & Yoshida, R. Bayesian molecular design
202 with a chemical language model. *J. Comput. Aided. Mol. Des.* (2017).
203 doi:10.1007/s10822-016-0008-z

205 **Figures**



206

207 **Figure 1:** The relationships as smoothed averages between species richness and diversity from
208 16S metabarcoding and functional richness or diversity of eggNOG functions from
209 metagenomes in 20L seawater samples from 68 locations waters across to globe. The adjusted
210 coefficient of determination (R^2) is given for the best fitting model for each equation: quadratic
211 (a), logarithmic (b), quadratic (c) and linear (d). The P-value was determined by Spearman's
212 rank correlation.



213
214 **Figure 2:** Global distribution of functional richness as eggNOG gene families from metagenomes
215 in 20L seawater samples from 68 locations waters across to globe using non-parametric
216 smoothing for 1x1 grid cells by additive second-degree polynomial models for latitude,
217 longitude, temperature, concentration of nitrate, oxygen and phosphate, and generation time.
218 Including nitrate concentration (AIC =1,138.6) showed the lowest AIC compared to the basic
219 model with latitude and longitude (AIC =1,360.4) and the model with all environmental
220 variables (AIC =1,115.7). The additive pairing of first- or second-degree polynomial terms for
221 latitude, longitude and nitrate concentration showed the lowest AIC value when a first-degree
222 polynomial term is used for nitrate concentration combined with second-degree polynomial
223 terms for latitude and longitude (AIC =1,136.7). The functional richness is also shown based on
224 the region of the different waters with the number of samples as numbers. Groups followed by
225 the same letter are not significantly different according to the HSD test (P-value >0.05).

226 **Tables**

227 **Table 1:** AICs and pseudo R²s of the linear (ln), the logarithmic (lg) and the quadratic (qu) model
228 to describe the relationship between functionality as eggNOG gene families from metagenomes
229 as eggNOG richness and eggNOG diversity to taxonomy as species from 16S metabarcoding and
230 functionality in the form of richness and Shannon diversity. The best fitting model is highlighted
231 in bold.

	eggNOG richness vs species richness		eggNOG diversity vs species richness		eggNOG richness vs species diversity		eggNOG richness vs species diversity	
	AIC	pseudo R ²	AIC	pseudo R ²	AIC	pseudo R ²	AIC	pseudo R ²
ln	2452.71	0.64	-469.09	0.14	2567.19	0.18	-448.27	<0.01
lg	2454.22	0.63	-470.65	0.15	2568.86	0.26	-448.19	<0.01
qu	2454.22	0.64	-470.72	0.16	2553.80	0.17	-461.90	0.11

232

233 **Table 1:** AICs of the basic model (Lat^{2nd}, Long^{2nd}) combined with single environmental variables (DE -
 234 depths, GT - generation time, Ni - nitrate concentration, Ox - oxygen concentration, Ph - Phosphate
 235 concentration and T - temperature) or altogether (Lat^{2nd}, Long^{2nd}, DE^{2nd}, GT^{2nd}, Ni^{2nd}, Ox^{2nd}, Ph^{2nd}, SA^{2nd},
 236 T^{2nd}) to describe the global distribution of functional richness as number of different eggNOG
 237 gene families from 139 seawater metagenomes in 1x1 grid cells. In cells containing multiple
 238 samples, the sample with the highest functional richness was used (n =74). The best fitting
 239 model for the comparison of individual environmental factors to the combination is shown in
 240 bold. Then, each combination of first- and second-degree polynomial terms for latitude,
 241 longitude and nitrate concentration was tested and the best fitting model shown in bold used
 242 to predict the global distribution of functional richness.

Model (variable ^{degree})	AIC
Lat ^{2nd} , Long ^{2nd}	1360.369
Lat ^{2nd} , Long ^{2nd} , DE ^{2nd}	1357.532
Lat ^{2nd} , Long ^{2nd} , GT ^{2nd}	1362.422
Lat^{2nd}, Long^{2nd}, Ni^{2nd}	1138.57
Lat ^{2nd} , Long ^{2nd} , Ox ^{2nd}	1339.934
Lat ^{2nd} , Long ^{2nd} , Ph ^{2nd}	1285.342
Lat ^{2nd} , Long ^{2nd} , T ^{2nd}	1336.751
Lat ^{2nd} , Long ^{2nd} , DE ^{2nd} , GT ^{2nd} , Ni ^{2nd} , Ox ^{2nd} , Ph ^{2nd} , SA ^{2nd} , T ^{2nd}	1115.652
Lat ^{1st} , Long ^{1st} , Ni ^{2nd}	1147.529
Lat ^{1st} , Long ^{2nd} , Ni ^{1st}	1147.579
Lat ^{2nd} , Long ^{1st} , Ni ^{1st}	1137.43
Lat ^{1st} , Long ^{2nd} , Ni ^{2nd}	1149.528
Lat ^{2nd} , Long ^{1st} , Ni ^{2nd}	1139.336
Lat^{2nd}, Long^{2nd}, Ni^{1st}	1136.7
Lat ^{2nd} , Long ^{2nd} , Ni ^{2nd}	1138.57
Lat ^{1st} , Long ^{1st} , Ni ^{1st}	1145.579

243

244 **Acknowledgements**

245 I acknowledge my colleague Daniel Morais for showing me the *Tara* ocean data and Petr Capek
246 for his advice on modelling and statistical analysis. This work was supported by the Czech
247 Science Foundation (20-02022Y).

248 **Author information**

249 **Affiliations**

250 *Laboratory of Environmental Microbiology, Institute of Microbiology of the Czech Academy of*
251 *Sciences, Prague, Czech Republic*

252 Robert Starke

253 **Contributions**

254 RS designed the study, analyzed the data and wrote the manuscript.