

1 The Reconstruction of 2,631 Draft Metagenome-Assembled Genomes from the Global 2 Oceans

3 Authors

4 Benjamin J. Tully^{1*}, Elaina D. Graham², John F. Heidelberg^{1,2}

5
6
7 1. Center for Dark Energy Biosphere Investigations, University of Southern California, Los
8 Angeles, CA, USA

9 2. Department of Biological Sciences, University of Southern California, Los Angeles, CA,
10 USA

11 *corresponding author: tully.bj@gmail.com

12 Abstract

13
14 Microorganisms play a crucial role in mediating global biogeochemical cycles in the marine
15 environment. By reconstructing the genomes of environmental organisms through
16 metagenomics, researchers are able to study the metabolic potential of Bacteria and Archaea that
17 are resistant to isolation in the laboratory. Utilizing the large metagenomic dataset generated
18 from 234 samples collected during the *Tara Oceans* circumnavigation expedition, we were able
19 to assemble 102 billion paired-end reads into 562 million contigs, which in turn were co-
20 assembled and consolidated into 7.2 million contigs ≥ 2 kb in length. Approximately 1 million of
21 these contigs were binned to reconstruct draft genomes. In total, 2,631 draft genomes with an
22 estimated completion of $\geq 50\%$ were generated (1,491 draft genomes $> 70\%$ complete; 603 high-
23 quality genomes $> 90\%$ complete). A majority of the draft genomes were manually assigned
24 phylogeny based on sets of concatenated phylogenetic marker genes and/or 16S rRNA gene
25 sequences. The draft genomes are now publically available for the research community at-large.
26

27 Background & Summary

28 The global oceans are a vast environment in which many key biogeochemical cycles are
29 performed by microorganisms, specifically the Bacteria and Archaea^{1,2}. Assessing the role of
30 individual microorganisms has been confounded due to limitations in growing and maintaining
31 ‘wild’ organisms in the laboratory environment³. The advent of “-omic” techniques,
32 metagenomics, metatranscriptomics, metaproteomics, and metabolomics, has provided an avenue
33 for exploring microbial diversity and function by skipping the necessity of culturing organisms,
34 thus allowing researchers to study organisms for which growth conditions cannot be replicated.
35 Specifically, the application of metagenomics, the sampling and sequencing of genetic material
36 directly from environment, provides an avenue for reconstructing the genomic sequences of
37 environmental Bacteria and Archaea⁴⁻⁶.

38 Through the *Tara Oceans* Expedition (2003-2010), thousands of samples were collected
39 of marine life⁷, including more than 200 metagenomic samples targeting the viral and microbial
40 components of the marine ecosystem from around the globe^{8,9}. Several studies have started the
41 process of reconstructing microbial genomes from these metagenomics samples, utilizing
42 samples from the Mediterranean¹⁰ and the bacterial size fraction (0.2-3 μ m)¹¹. Here, we present
43 $> 2,000$ additional draft genomes from the Bacteria and Archaea estimated to be $> 50\%$ complete
44 reconstructed from 102 billion metagenomic sequences generated from multiple size fractions
45 and depths at the 61 stations sampled during the *Tara Oceans* circumnavigation of the globe.
46 Phylogenomic analysis suggests that this set of draft genomes includes highly sought after

47 genomes that lack cultured representatives, such as: Group II (149) and Group III (12)
48 Euryarchaeota, the Candidate Phyla Radiation (30), the SAR324 (18), the *Pelagibacteraceae*
49 (32), and the *Marinimicrobia* (111).

50 We envision that these draft genomes will provide a resource for downstream analysis
51 acting as references for metatranscriptomic¹² and metaproteomic¹³ projects, providing the data
52 necessary for large-scale comparative genomics within globally vital phylogenetic groups¹⁴, and
53 allowing for the exploration of novel microbial metabolisms¹⁵. Non-redundant draft
54 metagenome-assembled genomes have been deposited into the National Center for
55 Biotechnology Information (NCBI) database, along with publically accessible datasets for
56 examining metagenomic information that was not incorporated in to the draft genomes.

57 58 **Methods**

59 These methods have been described in part previously¹⁵, but have not been applied to full dataset
60 discussed below.

61 **Gathering Metagenomics Sequences & Assembly**

62 An example of the methodology used to assemble the *Tara Oceans* metagenomes is
63 available on Protocols.io (dx.doi.org/10.17504/protocols.io.hfqb3mw). All metagenomic
64 sequences generated for 234 samples collected from 61 stations during the *Tara Oceans*
65 expedition were accessed from the European Molecular Biology Laboratory-European
66 Bioinformatics Institute (EMBL-EBI)^{8,9}. Generally, samples were collected from multiple size
67 fractions, commonly ‘viral’ (<0.22 μ m), ‘girus’ (0.22-0.8 μ m), ‘bacterial’ (0.22-1.6 μ m), and
68 ‘protistan’ (0.8-5.0 μ m), at multiple depths, commonly at the surface (~5-m), deep chlorophyll
69 maximum (DCM), and mesopelagic, from each station. Each sample was assembled individually
70 using Megahit¹⁶ (v.1.0.3; parameters: --preset meta-sensitive). In total, over 102 billion paired-
71 end reads were assembled into >562 million contigs (Table 1; referred to as primary contigs).
72 Primary contigs <2kb in length were not used in downstream analysis. Primary contigs \geq 2kb in
73 length were processed using CD-HIT-EST¹⁷ (v4.6; parameter: -c 0.99) to reduce the
74 computational load required for the secondary assembly by combining contigs with \geq 99% semi-
75 global identity. Primary contigs for stations from the same oceanographic province were co-
76 assembled using Minimus2¹⁸ (Figure 1; AMOS v3.1.0; parameters: -D OVERLAP=100
77 MINID=95). Combining the Minimus2 generated contigs and the primary contigs that did not
78 assemble with Minimus2, approximately 7.2 million contigs were generated for downstream
79 analysis (Table 2; referred to as secondary contigs).

80 **Binning**

81 An example of methodology used to bin the *Tara Oceans* metagenomes is available on
82 Protocols.io (dx.doi.org/10.17504/protocols.io.iwgcfbw). Metagenomic reads from each sample
83 in a province were recruited against the set of secondary contigs generated for that province
84 using Bowtie2¹⁹ (v4.1.2; default parameters). Utilizing the BinSanity²⁰ workflow²¹ (BinSanity-
85 profile), a reads·bp⁻¹ coverage value was generated for each contig and coverage values were
86 multiplied by 100 and log normalized (parameter: --transform scale). Then due to computational
87 limitations imposed during the BinSanity binning method, the secondary contigs from each
88 province were size selected (4-14kb cutoffs) to choose approximately 100,000 contigs for
89 binning (Table 2). Approximately 6 million secondary contigs remain un-binned and are
90 available for analysis. The binning using BinSanity was performed iteratively six times, with
91 changes to the preference value after the first three iterations and a set parameter for iterations 4-
92 6 in order to influence the degree of clustering (v0.2.5.5; parameters : -p [(1) -10, (2) -5, (3) -3,

93 (4-6) -3] -m 4000 -v 400 -d 0.95 -kmer 4). Bins generated during the first five iterations were
94 processed with the BinSanity-refinement script utilizing a set preference value (parameter: -p -
95 25). After the six iteration, bins with high contamination (>10% contamination; see below) were
96 processed two more times with BinSanity-refinement using variable preference values
97 (parameter: -p [(6) -10, (7) -3]). After each refinement step, bins were assessed using CheckM²²
98 (v1.0.3; parameters: lineage_wf) for completion and contamination estimates, which were used
99 as cutoffs for inclusion in the final dataset. Bins were reassigned as a draft genome if: >90%
100 complete with <10% contamination, 80-90% complete with <5% contamination, or 50-80%
101 complete with <2% contamination. Bins that did not meet these criteria were combined for the
102 next iteration of binning, except after the six iteration (see above). In total, 2,631 draft genomes
103 were generated, with 1,491 of the genomes >70% complete, and 420 genomes meeting a high-
104 quality threshold of >90% complete and <5% contamination (Table 3). Genomes were provided
105 identifiers with the format *Tara Oceans* Binned Genome (TOBG) – Province Abbreviation –
106 Numeric ID (e.g., TOBG_NAT-221).

107 An additional 15,557 bins were generated containing at least five contigs that did not
108 meet the criteria for reclassification as a draft genome. These bins may offer pertinent
109 information for different downstream analyses. Bins of interest with high completion and high
110 contamination can be manually assessed using tools, such as Anvi'o²³, to generate a more
111 accurate draft genome. For bins with <50% completion, it may be possible to combine two or
112 more bins to generate a draft genome. And for bins with minimal or no phylogenetic markers
113 assessment may reveal that they represent viral, episomal, or eukaryotic DNA sequences.

114 **Phylogenetic Assignment**

115 A multi-pronged approach was used to provide a phylogenetic assignment to all of the
116 draft genomes. All of the secondary contigs had putative coding DNA sequences (CDSs)
117 predicted using Prodigal²⁴ (v2.6.2; -m -p meta). Contigs assigned to draft genomes and 7,041
118 complete and partial reference genomes (Supplemental Table 1) accessed from NCBI GenBank²⁵
119 and searched for phylogenetic markers. Protein phylogenetic markers were detected using hidden
120 Markov models (HMMs) collected from the Pfam database²⁶ (Accessed March 2017) and
121 identified using HMMER²⁷ (v3.1b2; parameters: hmmsearch -E 1e-10). Two sets of single-copy
122 markers recalcitrant to horizontal gene transfer were identified and used to construct
123 phylogenetic trees; a set of 16 generally syntenic markers identified in Hug²⁸, *et al.* (2016) and
124 an alternative set of 25 markers (Supplemental Table 2). Draft and reference genomes were
125 required to possess ≥ 10 and ≥ 15 markers for the Hug, *et al.* and alternative marker sets,
126 respectively, to be included in downstream analysis. If multiple copies of the same marker were
127 detected, neither copy was considered for further analysis. Each marker was aligned using
128 MUSCLE²⁹ (v3.8.31; parameter: -maxiters 8), trimmed using trimAL³⁰ (v.1.2rev59; parameter: -
129 automated1), and manually assessed. Alignments for each set of markers were concatenated. A
130 maximum likelihood tree using the LGGAMMA model was generated using FastTree³¹
131 (v.2.1.10; parameters: -lg -gamma; Supplemental Information 1 and 2). Phylogenies were
132 determined manually for 2,009 and 95 draft genomes for the Hug, *et al.* and alternative marker
133 sets, respectively (Table 4). A simplified phylogenetic tree of the Hug, *et al.* phylogenetic marker
134 set was constructed using the same parameters with only the alignments of the draft genomes for
135 Fig. 2.

136 16S rRNA genes were predicted from draft genomes using RNAmmer³² (v1.2;
137 parameters: -S bac -m ssu). 276 16S rRNA genes were detected and aligned using the SINA web
138 portal aligner³³ (<https://www.arb-silva.de/aligner/>). Aligned 16S rRNA gene sequences were

139 added to the non-redundant 16S rRNA gene database (SSURef128 NR99) in ARB³⁴ (v6.0.3)
140 using the Parsimony (Quick) tool (default parameters). Each 16S rRNA gene sequence from a
141 draft genome was assigned a putative phylogeny based on placement on the SSURef128 NR99
142 guide tree (Table 4).

143 For the draft genomes, 81.3% were manually assigned a phylogeny based on the Hug, *et*
144 *al.* marker set (2,009 draft genomes), the alternative marker set (95 draft genomes), or the 16S
145 rRNA gene tree (35 draft genomes). The remaining 492 draft genomes were provided a putative
146 phylogeny based on CheckM (Table 4).

147 **Relative Abundance**

148 Several of the size fractions used to reconstruct bacterial and archaeal draft genomes were
149 specifically designed to target different biological entities, such as double-stranded DNA viruses,
150 giant viruses (girus), and protists. In order to estimate the relative abundance of the draft
151 genomes compared to only the total bacterial and archaeal community, a set of 100 previously
152 identified HMMs for predominantly single-copy bacterial and archaeal markers^{35,36} were
153 searched against the putative CDS of the secondary contigs from each province using HMMER
154 (parameters: `hmmsearch --cut_tc`). From each province, the set of CDS identified by the marker
155 HMMs could be used to approximate the total bacterial and archaeal community. Markers
156 belonging to the draft genomes were identified. Based on the metagenomic reads recruited to the
157 secondary contigs for each sample, the number of reads aligned to each marker in a sample was
158 determined using BEDTools³⁷ (v2.17.0; multicov default parameters). A length-normalized
159 estimate of relative abundance for each draft genome in each sample in a province was
160 determined using the following equation:

$$161 \frac{\sum \text{Reads bp}^{-1} \text{ TOBG markers}}{\sum \text{Reads bp}^{-1} \text{ all province markers}} \times 100$$

162 The relative abundance estimates of draft genomes indicate that the genomes generated
163 for this study constitute only a small percentage of the total bacterial and archaeal abundance in
164 each sample (Table 5; Figure 3). The draft genomes account for a higher percentage of the viral
165 size fraction compared to other size fractions, accounting for ~60% of the total bacterial and
166 archaeal community in that size fraction. This is likely due to the fact that the number of
167 microbial organisms capable of passing through a 0.22 μ m filter is limited and the overall
168 microbial community in these samples is less complex. On average, the draft genomes in the
169 girus, bacterial, and protistan size fractions account for 14-19% of the total bacterial and archaeal
170 communities. As such, the application of alternative binning methods to this same dataset should
171 generate additional draft genomes³⁸.

173 **Data Records**

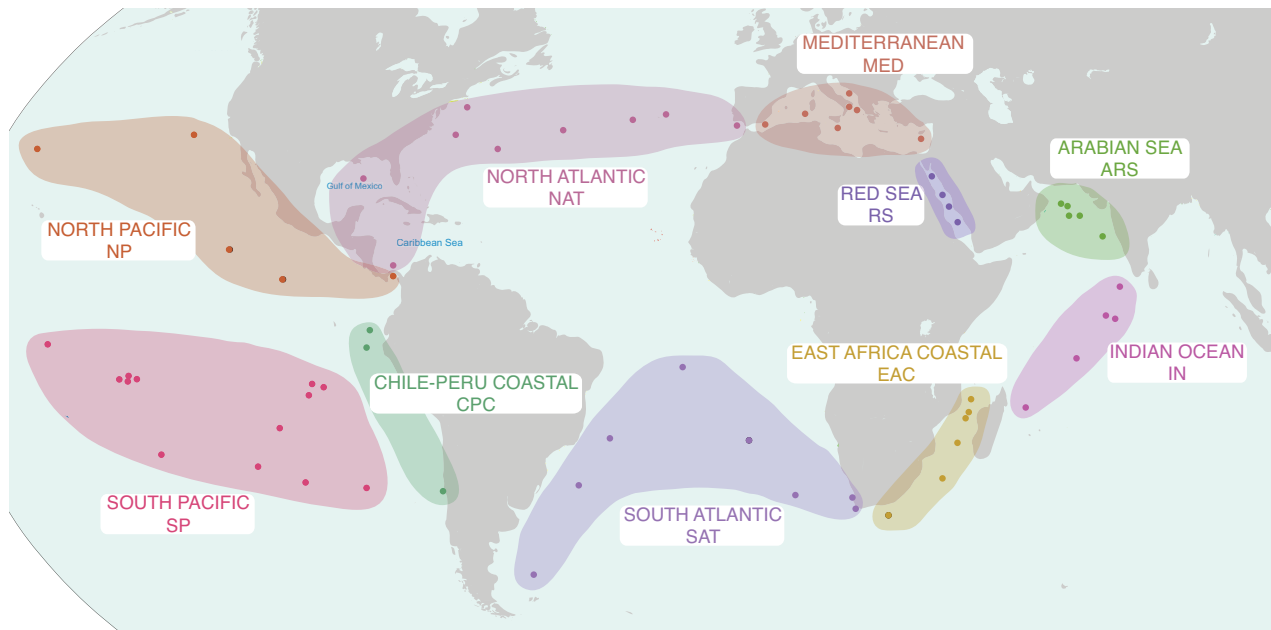
174 This project has been deposited at DDBJ/ENA/GenBank under the BioProject accession no.
175 PRJNA391943 and drafts of genomes are available with accession no. XXX-XXX [submission
176 to NCBI is ongoing – draft genomes can be found on provided figshare link] (Data Citation 1).
177 Additional data is available through figshare, including all draft genomes, all secondary contigs,
178 read count data for each secondary contig from each sample (Data Citation 2). The set of 100
179 HMMs for predominantly single-copy bacterial and archaeal markers from Albertsen, *et al.*
180 (2013) is available on GitHub (Data Citation 3).

182 **Data Usage**

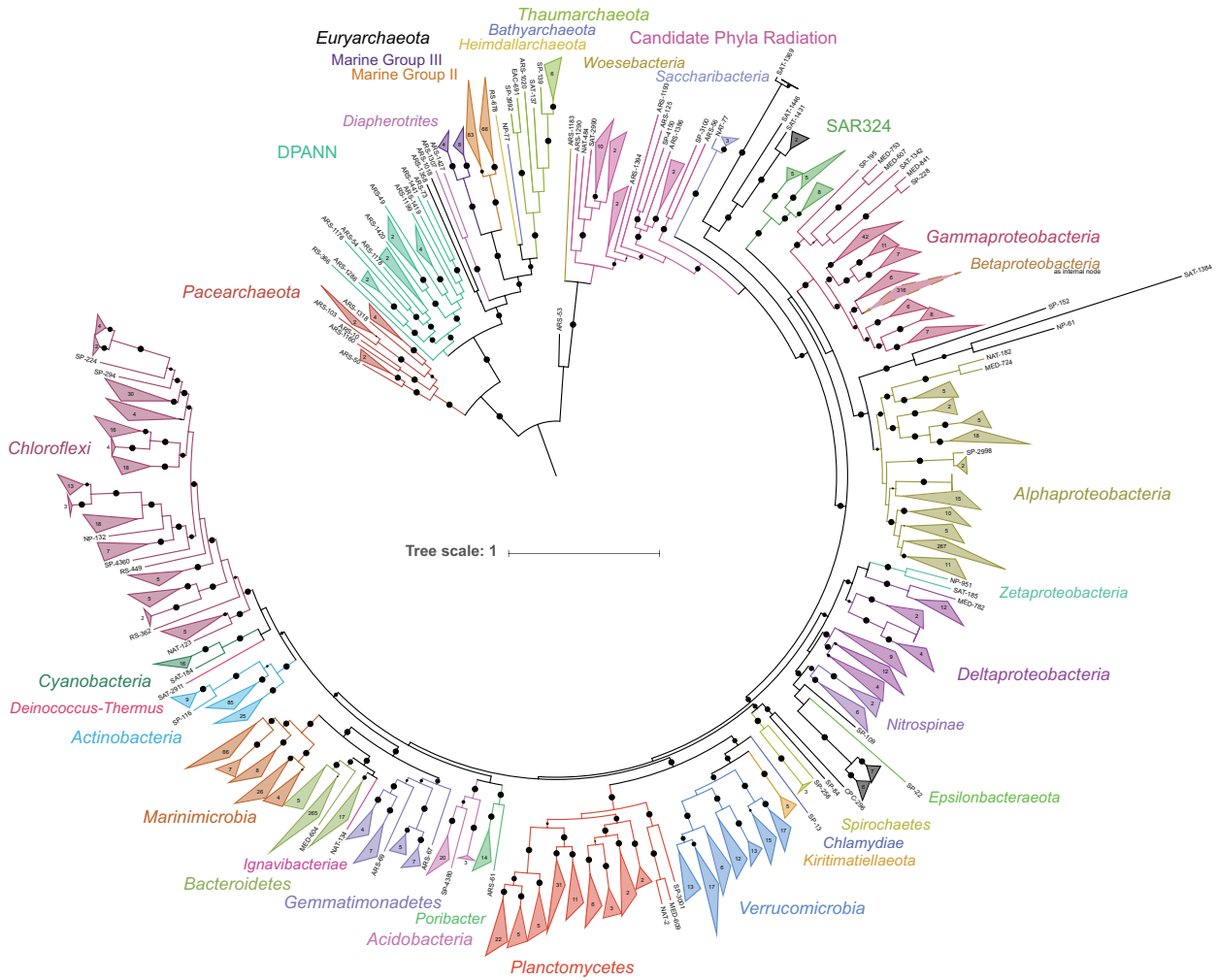
183 Due to the draft nature of the TOBG genomes, all downstream research should
184 independently assess the accuracy of genes, contigs, and phylogenetic assignments for organisms
185 of interest. Several of the draft genomes generated through this methodology appear to be
186 identical, based on the Hug marker set phylogenomic tree, to genomes generated by Tully, *et al.*
187 (2017) and Delmont, *et al.* (2017), these genomes have been identified (Table 3) and in most
188 cases duplicate genomes were not submitted to NCBI. In total, 186 draft genomes from this
189 dataset, 68 from Tully, *et al.* (2017) and 118 from Delmont, *et al.* (2017), were determined to be
190 identical to the previous work and not submitted to NCBI. However, draft genomes from this
191 study that were estimated to be more complete than available through Delmont, *et al.* (2017)
192 were submitted (n = 198) to NCBI. In providing official nomenclature for submission to NCBI,
193 priority was given to the Hug marker assignment, followed by the 16S rRNA assignment, then
194 alternative marker assignment, and, finally, the CheckM assignment.

195
196
197

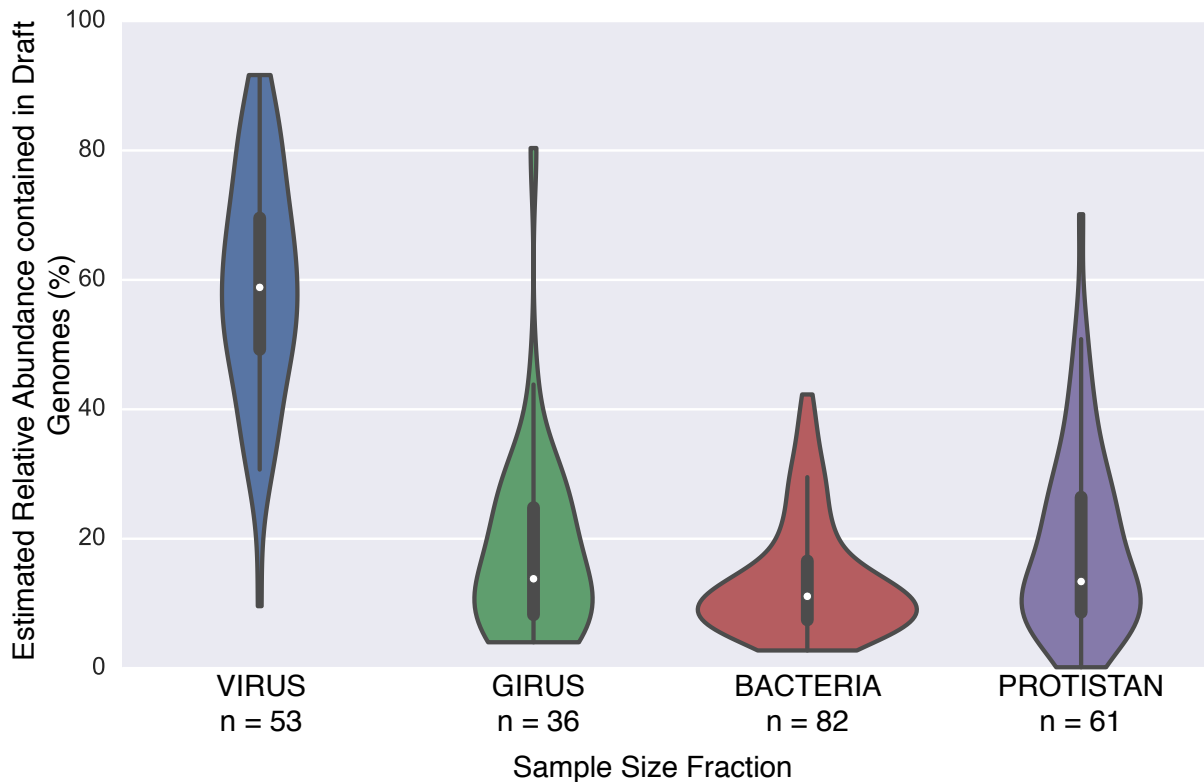
Figures



198
199 Figure 1.
200



201
202 Figure 2.
203



204
205 Figure 3.

206
207 **Figure Legends**

208 **Figure 1.** A map depicting the approximate locations of the *Tara Oceans* sampling stations from
209 which metagenomics data was collected. Stations are grouped in to larger provinces based on
210 Longhurst Provinces and site proximity. Province abbreviations are used for draft genome IDs.
211 The map in Figure 1 were modified under a CC BY-SA 3.0 license from ‘Oceans and Seas
212 boundaries map’ by Pinpin.

213 **Figure 2.** A maximum likelihood tree of the TOBG draft genomes based on 16 concatenated
214 single-copy phylogenetic markers. Bootstrap values >0.75 are shown. Circle size representing
215 the bootstrap value is scaled from 0.75-1.0. Nodes where the average branch length distance is
216 <0.5 were collapsed and the number of draft genomes in each node are provided. The image was
217 generated using the Interactive Tree of Life (iTOL; <http://itol.embl.de/>).

218 **Figure 3.** Violin plots illustrating the fraction of the estimated total bacterial and archaeal
219 community represented by the draft genomes for samples from the different size fractions.

220
221 **Tables**

222 **Table 1.** Statistics for the primary contigs generated for each of the 234 sample fractions. (Due
223 to size limitations in a bioRxiv submission, this table is included on the figshare page)

224 **Table 2.** Statistics for each province on the number secondary contigs generated, the number of
225 contigs binned and corresponding length cutoff, and the number of draft genomes reconstructed.

Province	No. of Secondary Contigs	Size Cutoff (kb)	No. of Binned Contigs	No. of Draft Genomes
Mediterranean	660,937	7.5	95,506	360

Red Sea	328,325	5.0	84,936	180
Arabian Sea	525,636	6.0	99,649	194
Indian Monsoon	285,238	4.0	93,760	72
East Africa Coastal Current	613,778	7.0	91,053	208
South Atlantic	1,373,173	11.5	96,972	360
Chile Peru Coastal	857,548	5.5	95,557	146
South Pacific	807,193	14.0	104,598	536
North Pacific	943,809	7.0	96,396	254
North Atlantic	804,316	8.5	104,848	321
SUM	7,199,953	-	963,275	2,631

226

227 **Table 3.** Statistics for each of the 2,631 draft genomes, including completion and contamination.
228 (Due to size limitations in a bioRxiv submission, this table is included on the figshare page)

229 **Table 4.** Phylogenetic assignment for each of the draft genomes as determined by the four
230 methodologies outlined in the manuscript (Assignments for the Hug *et al.* marker gene set,
231 alternative marker gene set, 16S rRNA gene, and CheckM). (Due to size limitations in a bioRxiv
232 submission, this table is included on the figshare page)

233 **Table 5.** Estimated relative abundance value for all draft genomes in each sample fraction from
234 each province. (Due to size limitations in a bioRxiv submission, this table is included on the
235 figshare page)

236

237 **References**

- 238 1. Moran, M. A. The global ocean microbiome. *Science* **350**, aac8455–aac8455 (2015).
- 239 2. Falkowski, P. G., Fenchel, T. & DeLong, E. F. The Microbial Engines That Drive Earth's
240 Biogeochemical Cycles. *Science* **320**, 1034–1039 (2008).
- 241 3. Staley, J. T. & Konopka, A. Measurement of in situ activities of nonphotosynthetic
242 microorganisms in aquatic and terrestrial habitats. *Annual Reviews in Microbiology* 321–
243 346 (1985).
- 244 4. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of
245 microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
- 246 5. Seitz, K. W., Lazar, C. S., Hinrichs, K.-U., Teske, A. P. & Baker, B. J. Genomic
247 reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for
248 acetogenesis and sulfur reduction. *ISME J* 1–10 (2016). doi:10.1038/ismej.2015.233
- 249 6. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected
250 biogeochemical processes in an aquifer system. *Nature Communications* **7**, 13219 (2016).
- 251 7. Karsenti, E. *et al.* A Holistic Approach to Marine Eco-Systems Biology. *Plos Biol* **9**,
252 e1001177–5 (2011).
- 253 8. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans
254 data. *Sci. Data* **2**, 150023–16 (2015).
- 255 9. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean
256 microbiome. *Science* **348**, 1261359–1261359 (2015).
- 257 10. Tully, B. J., Sachdeva, R., Graham, E. D. & Heidelberg, J. F. 290 metagenome-assembled
258 genomes from the Mediterranean Sea: a resource for marine microbiology. *PeerJ* **5**,

- 259 e3558–15 (2017).
- 260 11. Delmont, T. O. *et al.* Nitrogen-Fixing Populations Of Planctomycetes And Proteobacteria
261 Are Abundant In The Surface Ocean. *bioRxiv* 1–16 (2017). doi:10.1101/129791
- 262 12. Gifford, S. M., Sharma, S., Booth, M. & Moran, M. A. Expression patterns reveal niche
263 diversification in a marine microbial assemblage. **7**, 281–298 (2012).
- 264 13. Saito, M. A. *et al.* Multiple nutrient stresses at intersecting Pacific Ocean biomes detected
265 by protein biomarkers. *Science* **345**, 1173–1177 (2014).
- 266 14. Farrant, G. K. *et al.* Delineating ecologically significant taxonomic units from global
267 patterns of marine picocyanobacteria. *Proc. Natl. Acad. Sci. U.S.A.* 201524865–10 (2016).
268 doi:10.1073/pnas.1524865113
- 269 15. Graham, E. D., Heidelberg, J. F. & Tully, B. Undocumented Potential For Primary
270 Productivity In A Globally-Distributed Bacterial Photoautotroph. *bioRxiv* 1–17 (2017).
271 doi:10.1101/140715
- 272 16. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by
273 advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
- 274 17. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-
275 generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
- 276 18. Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S. & Pop, M. Next generation
277 sequence assembly with AMOS. *Curr Protoc Bioinformatics* **Chapter 11**, Unit 11.8
278 (2011).
- 279 19. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9**,
280 357–359 (2012).
- 281 20. Graham, E. D., Heidelberg, J. F. & Tully, B. J. BinSanity: unsupervised clustering of
282 environmental microbial assemblies using coverage and affinity propagation. *PeerJ* **5**,
283 e3035–19 (2017).
- 284 21. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for
285 assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
- 286 22. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM:
287 assessing the quality of microbial genomes recovered from isolates, single cells, and
288 metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
- 289 23. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics
290 data. *PeerJ* **3**, e1319 (2015).
- 291 24. Hyatt, D., LoCascio, P. F., Hauser, L. J. & Uberbacher, E. C. Gene and translation
292 initiation site prediction in metagenomic sequences. *Bioinformatics* **28**, 2223–2230
293 (2012).
- 294 25. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **28**, 15–18 (2000).
- 295 26. Bateman, A. *et al.* The Pfam Protein Families Database. *Nucleic Acids Res.* **30**, 276–280
296 (2002).
- 297 27. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence
298 similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
- 299 28. Hug, L. A. *et al.* A new view of the tree of life. *Nature Microbiology* **1**, 16048 (2016).
- 300 29. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high
301 throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
- 302 30. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated
303 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973
304 (2009).

- 305 31. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood
306 trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
- 307 32. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes.
308 *Nucleic Acids Res.* **35**, 3100–3108 (2007).
- 309 33. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: accurate high-throughput multiple
310 sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829 (2012).
- 311 34. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**,
312 1363–1371 (2004).
- 313 35. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by
314 differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**, 533–538
315 (2013).
- 316 36. Tully, B. J. & Heidelberg, J. F. Potential Mechanisms for Microbial Energy Acquisition in
317 Oxic Deep-Sea Sediments. *Appl. Environ. Microbiol.* **82**, 4232–4243 (2016).
- 318 37. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
319 features. *Bioinformatics* **26**, 841–842 (2010).
- 320 38. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication,
321 aggregation, and scoring strategy. *bioRxiv* 1–24 (2017). doi:10.1101/107789
322

323 Data Citations

- 324 1. Tully, B. J. *NCBI BioProject* PRJNA391943 (2017)
- 325 2. Tully, B.J. *figshare* <http://dx.doi.org/10.6084/m9.figshare.5188273> (2017)
- 326 3. Albertsen, M. *GitHub* [https://github.com/MadsAlbertsen/multi-](https://github.com/MadsAlbertsen/multi-metagenome/blob/master/R.data.generation/essential.hmm)
327 [metagenome/blob/master/R.data.generation/essential.hmm](https://github.com/MadsAlbertsen/multi-metagenome/blob/master/R.data.generation/essential.hmm) (2011)
328

329 Author Contribution

330 BJT conceived and designed the methodology, performed the analysis, wrote the paper, and
331 prepared the figure and tables. EDG performed the analysis and reviewed drafts of the paper.
332 JHF provided funding and resources to perform the analysis and reviewed drafts of the paper.
333

334 Acknowledgements

335 Funding was provided by the Center for Dark Energy Biosphere Investigations (C-DEBI) to BJT
336 and JFH (OCE-0939654). As we have stated before, this project would have not been possible if
337 not for the diligent commitment by the *Tara Oceans* consortium to allow for the open access of
338 the data collected during the expedition. We only hope that this small dataset can be used by the
339 scientific community at-large to increase the impact of this transformational research project.
340 This is C-DEBI Contribution XXX.