

# The sponge microbiome project

Lucas Moitinho-Silva<sup>1</sup>, Shaun Nielsen<sup>1</sup>, Amnon Amir<sup>2</sup>, Antonio Gonzalez<sup>2</sup>, Gail L. Ackermann<sup>2</sup>, Carlo Cerrano<sup>3</sup>, Carmen Astudillo-Garcia<sup>4</sup>, Cole Easson<sup>5</sup>, Detmer Sipkema<sup>6</sup>, Fang Liu<sup>7</sup>, Georg Steinert<sup>6</sup>, Giorgos Kotoulas<sup>7</sup>, Grace P. McCormack<sup>8</sup>, Guofang Feng<sup>9</sup>, James J. Bell<sup>10</sup>, Jan Vicente<sup>11</sup>, Johannes R Björk<sup>12</sup>, Jose M. Montoya<sup>13</sup>, Julie B. Olson<sup>14</sup>, Julie Reveillaud<sup>15</sup>, Laura Steindler<sup>16</sup>, Mari-Carmen Pineda<sup>17</sup>, Maria V. Marra<sup>9</sup>, Micha Ilan<sup>18</sup>, Michael W. Taylor<sup>3</sup>, Paraskevi Polymenakou<sup>8</sup>, Patrick M. Erwin<sup>19</sup>, Peter J. Schupp<sup>20</sup>, Rachel L. Simister<sup>21</sup>, Rob Knight<sup>2,22</sup>, Robert W. Thacker<sup>23</sup>, Rodrigo Costa<sup>24</sup>, Russell T. Hill<sup>25</sup>, Susanna Lopez-Legentil<sup>19</sup>, Thanos Dailianis<sup>8</sup>, Timothy Ravasi<sup>26</sup>, Ute Hentschel<sup>27</sup>, Zhiyong Li<sup>6</sup>, Nicole S. Webster<sup>17,28</sup> and Torsten Thomas<sup>1,\*</sup>

<sup>1</sup>Centre for Marine Bio-Innovation and School of Biological, Earth and Environmental Sciences, The University of New South Wales, Sydney, 2052, Australia

<sup>2</sup>Department of Pediatrics, University of California - San Diego, La Jolla, CA 92093, USA

<sup>3</sup>Department of Life and Environmental Sciences, Polytechnic University of Marche, Ancona, 60131, Italy

<sup>4</sup>School of Biological Sciences, University of Auckland, Auckland, New Zealand

<sup>5</sup>Halmos College of Natural Sciences and Oceanography, Nova Southeastern University, Dania Beach, FL 33004, USA

<sup>6</sup>Wageningen University, Laboratory of Microbiology, Stippeneng 4, 6708 WE Wageningen, The Netherlands

<sup>7</sup>Hellenic Centre for Marine Research, Institute of Marine Biology, Biotechnology and Aquaculture, Thalassocosmos 71500 Heraklion Greece

<sup>8</sup>Zoology, School of Natural Sciences, Ryan Institute, National University of Ireland Galway, University Rd., Galway, Ireland

<sup>9</sup>State Key Laboratory of Microbial Metabolism and School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, P.R. China

<sup>10</sup>School of Biological Sciences, Victoria University of Wellington, Wellington, New Zealand

<sup>11</sup>Hawaii Institute of Marine Biology, 46-007 Lilipuna Road, Kaneohe, HI 96744-1346

<sup>12</sup>Galvin Life Science Center, University of Notre Dame, Notre Dame, IN 46556, USA and Ecological Networks and Global Change Group, Theoretical and Experimental Ecology Station, CNRS, Moulis, France

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65
- 34 <sup>13</sup>Ecological Networks and Global Change Group, Theoretical and Experimental Ecology  
Station, CNRS and Paul Sabatier University, Moulis, France
- 35
- 36 <sup>14</sup>Department of Biological Sciences, University of Alabama, Tuscaloosa, AL 35487, USA
- 37 <sup>15</sup>INRA, UMR1309 CMAEE; Cirad, UMR15 CMAEE, 34398 Montpellier, France
- 38 <sup>16</sup>Department of Marine Biology, Leon H. Charney School of Marine Sciences, University of  
Haifa, Haifa, Israel
- 39
- 40 <sup>17</sup>Australian Institute of Marine Science (AIMS), Townsville, 4810, Queensland, Australia
- 41 <sup>18</sup>Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel  
Aviv 69978, Israel
- 42
- 43 <sup>19</sup>Department of Biology and Marine Biology, University of North Carolina Wilmington,  
Wilmington NC 28409, USA
- 44
- 45 <sup>20</sup>Institute for Chemistry and Biology of the Marine Environment (ICBM), Carl-von-Ossietzky  
and University Oldenburg, Schleusenstr. 1, 26382 Wilhelmshaven, Germany
- 46
- 47 <sup>21</sup>Department of microbiology and immunology, University of British Columbia, Canada, V6T  
1Z3
- 48
- 49 <sup>22</sup>Department of Computer Science and Engineering, and Center for Microbiome Innovation,  
University of California - San Diego, La Jolla, CA 92093, USA
- 50
- 51 <sup>23</sup>Department of Ecology and Evolution, Stony Brook University, Stony Brook NY 11794, USA
- 52 <sup>24</sup>Institute for Bioengineering and Biosciences (IBB), Department of Bioengineering, IST,  
Universidade de Lisboa, Lisbon, Portugal
- 53
- 54 <sup>25</sup>Institute of Marine and Environmental Technology, University of Maryland Center for  
Environmental Science, 701 East Pratt Stree, Baltimore, MD 21202, USA
- 55
- 56 <sup>26</sup>KAUST Environmental Epigenetic Program (KEEP), Division of Biological and Environmental  
Sciences & Engineering, King Abdullah University of Science and Technology, Thuwal,  
Kingdom of Saudi Arabia
- 57
- 58 <sup>27</sup>RD3 Marine Microbiology, GEOMAR Helmholtz Centre for Ocean Research, Kiel, and  
Christian-Albrechts-University of Kiel, Germany
- 59
- 60 <sup>28</sup>Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences,  
University of Queensland, St Lucia, QLD, Australia
- 61
- 62
- 63
- 64
- 65

\*Corresponding author: [t.thomas@unsw.edu.au](mailto:t.thomas@unsw.edu.au); Centre for Marine Bio-Innovation,  
UNSW, Sydney, Australia

68           **Abstract**

69           *Background:* Marine sponges (phylum Porifera) are a diverse, phylogenetically deep-  
70 branching clade known for forming intimate partnerships with complex communities of  
71 microorganisms. To date, 16S rRNA gene sequencing studies have largely utilised different extraction  
72 and amplification methodologies to target the microbial communities of a limited number of sponge  
73 species, severely limiting comparative analyses of sponge microbial diversity and structure. Here, we  
74 provide an extensive and standardised dataset that will facilitate sponge microbiome comparisons  
75 across large spatial, temporal and environmental scales.

76           *Findings:* Samples from marine sponges (n=3568 specimens), seawater (n=370), marine  
77 sediments (n=65) and other environments (n=29) were collected from different locations across the  
78 globe. This dataset incorporates at least 269 different sponge species, including several yet  
79 unidentified taxa. The V4 region of the 16S rRNA gene was amplified and sequenced from extracted  
80 DNA using standardised procedures. Raw sequences (total of 1.1 billion sequences) were processed  
81 and clustered with a) a standard protocol using QIIME closed-reference picking resulting in 39,543  
82 Operational Taxonomic Units (OTU) at 97% sequence identity, b) a *de novo* protocol using Mothur  
83 resulting in 518,246 OTUs, and c) a new high-resolution Deblur protocol resulting in 83,908 unique  
84 bacterial sequences. Abundance tables, representative sequences, taxonomic classifications and  
85 metadata are provided.

86           *Conclusions:* This dataset represents a comprehensive resource of sponge-associated  
87 microbial communities based on 16S rRNA gene sequences that can be used to address overarching  
88 hypotheses regarding host-associated prokaryotes, including host-specificity, convergent evolution,  
89 environmental drivers of microbiome structure and the sponge-associated rare biosphere.

90

91           **Keywords:** Marine sponges, Archaea, Bacteria, Symbiosis, Microbiome, 16S rRNA gene,  
92 Microbial diversity

93

## 94 Data Description

### 95 *Purpose of data acquisition*

96 Sponges (phylum Porifera) are an ancient metazoan clade [1], with more than 8,500 formally  
97 described species [2]. Sponges are benthic organisms that have important ecological functions in  
98 aquatic habitats [3, 4]. Marine sponges are often found in symbiotic association with  
99 microorganisms and these microbial communities can be very diverse and complex [5, 6]. Sponge  
100 symbionts perform a wide range of functional roles, including vitamin synthesis, production of  
101 bioactive compounds and biochemical transformations of nutrients or waste products [7-9]. The  
102 diversity of microorganisms associated with sponges has been the subject of intense study (the  
103 search of “sponge microbial diversity” returned 348 publications in Scopus database [10]. Most of  
104 these studies were performed on individual species from restricted geographic regions [e.g., 11, 12].  
105 A comparative assessment of these studies is often hindered by differences in sample processing  
106 and 16S rRNA gene sequencing. However, two recent studies incorporating a large number of  
107 sponge microbiomes (> 30) [5, 13] revealed the potential of large-scale, standardised, high-  
108 throughput sequencing for gaining unique insights into the diversity and structure of sponge-  
109 associated microbial communities. The purpose of this global dataset is to provide a comprehensive  
110 16S rRNA gene-based resource for investigating and comparing microbiomes more generally across  
111 the phylum Porifera.

### 112 *Sample collection, processing and 16S rRNA gene sequencing*

113 Sample collection and processing, species identification and DNA extractions were  
114 conducted as previously described [13]. A total of 3568 sponge specimens were collected,  
115 representing at least 268 species, including several yet unidentified taxa (hereafter collectively  
116 referred to as species) (Supplementary Table S1). Of the total species, 213 were represented by at  
117 least three specimens. *Carteriospongia foliascens* had the highest replication comprising 150  
118 individuals. Seawater (n=370), sediment (n=65), algae (n=1) and echinoderm (n=1) samples as well as  
119 biofilm swabs (n=21) of rock surfaces were collected in close proximity to the sponges for  
120 comparative community analysis. Six negative control samples (sterile water) were processed to  
121 identify any potential contaminations. Of the samples included in this current dataset, 973 samples  
122 had been analysed previously [13]. Samples were collected from a wide range of geographical  
123 locations (Figure 1 and Supplementary Table S1). Total DNA was extracted as previously described  
124 [13] and used as templates to amplify and sequence the V4 region of the 16S rRNA gene using the  
125 standard procedures of the EMP [14, 15].

126 *Processing of sequencing data*

127 Clustering using the EMP standard protocols in QIIME:

128 Quality-filtered, demultiplexed fastq files were processed using the default closed-reference  
129 pipeline from QIIME v. 1.9.1, providing the EMP standard method for cross-dataset comparisons and  
130 allowing direct comparison with the tens of thousands of other samples processed in the EMP and  
131 available via the Qiita database [16]. Briefly, sequences were matched against GreenGenes (v. 13\_8  
132 at 97% similarity) reference database. Sequences that failed to align were discarded. Taxonomy for  
133 each sequence was taken from the cluster to which it aligned.

134  
135 Clustering using Mothur:

136 Quality-filtered, demultiplexed fastq files were also processed using mothur v. 1.37.6 [17]  
137 and Python v. 2.7 [18] custom scripts with modifications from previously established protocols [13].  
138 Detailed descriptions and command outputs are available at the project notebook (see Availability of  
139 supporting data). Briefly, sequences were quality-trimmed to a maximum length of 100 bp. To  
140 minimize computational effort, the dataset was reduced to unique sequences, retaining total  
141 sequence counts. Sequences were aligned to the V4 region of the 16S rRNA gene sequences from  
142 the SILVA v 123 database [19]. Sequences that aligned at the expected positions were kept and this  
143 dataset was again reduced to unique sequences. Further, singletons were removed from the dataset  
144 and remaining sequences were pre-clustered if they differed by one nucleotide position. Sequences  
145 classified as eukaryote, chloroplast, mitochondria or unknown according to the Greengenes (v. 13\_8  
146 at 99% similarity) [20] and SILVA taxonomies [21] were removed. Chimeras were identified with  
147 UCHIME [22] and removed. Finally, sequences were *de novo* clustered into Operational Taxonomic  
148 Units (OTUs) using the furthest neighbour method at 97% similarity. Representative sequences of  
149 OTUs were retrieved based on the mean distance among the clustered sequences. Consensus  
150 taxonomies based on the SILVA, Greengenes and RDP (v. 14\_032015) [23] databases were obtained  
151 based on the classification of sequences clustered within each OTU.

152  
153 De-noising using Deblur:

154 Recently, sub-OTU methods that allow views of the data at single-nucleotide resolution have  
155 become available. One such methods is Deblur [24], which is a denoising algorithm for identification  
156 of actual bacterial sequences present in a sample. Using an upper bound on the PCR and read-error

157 rates, Deblur processes each sample independently and outputs the list of sequences and their  
158 frequencies in each sample, enabling single nucleotide resolution. For creating the deblurred biom  
159 table, quality filtered, demultiplexed fasta files were used as input to Deblur using a trim length of  
160 100, and min-reads of 25 (removing sOTUs with < 25 reads total in all samples combined). Taxonomy  
161 was added to resulting biom table using QIIME [25], RDP classifier [26] and Greengenes 13.8 [20].

#### 163 Database metadata category enrichment:

164 For enrichment analysis of metadata terms in a set of sequences, each unique metadata  
165 value is tested using a binomial test. For a bacterial sequence  $s$  and metadata value  $v$ , denote  $N$  the  
166 total number of samples,  $O(s)$  the number of samples where  $s$  is present,  $K_v(s)$  the number of sample  
167 with value  $v$  where  $s$  is present, and  $T(v)$  the total number of samples with value  $v$ . The p-value for  
168 enrichment was then calculated as:

$$169 \quad p\text{-value} = \text{binomial\_cdf} ( T(v)-K_v(s), T(v), P_{\text{Null}}(s) )$$

$$170 \quad \text{where } P_{\text{Null}}(s) = O(s) / N$$

171 We have set up a webserver ([www.spongeemp.com](http://www.spongeemp.com)) that performs this enrichment analysis for  
172 user-defined sequence submissions. The code for the webserver is also available in Github [27] for a  
173 local installation.

#### 175 *Data description*

176 The dataset covers 4032 samples with a total of 1,167,226,701 raw sequence reads. These  
177 sequence reads clustered into 39,543 OTUs using QIIME's closed-reference processing, 518,246  
178 OTUs from *de novo* clustering using Mothur (not filtered for OTU abundances), and 83,908 sOTUs  
179 using Deblur (with a filtering of at least 25 reads total per sOTU). We recommend that data users  
180 consider the differences in sequencing depths per sample and abundance filtering for certain  
181 downstream analyses, such as when calculating diversity estimates [28] and comparing OTU  
182 abundances across samples [29]. In terms of taxonomic diversity, most Mothur OTUs were assigned  
183 to the phylum Proteobacteria, although more than 60 different microbial phyla were recovered from  
184 the marine sponge samples according to SILVA ( $n=63$ ) and Greengenes classifications ( $n=72$ ) (Figure  
185 2).

187           **Potential uses**

1  
2  
3 188           This dataset can be utilised to assess a broad range of ecological questions pertaining to  
4  
5 189   host-associated microbial communities generally or to sponge microbiology specifically. These  
6  
7 190   include: i) the degree of host-specificity, ii) the existence of biogeographic or environmental  
8  
9 191   patterns, iii) the relation of microbiomes to host phylogeny, iv) the variability of microbiomes within  
10  
11 192   or between host species, v) symbiont co-occurrence patterns as well as vi) assessing the existence of  
12  
13 193   a core sponge microbiome. An example of this type of analysis is shown in Figure 3, where samples  
14  
15 194   were clustered using unweighted UniFrac data [30] with a Principal Coordinate Analysis and  
16  
17 195   visualization in Emperor [31] based on their origins from sponges, seawater or kelps [32].  
18  
19  
20  
21

22 196

25 197           **Availability and requirements**

28 198           Project name: The Sponge Microbiome Project

30  
31 199           Project     home     page:     GigaScience     repository;     www.spongeemp.com;  
32 200   <https://github.com/amnona/SpongeEMP>  
33

35 201           Operating system(s): Unix

37 202           Programming language: Python and R

39  
40 203           Other requirements: Python v. 2.7, Biopython v. 1.65, Python 3.5, R v. 3.2.2, mothur v.  
41 204   1.37.6, QIIME v. 1.9.1, Deblur  
42

44 205           License: MIT

46 206           Any restrictions to use by non-academics: None  
47  
48

49 207

52 208           **Availability of supporting data**

55 209           Raw sequence data were deposited in the European Nucleotide Archive (accession numbers:  
56 210   ERP020690). Quality-filtered, demultiplexed fastq files, Deblur and QIIME resulting OTU tables are  
57  
58 211   available at Qiita database [16] (Study ID: 10793). The additional datasets that support the results of  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

212 this article are available in the GigaScience repository (**DOI of the dataset**) and include an OTU  
213 abundance matrix (the output “.shared” file from mothur, which is tab delimited), an OTU taxonomic  
214 classification table (tab delimited text file), an OTU representative sequence FASTA file, and a table  
215 of samples’ metadata. The project workflow, mothur commands and additional scripts are available  
216 as HTML TiddlyWiki notebook [33], which is viewed in any browser (**DOI of the workflow**).

217 The deblurred dataset has also been uploaded to an online server [34] that supplies both  
218 html and REST-API access for querying bacterial sequences and obtaining the observed prevalence  
219 and enriched metadata categories where the sequence is observed (Figure 4). This allows an  
220 interactive view of which sequences are associated with which specific parameters, such as depth or  
221 salinity.

#### 222 223 224 **List of abbreviations**

225 bp: base pairs

226 OTU: operational taxonomic unit

227 rRNA: ribosomal RNA

#### 228 **Competing interests**

229 The authors declare that they have no competing interests.

#### 230 **Funding**

231 T.T. and N.S.W were funded by an Australian Research Council Future Fellowship  
232 FT140100197 and FT120100480, respectively. T.T. received funds from the Gordon and Betty Moore  
233 Foundation. This work was also supported in part by the W.M. Keck Foundation and the John  
234 Templeton Foundation. R.K. received funding as a Howard Hughes Medical Institute Early Career  
235 Scientist.



236

237 **Authors' contributions**

238 L.M.-S., N.S.W. and T.T. designed the study. C.A.G., D.S., F.L., G.S., G.K., G.McC., G.-F. F, J.J.B.,  
239 J.V., J.R.B., J.M.M., J.R., L.S., M.C.P, M.V.M., M.W.T., N.S.W., P.P., P.M.E., P.J.S., R.L.S, R.W.T., R.C.,  
240 R.T.H., S.L-L., T.D., T.R., U.H. and Z-Y. L. collected samples. C.A.G., D.S., J.V., J.R.B., L.S., M.C.P.,  
241 M.W.T., N.S.W., P.M.E., R.L.S, R.W.T., S.L-L. and U.H. extracted DNA. G.L.A. and R.K. sequenced DNA.  
242 L.M.-S., S.N., A.A., A.G., G.L.A. and T.T. performed data processing and analysis. L.M.-S., N.S.W. and  
243 T.T. wrote the manuscript. All authors contributed to the writing of the manuscript.

245 **References**

- 246 1. Li CW, Chen JY and Hua TE. Precambrian sponges with cellular structures. *Science*. 1998;279  
247 5352:879-82.
- 248 2. Van Soest RW, Boury-Esnault N, Vacelet J, Dohrmann M, Erpenbeck D, De Voogd NJ, et al.  
249 Global diversity of sponges (Porifera). *PLoS One*. 2012;7 4:e35105.  
250 doi:10.1371/journal.pone.0035105.
- 251 3. Bell JJ. The functional roles of marine sponges. *Estuar Coast Shelf S*. 2008;79 3:341-53.
- 252 4. de Goeij JM, van Oevelen D, Vermeij MJ, Osinga R, Middelburg JJ, de Goeij AF, et al. Surviving  
253 in a marine desert: the sponge loop retains resources within coral reefs. *Science*. 2013;342  
254 6154:108-10. doi:10.1126/science.1241981.
- 255 5. Schmitt S, Tsai P, Bell J, Fromont J, Ilan M, Lindquist N, et al. Assessing the complex sponge  
256 microbiota: core, variable and species-specific bacterial communities in marine sponges. *ISME*  
257 *J*. 2012;6 3:564-76. doi:10.1038/ismej.2011.116.
- 258 6. Webster NS, Taylor MW, Behnam F, Lucker S, Rattei T, Whalan S, et al. Deep sequencing  
259 reveals exceptional diversity and modes of transmission for bacterial sponge symbionts.  
260 *Environ Microbiol*. 2010;12 8:2070-82. doi:10.1111/j.1462-2920.2009.02065.x.
- 261 7. Siegl A, Kamke J, Hochmuth T, Piel J, Richter M, Liang C, et al. Single-cell genomics reveals the  
262 lifestyle of Poribacteria, a candidate phylum symbiotically associated with marine sponges.  
263 *ISME J*. 2011;5 1:61-70. doi:10.1038/ismej.2010.95.
- 264 8. Taylor MW, Radax R, Steger D and Wagner M. Sponge-associated microorganisms: evolution,  
265 ecology, and biotechnological potential. *Microbiol Mol Biol Rev*. 2007;71 2:295-347.  
266 doi:10.1128/MMBR.00040-06.
- 267 9. Wilson MC, Mori T, Ruckert C, Uria AR, Helf MJ, Takada K, et al. An environmental bacterial  
268 taxon with a large and distinct metabolic repertoire. *Nature*. 2014;506 7486:58-62.  
269 doi:10.1038/nature12959.
- 270 10. Scopus database. <https://www.scopus.com/>. Accessed 1 Dec 2016.
- 271 11. Moitinho-Silva L, Bayer K, Cannistraci CV, Giles EC, Ryu T, Seridi L, et al. Specificity and  
272 transcriptional activity of microbiota associated with low and high microbial abundance  
273 sponges from the Red Sea. *Mol Ecol*. 2014;23 6:1348-63. doi:10.1111/mec.12365.
- 274 12. Montalvo NF and Hill RT. Sponge-associated bacteria are strictly maintained in two closely  
275 related but geographically distant sponge hosts. *Appl Environ Microbiol*. 2011;77 20:7207-16.  
276 doi:10.1128/AEM.05285-11.

- 277 13. Thomas T, Moitinho-Silva L, Lurgi M, Bjork JR, Easson C, Astudillo-Garcia C, et al. Diversity,  
1 278 structure and convergent evolution of the global sponge microbiome. *Nat Commun.*  
2 279 2016;7:11870. doi:10.1038/ncomms11870.
- 3 280 14. Gilbert JA, Jansson JK and Knight R. The Earth Microbiome project: successes and aspirations.  
4 281 *BMC Biol.* 2014;12:69. doi:10.1186/s12915-014-0069-1.
- 5 282 15. EMP 16S Illumina amplicon protocol. [http://www.earthmicrobiome.org/emp-standard-  
6 283 protocols/16s/](http://www.earthmicrobiome.org/emp-standard-protocols/16s/). Accessed 1 Dec 2016.
- 7 284 16. Qiita database <http://qiita.microbio.me/>. Accessed 31 Mar 2017.
- 8 285 17. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing  
9 286 mothur: open-source, platform-independent, community-supported software for describing  
10 287 and comparing microbial communities. *Appl Environ Microbiol.* 2009;75 23:7537-41.  
11 288 doi:10.1128/AEM.01541-09.
- 12 289 18. Python. <https://www.python.org/>. Accessed 31 Mar 2017.
- 13 290 19. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA  
14 291 gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*  
15 292 2013;41 Database issue:D590-6. doi:10.1093/nar/gks1219.
- 16 293 20. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a  
17 294 chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ  
18 295 Microbiol.* 2006;72 7:5069-72. doi:10.1128/AEM.03006-05.
- 19 296 21. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, et al. The SILVA and "All-species  
20 297 Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* 2014;42 Database  
21 298 issue:D643-8. doi:10.1093/nar/gkt1209.
- 22 299 22. Edgar RC, Haas BJ, Clemente JC, Quince C and Knight R. UCHIME improves sensitivity and  
23 300 speed of chimera detection. *Bioinformatics.* 2011;27 16:2194-200.  
24 301 doi:10.1093/bioinformatics/btr381.
- 25 302 23. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data  
26 303 and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 2014;42 Database issue:D633-  
27 304 42. doi:10.1093/nar/gkt1244.
- 28 305 24. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur Rapidly  
29 306 Resolves Single-Nucleotide Community Sequence Patterns. *mSystems.* 2017;2 2  
30 307 doi:10.1128/mSystems.00191-16.
- 31 308 25. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME  
32 309 allows analysis of high-throughput community sequencing data. *Nat Methods.* 2010;7 5:335-  
33 310 6. doi:10.1038/nmeth.f.303.
- 34 311 26. Wang Q, Garrity GM, Tiedje JM and Cole JR. Naive Bayesian classifier for rapid assignment of  
35 312 rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol.* 2007;73 16:5261-  
36 313 7. doi:10.1128/AEM.00062-07.
- 37 314 27. SpongeEMP GitHub. <https://github.com/amnona/SpongeEMP>. Accessed 31 Mar 2017.
- 38 315 28. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, et al. Quality-filtering  
39 316 vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods.*  
40 317 2013;10 1:57-9. doi:10.1038/nmeth.2276.
- 41 318 29. McMurdie PJ and Holmes S. Waste not, want not: why rarefying microbiome data is  
42 319 inadmissible. *PLoS Comput Biol.* 2014;10 4:e1003531. doi:10.1371/journal.pcbi.1003531.
- 43 320 30. Lozupone C and Knight R. UniFrac: a new phylogenetic method for comparing microbial  
44 321 communities. *Appl Environ Microbiol.* 2005;71 12:8228-35. doi:10.1128/AEM.71.12.8228-  
45 322 8235.2005.
- 46 323 31. Vazquez-Baeza Y, Pirrung M, Gonzalez A and Knight R. EMPeror: a tool for visualizing high-  
47 324 throughput microbial community data. *Gigascience.* 2013;2 1:16. doi:10.1186/2047-217X-2-  
48 325 16.
- 49 326 32. Marzinelli EM, Campbell AH, Zozaya Valdes E, Verges A, Nielsen S, Wernberg T, et al.  
50 327 Continental-scale variation in seaweed host-associated bacterial communities is a function of

328  
1 329  
2 330  
3 331  
4 332  
5  
6 333  
7  
8  
9 334  
10  
11  
12 335  
13  
14 336  
15  
16  
17 337  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31 338  
32  
33  
34 339  
35  
36 340  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

host condition, not geography. *Environ Microbiol.* 2015;17 10:4078-88. doi:10.1111/1462-2920.12972.  
33. TiddlyWiki. <http://tiddlywiki.com/>. Accessed 31 Mar 2017.  
34. Sponge microbiome project deblurred dataset online server. [www.spongeemp.com](http://www.spongeemp.com). Accessed 31 Mar 2017.

## Figures

Figure 1.

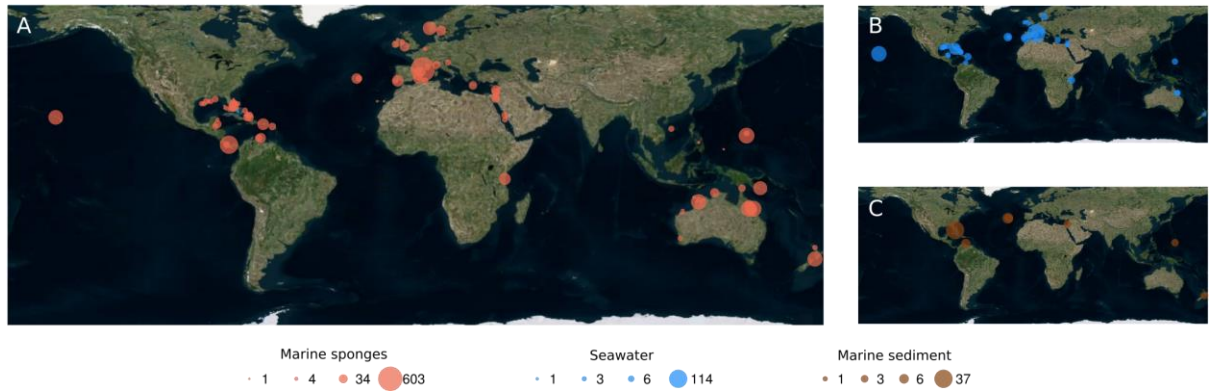
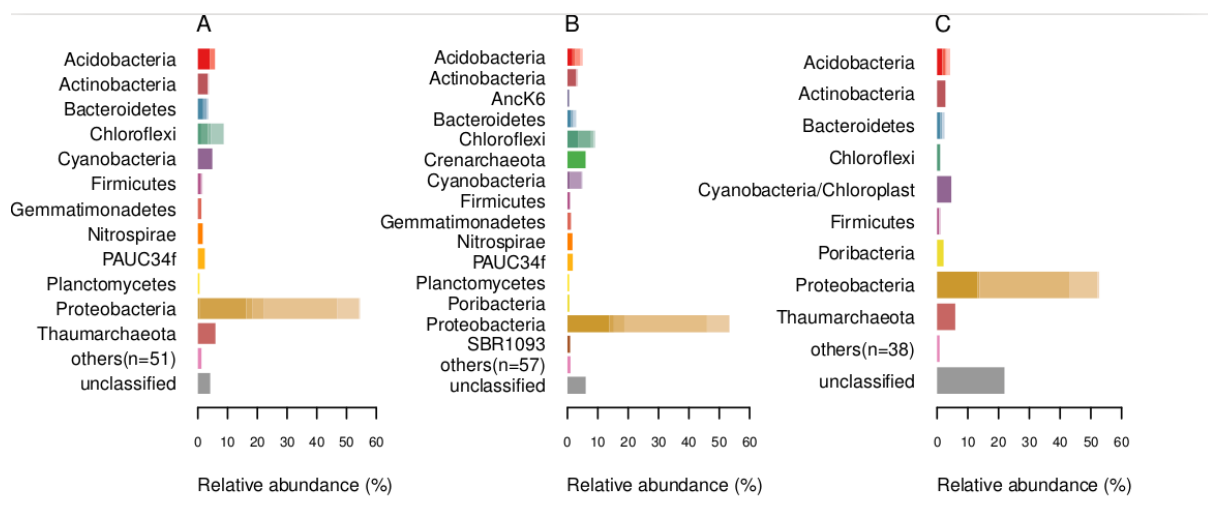


Figure 2.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65



**SILVA**

- Acidobacteria;Acidobacteria
- Acidobacteria;Holophagae
- Actinobacteria;Acidimicrobiia
- Bacteroidetes;Cytophagia
- Bacteroidetes;Flavobacteriia
- Chloroflexi;Anaerolineae
- Chloroflexi;Caldilineae
- Chloroflexi;Chloroflexi\_unclassified
- Chloroflexi;SAR202\_clade
- Cyanobacteria;Cyanobacteria
- Firmicutes;Clostridia
- Gemmatimonadetes;Gemmatimonadetes
- Nitrospirae;Nitrospira
- PAUC34f;PAUC34f\_unclassified
- Proteobacteria;Alphaproteobacteria
- Proteobacteria;Betaproteobacteria
- Proteobacteria;Deltaproteobacteria
- Proteobacteria;Gammaproteobacteria
- Proteobacteria;Proteobacteria\_unclassified
- Thaumarchaeota;Marine\_Group\_I

**Greengenes**

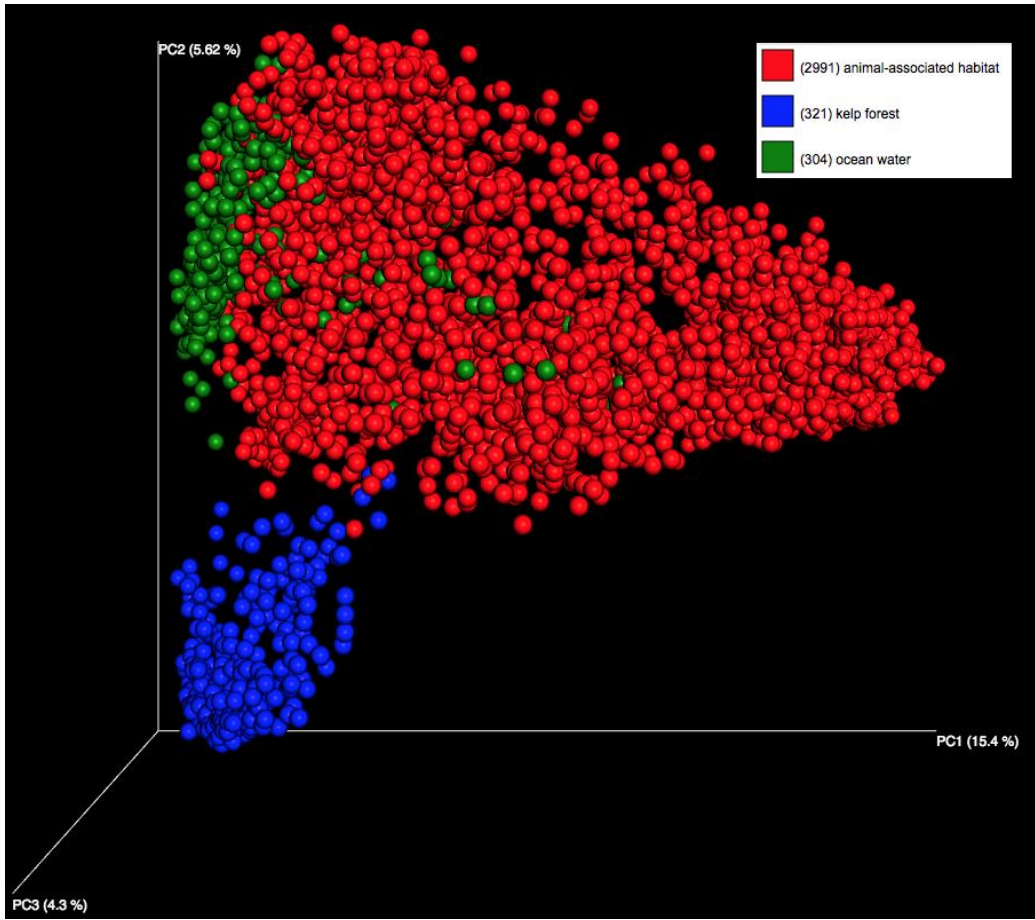
- Acidobacteria;Acidobacteria-6
- Acidobacteria;Sva0725
- Actinobacteria;Acidimicrobiia
- Bacteroidetes;Flavobacteriia
- Chloroflexi;Anaerolineae
- Chloroflexi;SAR202
- Crenarchaeota;Thaumarchaeota
- Cyanobacteria;Synechococcophycideae
- Gemmatimonadetes;Gemm-2
- Nitrospirae;Nitrospira
- PAUC34f;PAUC34f\_unclassified
- Proteobacteria;Alphaproteobacteria
- Proteobacteria;Betaproteobacteria
- Proteobacteria;Deltaproteobacteria
- Proteobacteria;Gammaproteobacteria
- Proteobacteria;Proteobacteria\_unclassified

**RDP**

- Acidobacteria;Acidobacteria\_Gp10
- Acidobacteria;others
- Actinobacteria;Actinobacteria
- Bacteroidetes;Flavobacteriia
- Cyanobacteria;Chloroplast;Cyanobacteria
- Poribacteria;Poribacteria\_unclassified
- Proteobacteria;Alphaproteobacteria
- Proteobacteria;Gammaproteobacteria
- Proteobacteria;Proteobacteria\_unclassified
- Thaumarchaeota;Nitrosopumilales

341

342 Figure 3.



343

344

345 Figure 4.

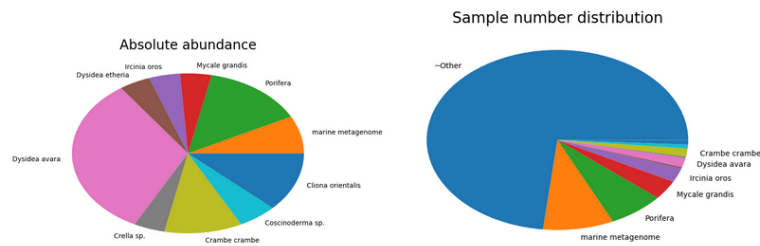


### Search results

taxonomy: k\_Bacteria;p\_Proteobacteria;c\_Alphaproteobacteria;o\_Rhizobiales

sequence:  
TACGAAGGGGGCTAGCGTTGTTCGGAATCACTGGGCGTAAAGCGCACGTAGGCGGACTTTTAAGTCAGGGGTGAAATCCCGGGGCTCAACCCCGGAACCTG  
[More info from dbBact](#)  
Present in 0.034474 of samples (132 / 3829)

▼ host\_scientific\_name (6 significant)



**Significant enrichment:**  
 host\_scientific\_name:Dysidea avara (30/64)  
 host\_scientific\_name:Crella sp. (4/9)  
 host\_scientific\_name:Dysidea etheria (4/10)  
 host\_scientific\_name:Cliona orientalis (11/31)  
 host\_scientific\_name:Coscinoderma sp. (5/27)  
 host\_scientific\_name:Crambe crambe (10/56)

- ▶ env\_feature (1 significant)
- ▶ country (3 significant)
- ▶ ALL (84 significant)

346

### Legends

347 Figure 1. Global sample collection sites. Bubbles indicate collection sites of (A) marine  
348 sponges, (B) seawater and (C) marine sediment samples. Bubble sizes are proportional to number of  
349 samples as indicated.

350  
351 Figure 2. Microbial taxonomic profile of marine sponge samples classified using (A) SILVA, (B)  
352 Greengenes and (C) RDP. OTU sequence counts were grouped according to phylum and class. Taxa  
353 with relative abundances  $\leq 0.5\%$  were grouped as 'others'. Classes with relative abundances  $> 1\%$   
354 are shown in the legend (phylum “;” class). Relative abundances are represented on the x-axes.

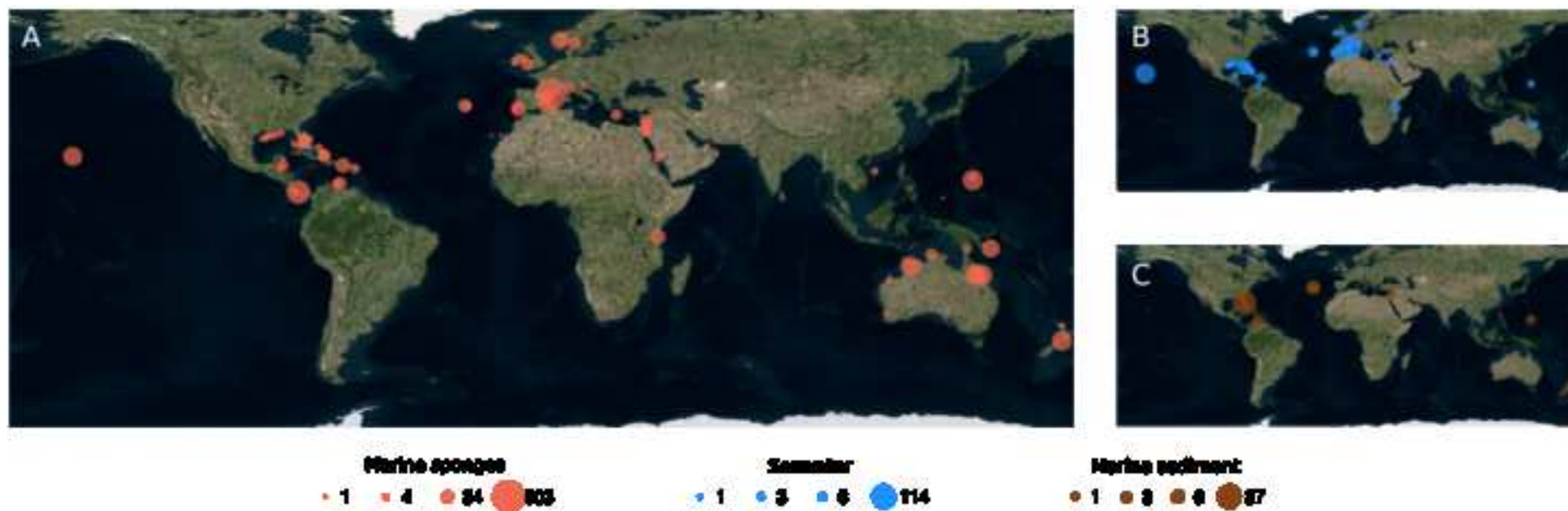
355 Figure 3. Unweighted UniFrac Principal Coordinates Analysis (PCA) of samples from sponges  
356 (“animal=associated”, red), ocean water (green) and kelp (blue). A separation can be seen between  
357 samples based to the environmental origin. Samples were rarefying to 10,000 sequences per sample.

358  
359 Figure 4. Output of the enrichment analysis through the online server  
360 www.spongeemp.com. Top line shows taxonomic assignment for the user-submitted sequence in  
361 the second line. Pie charts below show relative abundance and sample distribution plus their

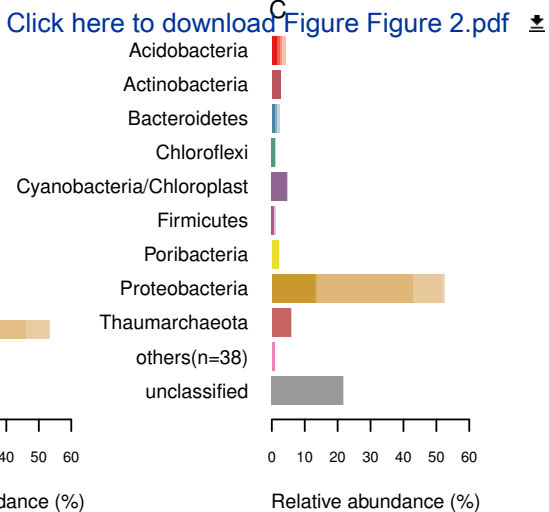
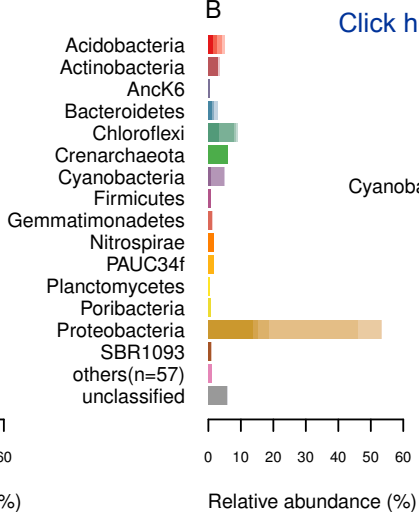
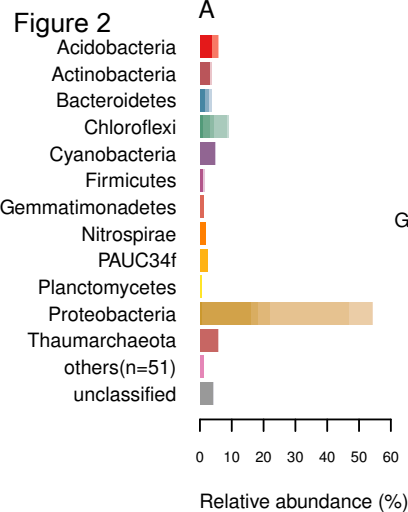
362 associated significant enrichment results for the submitted sequence based on the scientific names  
1 363 of the host. At the bottom, fields can be opened to show results of the enrichment analyses for  
2  
3 364 other metadata types (e.g. country).  
4

5  
6 365

7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65







[Click here to download Figure Figure 2.pdf](#)

### SILVA

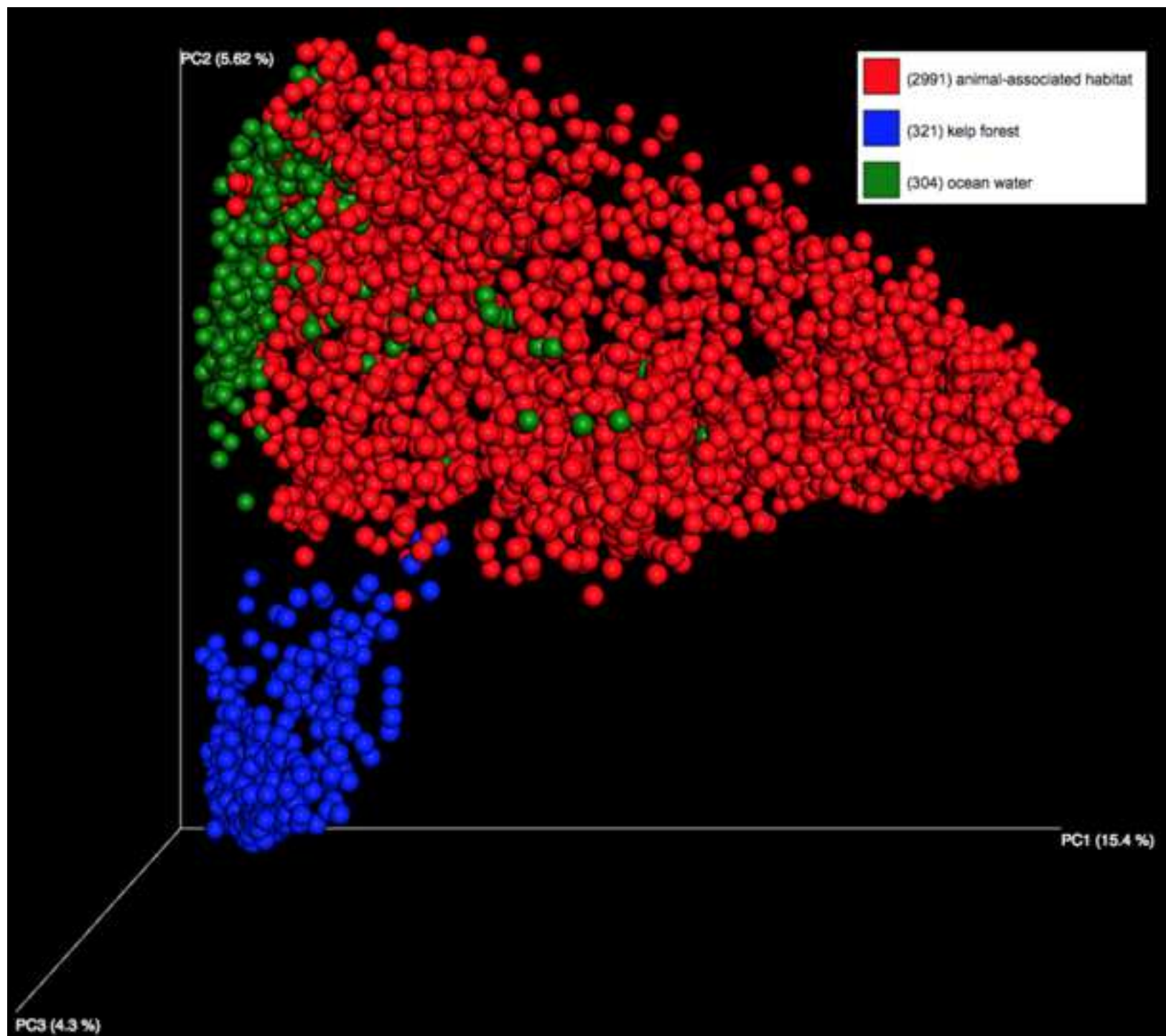
- Acidobacteria;Acidobacteria
- Acidobacteria;Holophagae
- Actinobacteria;Acidimicrobiia
- Bacteroidetes;Cytophagia
- Bacteroidetes;Flavobacteriia
- Chloroflexi;Anaerolineae
- Chloroflexi;Caldilineae
- Chloroflexi;Chloroflexi\_unclassified
- Chloroflexi;SAR202\_clade
- Cyanobacteria;Cyanobacteria
- Firmicutes;Clostridia
- Gemmatimonadetes;Gemmatimonadetes
- Nitrospirae;Nitrospira
- PAUC34f;PAUC34f\_unclassified
- Proteobacteria;Alphaproteobacteria
- Proteobacteria;Betaproteobacteria
- Proteobacteria;Deltaproteobacteria
- Proteobacteria;Gammaproteobacteria
- Proteobacteria;Proteobacteria\_unclassified
- Thaumarchaeota;Marine\_Group\_I

### Greengenes

- Acidobacteria;Acidobacteria-6
- Acidobacteria;Sva0725
- Actinobacteria;Acidimicrobiia
- Bacteroidetes;Flavobacteriia
- Chloroflexi;Anaerolineae
- Chloroflexi;SAR202
- Crenarchaeota;Thaumarchaeota
- Cyanobacteria;Synechococophycideae
- Gemmatimonadetes;Gemm-2
- Nitrospirae;Nitrospira
- PAUC34f;PAUC34f\_unclassified
- Proteobacteria;Alphaproteobacteria
- Proteobacteria;Betaproteobacteria
- Proteobacteria;Deltaproteobacteria
- Proteobacteria;Gammaproteobacteria
- Proteobacteria;Proteobacteria\_unclassified

### RDP

- Acidobacteria;Acidobacteria\_Gp10
- Acidobacteria;others
- Actinobacteria;Actinobacteria
- Bacteroidetes;Flavobacteriia
- Cyanobacteria;Chloroplast;Cyanobacteria
- Poribacteria;Poribacteria\_unclassified
- Proteobacteria;Alphaproteobacteria
- Proteobacteria;Gammaproteobacteria
- Proteobacteria;Proteobacteria\_unclassified
- Thaumarchaeota;Nitrosopumilales



## Search results

taxonomy: k\_Bacteria;p\_Proteobacteria;c\_Alphaproteobacteria;o\_Rhizobiales

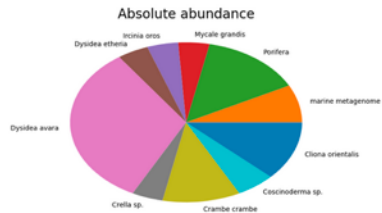
sequence:

TACGAAGGGGGCTAGCGTTGTTCCGAATCACTGGGCGTAAAGCGCACGTAGGCGGACTTTTAAGTCAGGGGTGAAATCCCGGGGCTCAACCCCGGAACTG

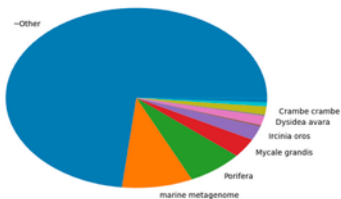
[More info from dbBact](#)

Present in 0.034474 of samples (132 / 3829)

▼ host\_scientific\_name (6 significant)



Sample number distribution



Significant enrichment:

host\_scientific\_name:Dysidea avara (30/64)  
 host\_scientific\_name:Cella sp. (4/9)  
 host\_scientific\_name:Dysidea etheria (4/10)  
 host\_scientific\_name:Cliona orientalis (11/31)  
 host\_scientific\_name:Coccinoderma sp. (5/27)  
 host\_scientific\_name:Crambe crambe (10/56)

► env\_feature (1 significant)


► country (3 significant)

► ALL (84 significant)

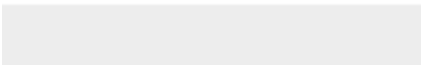



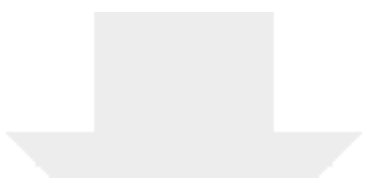
Click here to access/download  
**Supplementary Material**  
Supp.table1.tsv



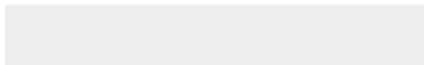



Click here to access/download  
**Supplementary Material**  
Data information.v2.docx





Click here to access/download  
**Supplementary Material**  
README.txt





Click here to access/download  
**Supplementary Material**  
[SMP.sequence.processing.html](#)





Click here to access/download  
**Supplementary Material**  
sample.metadata.csv

