

1 290 Metagenome-assembled Genomes from the Mediterranean Sea: a resource for marine 2 microbiology

3 4 Authors

5 Benjamin J. Tully^{1‡}, Rohan Sachdeva², Elaina D. Graham², and John F. Heidelberg^{1,2}

6
7 ‡ - corresponding author, tully.bj@gmail.com

8 1 – Center for Dark Energy Biosphere Investigations, University of Southern California, Los
9 Angeles, CA 90089

10 2 – Department of Biological Sciences, University of Southern California, Los Angeles, CA
11 90089

12

13 Abstract

14 The *Tara Oceans* Expedition has provided large, publicly-accessible microbial metagenomic
15 datasets from a circumnavigation of the globe. Utilizing several size fractions from the samples
16 originating in the Mediterranean Sea, we have used current assembly and binning techniques to
17 reconstruct 290 putative high-quality metagenome-assembled bacterial and archaeal genomes,
18 with an estimated completion of $\geq 50\%$, and an additional 2,786 bins, with estimated completion
19 of 0-50%. We have submitted our results, including initial taxonomic and phylogenetic
20 assignments, for the putative high-quality genomes to open-access repositories for the scientific
21 community to use in ongoing research.

22

23 Introduction

24 Microorganisms are a major constituent of the biology within the world's oceans and act
25 as the important linchpins in all major global biogeochemical cycles¹. Marine microbiology is
26 among the disciplines at the forefront of advancements in understanding how microorganisms
27 respond to and impact the local and large-scale environments. An estimated 10^{29} Bacteria and
28 Archaea² reside in the oceans and represent an immense amount of poorly constrained, and ever
29 evolving genetic diversity.

30 The *Tara Oceans* Expedition (2003-2010) encompassed a major endeavor to add to the
31 body of knowledge collected during previous global ocean surveys to sample the genetic
32 potential of microorganisms³. To accomplish this goal, *Tara Oceans* sampled planktonic
33 organisms (viruses to fish larvae) at two major depths, the surface ocean and the mesopelagic.
34 The amount of data collected was expansive and included 35,000 samples from 210 ecosystems³.
35 The *Tara Oceans* Expedition generated and publically released 7.2 Tbp of metagenomic data
36 from 243 ocean samples from throughout the global ocean, specifically targeting the smallest
37 members of the ocean biosphere, the viruses, Bacteria and Archaea, and picoeukaryotes⁴. Initial
38 work on these fractions produced a large protein database, totaling >40 million nonredundant
39 protein sequences and identified >35,000 microbial operational taxonomic units (OTUs)⁴.

40 Leveraging the publically available metagenomic sequences from the “girus” (giant virus;
41 0.22-1.6 μm), “bacteria” (0.22-1.6 μm), and “protist” (0.8-5 μm) size fractions, we have
42 performed a new joint assembly of these samples using current sequence assemblers (Megahit⁵)
43 and methods (combining assemblies from multiple sites using Minimus2⁶). These metagenomic
44 assemblies were binned using BinSanity⁷ in to 290 high-quality (low contamination) microbial
45 genomes, ranging from 50-100% estimated completion. Environmentally derived genomes are
46 imperative for a number of downstream applications, including comparative genomes,

47 metatranscriptomics, and metaproteomics. This series of genomic data can allow for the
48 recruitment of environmental “-omic” data and provide linkages between functions and
49 phylogenies. This method was initially performed on the seven sites from the Mediterranean Sea
50 containing microbial metagenomic samples (TARA007, -009, -018, -023, -025 and -030), but
51 will continue through the various Longhurst provinces⁸ sampled during the *Tara Oceans* project
52 (Figure 1). All of the assembly data is publically available, including the initial Megahit
53 assemblies for each site from the various size fractions and depths and putative (minimal quality
54 control) genomes.

55

56 **Materials and Methods**

57

58 A generalized version of the following workflow is presented in Figure 2.

59

60 *Sequence Retrieval and Assembly*

61 All sequences for the reverse and forward reads from each sampled site and depth within
62 the Mediterranean Sea were accessed from European Molecular Biology Laboratory (EMBL)
63 utilizing their FTP service (Table 1). Paired-end reads from different filter sizes from each site
64 and depth (e.g., TARA0007, girus filter fraction, sampled at the deep chlorophyll maximum)
65 were assembled using Megahit⁵ (v1.0.3; parameters: --preset, meta-sensitive). To keep consistent
66 with TARA sample nomenclature, “bacteria” or “BACT” will be used to encompass the size
67 fraction 0.22-1.6 μm . All of the Megahit assemblies were pooled in to two tranches based on
68 assembly size, $\leq 1,999\text{bp}$, and $\geq 2,000\text{bp}$. Longer assemblies ($\geq 2\text{kb}$) with $\geq 99\%$ semi-global
69 identity were combined using CD-HIT-EST (v4.6; -T 90 -M 500000 -c 0.99 -n 10). The reduced
70 set of contiguous DNA fragments (contigs) was then cross-assembled using Minimus2⁶ (AMOS
71 v3.1.0; parameters: -D OVERLAP=100 MINID=95).

72

73 *Metagenome-assembled Genomes*

74 Sequence reads were recruited against a subset of contigs ($\geq 7.5\text{kb}$) constructed during the
75 secondary assembly (Megahit + Minimus2) for each of the *Tara* samples using Bowtie2⁹ (v4.1.2;
76 default parameters). Utilizing the SAM file output, read counts for each contig were determined
77 using featureCounts¹⁰ (v1.5.0; default parameters). Coverage was determined for all contigs by
78 dividing the number of recruited reads by the length of the contig (reads/bp). Due to the low
79 coverage nature of the samples, in order to effectively delineate between contig coverage
80 patterns, the coverage values were transformed by multiplying by five (determined through
81 manual tuning). Transformed coverage values were then utilized to cluster contigs in to bins
82 utilizing BinSanity⁷ (parameters: -p -3, -m 4000, -v 400, -d 0.9). Bins were assessed for the
83 presence of putative microbial genomes using CheckM¹¹ (v1.0.3; parameters: lineage_wf). Bins
84 were split in to three categories: (1) putative high quality genomes ($\geq 50\%$ complete and $\leq 10\%$
85 cumulative redundancy [% contamination – (% contamination \times % strain heterogeneity \div 100]));
86 (2) bins with “high” contamination ($\geq 50\%$ complete and $\geq 10\%$ cumulative redundancy); and (3)
87 low completion bins ($< 50\%$ complete).

88 The high contamination bins containing approximately two genomes, three genomes, or
89 ≥ 4 genomes used the BinSanity refinement method (Binsanity-refine; -m 2000, -v 200, -d 0.9)
90 with variable preference values (-p) of -1000, -500, and -100, respectively. The resulting bins
91 were added to one of the three categories: putative high quality genomes, high contamination
92 bins, and low completion bins. The high contamination bins were processed for a third time with

93 the Binsanity-refine utilizing a preference of -100 (-p -100). These bins were given final
94 assignments to either the putative high quality genomes (some putative genomes had >10%
95 cumulative contamination, but have been designated) or low completion bins.

96 Any contigs not assigned to putative high-quality genomes were assessed using
97 BinSanity using raw coverage values. Two additional rounds of refinement were performed with
98 the first round of refinement using preference values based on the estimated number of
99 contaminating genomes (as above) and the second round using a set preference of -10 (-p -10).
100 Following this binning phase, contigs were assigned to high quality bins (e.g., *Tara*
101 **Mediterranean** genome 1, referred to as TMED1, etc.), low completion bins with at least five
102 contigs (0-50% complete; TMEDlc1, etc. lc, low completion), or were not placed in a bin
103 (Supplemental Table 1 & 2).

104 105 *Taxonomic and Phylogenetic Assignment of High Quality Genomes*

106 The bins representing the high quality genomes were assessed for taxonomy and
107 phylogeny using multiple methods to provide a quick reference for selecting genomes of interest.
108 Taxonomy as assigned using the putative placement provided via CheckM during the pplacer¹²
109 step of the analysis to the lowest taxonomic placement (parameters: tree_qa -o 2). This step was
110 also performed for all low completion bins.

111 Two separate attempts were made to assign the high quality genomes a phylogenetic
112 assignment. High quality genomes were searched for the presence of the full-length 16S rRNA
113 gene sequence using RNAmmer¹³ (v1.2; parameters: -S bac -m ssu). All full-length sequences
114 were aligned to the SILVA SSU reference database (Ref123) using the SINA web portal
115 aligner¹⁴ (<https://www.arb-silva.de/aligner/>). These alignments were loaded in to ARB¹⁵ (v6.0.3),
116 manually assessed, and added to the non-redundant 16S rRNA gene database (SSURef123
117 NR99) using ARB Parsimony (Quick) tool (parameters: default). A selection of the nearest
118 neighbors to the *Tara* genome sequences were selected and used to construct a 16S rRNA
119 phylogenetic tree. Genome-identified 16S rRNA sequences and SILVA reference sequences
120 were aligned using MUSCLE¹⁶ (v3.8.31; parameters: -maxiters 8) and processed by the
121 automated trimming program trimAL¹⁷ (v1.2rev59; parameters: -automated1). Automated
122 trimming results were assessed manually in Geneious¹⁸ (v6.1.8) and trimmed where necessary
123 (positions with >50% gaps) and re-aligned with MUSCLE (parameters: -maxiters 8). An
124 approximate maximum likelihood (ML) tree with pseudo-bootstrapping was constructed using
125 FastTree¹⁹ (v2.1.3; parameters: -nt -gtr -gamma; Figure 3).

126 High-quality genomes were assessed for the presence of the 16 ribosomal markers genes
127 used in Hug, *et al.* (2016)²⁰. Putative CDSs were determined using Prodigal (v2.6.3; parameters:
128 -m -p meta) and were searched using HMMs for each marker using HMMER²¹ (v3.1b2;
129 parameters: hmmsearch --cut_tc --notextw). If a genome had multiple copies of any single
130 marker gene, neither was considered, and only genomes with ≥ 8 markers were used to construct
131 a phylogenetic tree. Markers identified from the high quality genomes were combined with
132 markers from 1,729 reference genomes that represent the major bacterial phylogenetic groups (as
133 presented by IMG²²). Archaeal reference sequences were not included; however, none of the
134 putative archaeal environmental genomes had a sufficient number of markers for inclusion on the
135 tree. Each marker gene was aligned using MUSCLE (parameters: -maxiters 8) and automatically
136 trimmed using trimAL (parameters: -automated1). Automated trimming results were assessed (as
137 above) and re-aligned with MUSCLE, as necessary. Final alignments were concatenated and

138 used to construct an approximate ML tree with pseudo-bootstrapping with FastTree (parameters:
139 -gtr -gamma; Figure 4).

140

141 *Relative Abundance of High Quality Genomes*

142 To set-up a baseline that could approximate the “microbial” community (Bacteria, Archaea and
143 viruses) present in the various *Tara* metagenomes, which included filter sizes specifically
144 targeting both protists and giruses, reads were recruited against all contigs generated from the
145 Minimus2 and Megahit assemblies $\geq 2\text{kb}$ using Bowtie2 (default parameters). Some assumptions
146 were made that contigs $< 2\text{kb}$ would include, low abundance bacteria and archaea, bacteria and
147 archaea with high degrees of repeats/assembly poor regions, fragmented picoeukaryotic
148 genomes, and problematic read sequences (low quality, sequencing artefacts, etc.). All relative
149 abundance measures are relative to the number of reads recruited to the assemblies $\geq 2\text{kb}$. Read
150 counts were determined using featureCounts (as above). Length-normalized relative abundance
151 values were determined for each high quality genome for each sample:

$$\frac{\frac{\text{Reads}}{\text{bp}} \text{ per genome}}{\sum \frac{\text{Reads}}{\text{bp}} \text{ all genomes}} \times \frac{\sum \text{Recruited reads to genomes}}{\sum \text{Recruited reads to all contigs } (\geq 2\text{kb})} \times 100$$

152

153 *Data Availability*

154 This project has been deposited at DDBJ/ENA/GenBank under the BioProject accession
155 no. ##### and drafts of genomes are available with accession no. #####-#####. Additional files
156 have been provided and are available through FigShare
157 (<https://dx.doi.org/10.6084/m9.figshare.3545330>), such as: all contigs from Minimus2 + Megahit
158 output used for binning and community assessment; contig read counts per sample; the putative
159 genome contigs and Prodigal-predicted nucleotide and protein putative CDS FASTA files; the
160 ribosomal marker HMM profiles; reference genome markers; high quality genome markers; low
161 completion bins, and contigs without a bin. All contigs generated using Megahit from each
162 sample are available through iMicrobe (<http://data.imicrobe.us/project/view/261>).

163

164 **Results**

165 *Assembly*

166 The initial Megahit assembly was performed on the publicly available reads for *Tara*
167 stations 007, 009, 018, 023, 025, 030. Starting with 147-744 million reads per sample, the
168 Megahit assembly process generated 1.2-4.6 million assemblies with a mean N_{50} and longest
169 contig of 785bp and 537kb, respectively (Table 1). In general, the assemblies generated from the
170 *Tara* samples targeting the protist size fraction (0.8-5 μm) had a shorter N_{50} value than the
171 bacteria size fractions (mean: 554bp vs 892bp, respectively). Assemblies from the Megahit
172 assembly process were pooled and separated by length. Of the 42.6 million assemblies generated
173 during the first assembly, 1.5 million were $\geq 2\text{kb}$ in length (Table 2). Several attempts were made
174 to assemble the shorter contigs, but publicly available overlap-consensus assemblers (Newbler
175 [454 Life Sciences], cap3²³, and MIRA²⁴) failed on multiple attempts. Processing the $\geq 2\text{kb}$
176 assemblies from all of the samples through CD-HIT-EST reduced the total to 1.1 million contigs
177 $\geq 2\text{kb}$. This group of contigs was subjected to the secondary assembly through Minimus2,
178 generating 158,414 new contigs (all $\geq 2\text{kb}$). The secondary contigs were combined with the

179 Megahit contigs that were not assembled by Minimus2. This provided a contig dataset consisting
180 of 660,937 contigs, all ≥ 2 kb in length (Table 2; further referred to as data-rich-contigs).

181 182 *Binning*

183 The set of data-rich-contigs was used to recruit the metagenomic reads from each sample
184 using Bowtie2. The data-rich-contigs recruited 15-81% of the reads depending on the sample. In
185 general, the protist size fraction recruited substantially fewer reads than the girus and bacteria
186 size fractions (mean: 19.8% vs 75.0%, respectively) (Table 1). For the protist size fraction, the
187 “missing” data for these recruitments likely results from the poor assembly of more complex and
188 larger eukaryotic genomes. The fraction of the reads that do not recruit in the girus and bacterial
189 size fraction samples could be accounted for by the large number of low quality assemblies (200-
190 500bp) and reads that could not be assembled due to low abundance or high complexity (Table
191 2).

192 Unsupervised binning was performed using both transformed and raw coverage values
193 for a subset of 95,506 contigs from the data-rich-contigs that were ≥ 7.5 kb (referred to further as
194 binned-contigs). Binning using the transformed coverage data generated 237 putative high-
195 quality genomes (12 putative genomes are of slightly lower quality with $>10\%$ redundancy and
196 have been noted) containing 15,032 contigs (Supplemental Information S1). Contigs not in
197 putative genomes were re-binned based on raw coverage values, generating 53 additional
198 putative high-quality genomes encompassing 3,348 contigs. In total, 290 putative high-quality
199 genomes were generated with 50-100% completion (mean: 69%) with a mean length and number
200 of putative CDS of 1.7Mbp and 1,699, respectively (Supplemental Information S1). All other
201 contigs were grouped in to bins with at least five contigs, but with estimated completion of 0-
202 50% (2,786 low completion bins; 74,358 contigs; Supplemental Information S2) or did not bin
203 (2,732 contigs). Nearly a quarter of the low completion bins (24.7%) have an estimated
204 completion of 0%.

205 206 *Taxonomy, Phylogeny, & Potential Organisms of Interest*

207 The 290 putative high-quality genomes had a taxonomy assigned to it via CheckM during
208 the pplacer step. All of the genomes, except for 20, had an assignment to at least the Phylum
209 level, and 83% of the genomes had an assignment to at least the Class level (Supplemental
210 Information S1).

211 Phylogenetic information was determined for as many genomes as possible. Genomes
212 were assessed for the presence of full-length 16S rRNA genes. In total, 37 16S rRNA genes were
213 detected in 35 genomes. 16S rRNA genes can prove to be problematic during the assembly steps
214 due the high level of conservation that can break contigs²⁵ (Figure 3). Additionally, the
215 conserved regions of the 16S rRNA, depending on the situation, can over- or under-recruit reads,
216 resulting in coverage variations that can misplace contigs in to the incorrect genome. As such,
217 several of the 16S rRNA phylogenetic placements support the taxonomic assignments, while
218 some are contradictory. Further analysis should allow for the determination of the most
219 parsimonious result.

220 Beyond the 16S rRNA gene, genomes were searched for 16 conserved, syntenic
221 ribosomal markers. Sufficient markers (≥ 8) were identified in 193 of the genomes (67%) and
222 placed on a tree with 1,729 reference sequences (Figure 4). Phylogenies were then assigned to
223 the lowest taxonomic level that could be confidently determined. These putative results reveal a
224 number of genomes were generated that represent multiple clades for which environmental

225 genomic information remains limited, including: *Planctomycetes*, *Verrucomicrobia*,
226 *Marinimicrobia*, *Cyanobacteria*, and uncultured groups within the *Alpha*- and
227 *Gammaproteobacteria*.

228 229 *Relative Abundance*

230 A length-normalized relative abundance value was determined for each genome in each
231 sample based on the number of reads recruited to the data-rich-contigs. The relative abundance
232 for the individual genomes was determined based on this portion of the dataset (Supplemental
233 Information S3). In general, the genomes had low relative abundance (maximum relative
234 abundance = 1.9% for TMED155 a putative *Cyanobacteria* at site TARA023 from the protistan
235 size fraction sampled at the surface; Supplemental Table 1). The high-quality genomes
236 accounted for 1.57-25.16% of the approximate microbial community as determined by the data-
237 rich-contigs (mean = 13.69%), with the ten most abundant genomes in a sample representing
238 0.61-10.31% (Table 1).

239 240 **Concluding Statement**

241 The goal of this project was to provide preliminary putative genomes from the *Tara*
242 *Oceans* microbial metagenomic datasets. The 290 putative high-quality genomes and 2,786 low
243 completion bins were created using the 20 samples and six stations from the Mediterranean Sea.

244 Initial assessment of the phylogeny of these metagenomic-assembled genomes indicates
245 several new genomes from environmentally relevant organisms, including, approximately 14
246 new *Cyanobacteria* genomes within the genera *Prochlorococcus* and *Synechococcus* and 33 new
247 SAR11 genomes. Additionally, there are putative genomes from the marine *Euryarchaeota* ($n =$
248 11), *Verrucomicrobia* ($n = 17$), *Planctomycetes* ($n = 14$), and *Marinimicrobia* ($n \approx 5$).
249 Additionally, the low completion bins may house distinct viral genomes. Of particular interest
250 may be the 40 bins with 0% completion (based on single-copy marker genes), but that contain
251 >500kb of genetic material (including 3 bins with >1Mb). These large bins lacking markers may
252 be good candidates for research in to the marine “giant viruses” and episomal DNA sources
253 (plasmids, etc.).

254 It should be noted, researchers using this dataset should be aware that all of the genomes
255 generated from these samples should be used as a resource with some skepticism towards the
256 results being an absolute. Like all results for metagenome-assembled genomes, these genomes
257 represent a best-guess approximation of a taxon from the environment²⁶. Researchers are
258 encouraged to confirm all claims through various genomic analyses and accuracy may require
259 the removal of conflicting sequences.

260 261 **Acknowledgements**

262 We are indebted to the *Tara Oceans* project and team for their commitment to open-
263 access data that allows data aficionados to indulge in the data and attempt to add to the body of
264 science contained within. And we thank the Center for Dark Energy Biosphere Investigations (C-
265 DEBI) for providing funding to BJT and JFH (OCE-0939654).

266 267 **Author Contributions**

268 BJT conceived of the project, performed all of the methods and analyses, and wrote the
269 manuscript. RS provided the origins of the workflow and invaluable feedback during the
270 execution of the methods and analyses. EDG provided feedback and troubleshooting using the

271 pre-release version of BinSanity. JFH provided funding. RS and JFH contributed to manuscript
272 editing and polishing. All authors have read the submitted draft of the manuscript.

273

274 **Legends**

275

276 Table 1. Statistics for Megahit assemblies, recruitment to data-rich-contigs, and relative
277 abundance of high-quality genome results for each sample

278

279 Table 2. Assembly statistics at various steps during processing

280

281 Figure 1. Map illustrating the locations and size fractions sampled for the *Tara Oceans*
282 Mediterranean Sea datasets. Girus, ‘giant virus’ size fraction (0.22-1.6 μm). Bact, ‘bacteria’ size
283 fraction (0.22-1.6 μm). Prot, ‘protist’ size fraction (0.8-5.0 μm).

284

285 Figure 2. Workflow used to process *Tara Oceans* Mediterranean Sea metagenomic datasets.

286

287 Figure 3. FastTree approximate maximum-likelihood phylogenetic tree constructed with 37 and
288 785 16S rRNA genes from putative high-quality genomes and references, respectively.

289

290 Figure 4. Cladogram of a FastTree approximate maximum-likelihood phylogenetic tree
291 constructed using 16 syntenic, single-copy marker genes for 193 high-quality genomes and 1,729
292 reference genomes. Leaves denoting the position of the TMED genes have been indicated by
293 extending beyond the edge of the tree. Sequence alignment is available in Supplemental
294 Information S4. Phylogenetic tree newick file is available in Supplemental Information S5.

295

296 **Supplemental Information**

297

298 Supplemental Information S1. Statistics, taxonomic and phylogenetic assignments for the
299 putative high-quality genomes.

300

301 Supplemental Information S2. Statistics and CheckM taxonomy for low completion bins.

302

303 Supplemental Information S3. Relative abundance values determined for each genome based the
304 length-normalized fraction of reads recruited to the genome relative to reads recruited for the
305 data-rich-contigs.

306

307 Supplemental Information S4. Concatenated MUSCLE alignment file of 16 ribosomal marker
308 proteins used to construct Figure 4.

309

310 Supplemental Information S5. Newick file of concatenated 16 ribosomal marker proteins,
311 including FastTree determined local support values using the Shimodaira-Hasegawa test.

312

313 **References**

314

- 315 1. Falkowski, P. G., Fenchel, T. & DeLong, E. F. The Microbial Engines That Drive Earth's
316 Biogeochemical Cycles. *Science* **320**, 1034–1039 (2008).

- 317 2. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: The unseen majority. *Proc.*
318 *Natl. Acad. Sci. U.S.A.* **95**, 6578–6583 (1998).
- 319 3. Karsenti, E. *et al.* A Holistic Approach to Marine Eco-Systems Biology. *Plos Biol* **9**,
320 e1001177–5 (2011).
- 321 4. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean
322 microbiome. *Science* **348**, 1261359–1261359 (2015).
- 323 5. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by
324 advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
- 325 6. Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S. & Pop, M. Next generation
326 sequence assembly with AMOS. *Curr Protoc Bioinformatics* **Chapter 11**, Unit 11.8
327 (2011).
- 328 7. Graham, E. D., Heidelberg, J. F. & Tully, B. J. BinSanity: unsupervised clustering of
329 environmental microbial assemblies using coverage and affinity propagation. *PeerJ* **5**,
330 e3035–19 (2017).
- 331 8. Longhurst, A., Sathyendranath, S., Platt, T. & Caverhill, C. An Estimate of Global
332 Primary Production in the Ocean From Satellite Radiometer Data. *Journal of Plankton*
333 *Research* **17**, 1245–1271 (1995).
- 334 9. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9**,
335 357–359 (2012).
- 336 10. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for
337 assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
- 338 11. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM:
339 assessing the quality of microbial genomes recovered from isolates, single cells, and
340 metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
- 341 12. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood
342 and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC*
343 *Bioinformatics* **11**, 538 (2010).
- 344 13. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes.
345 *Nucleic Acids Res.* **35**, 3100–3108 (2007).
- 346 14. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: accurate high-throughput multiple
347 sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829 (2012).
- 348 15. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**,
349 1363–1371 (2004).
- 350 16. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high
351 throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
- 352 17. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated
353 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973
354 (2009).
- 355 18. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform
356 for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
- 357 19. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood
358 trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
- 359 20. Hug, L. A. *et al.* A new view of the tree of life. *Nature Microbiology* 1–6 (2016).
360 doi:10.1038/nmicrobiol.2016.48
- 361 21. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence
362 similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).

- 363 22. Markowitz, V. M. *et al.* The integrated microbial genomes (IMG) system. *Nucleic Acids*
364 *Res.* **34**, D344–8 (2006).
- 365 23. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res.* **9**,
366 868–877 (1999).
- 367 24. Chevreux, B. *et al.* Using the miraEST assembler for reliable and automated mRNA
368 transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* **14**, 1147–1159
369 (2004).
- 370 25. Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing
371 data. *Genomics* **95**, 315–327 (2010).
- 372 26. Sharon, I. & Banfield, J. F. Microbiology. Genomes from metagenomics. *Science* **342**,
373 1057–1058 (2013).
- 374

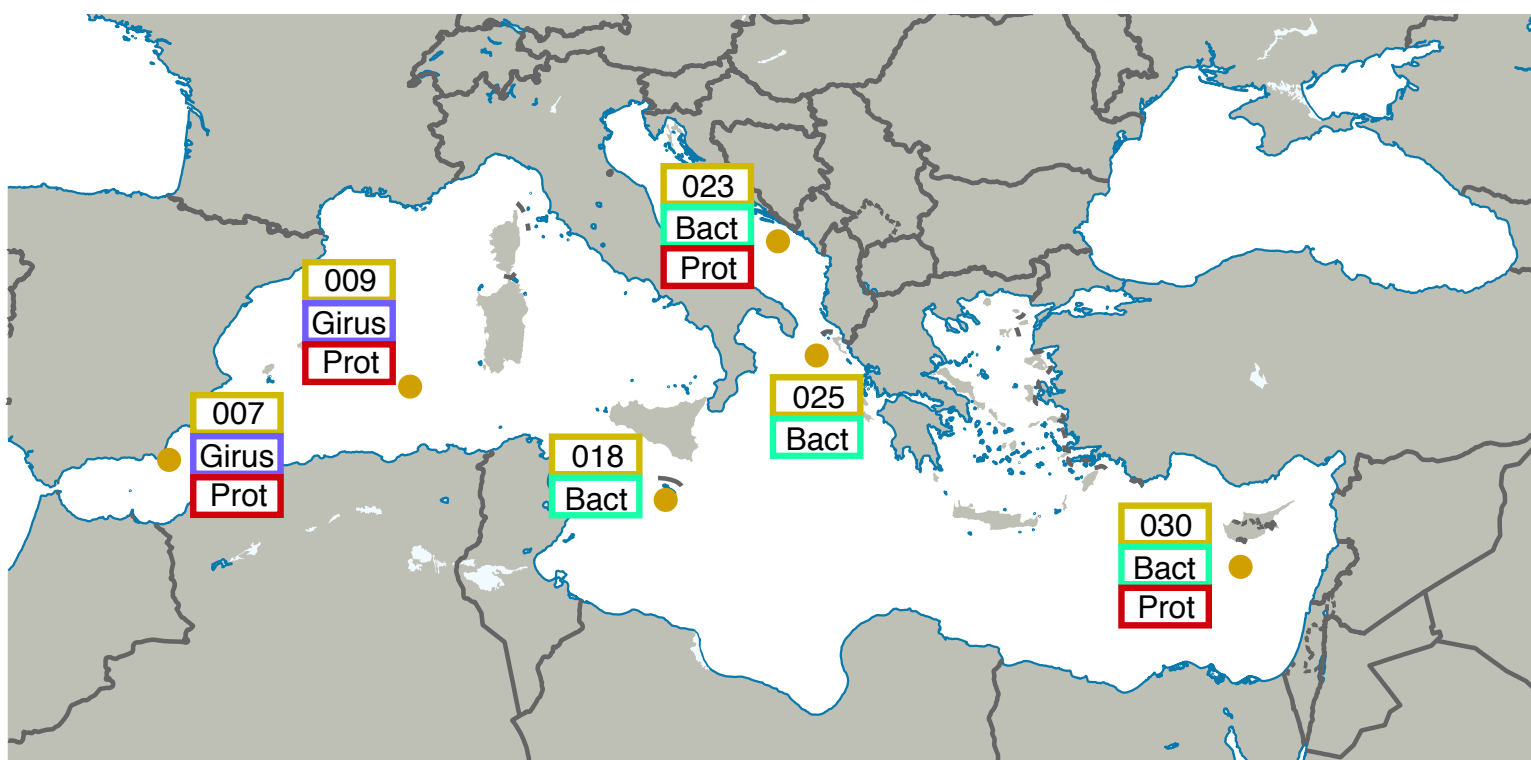


Figure 1. Map illustrating the locations and size fractions sampled for the Tara Oceans Mediterranean Sea datasets. Girus, 'giant virus' size fraction (0.22-1.6 μm). Bact, 'bacteria' size fraction (0.22-1.6 μm). Prot, 'protist' size fraction (0.8-5.0 μm)

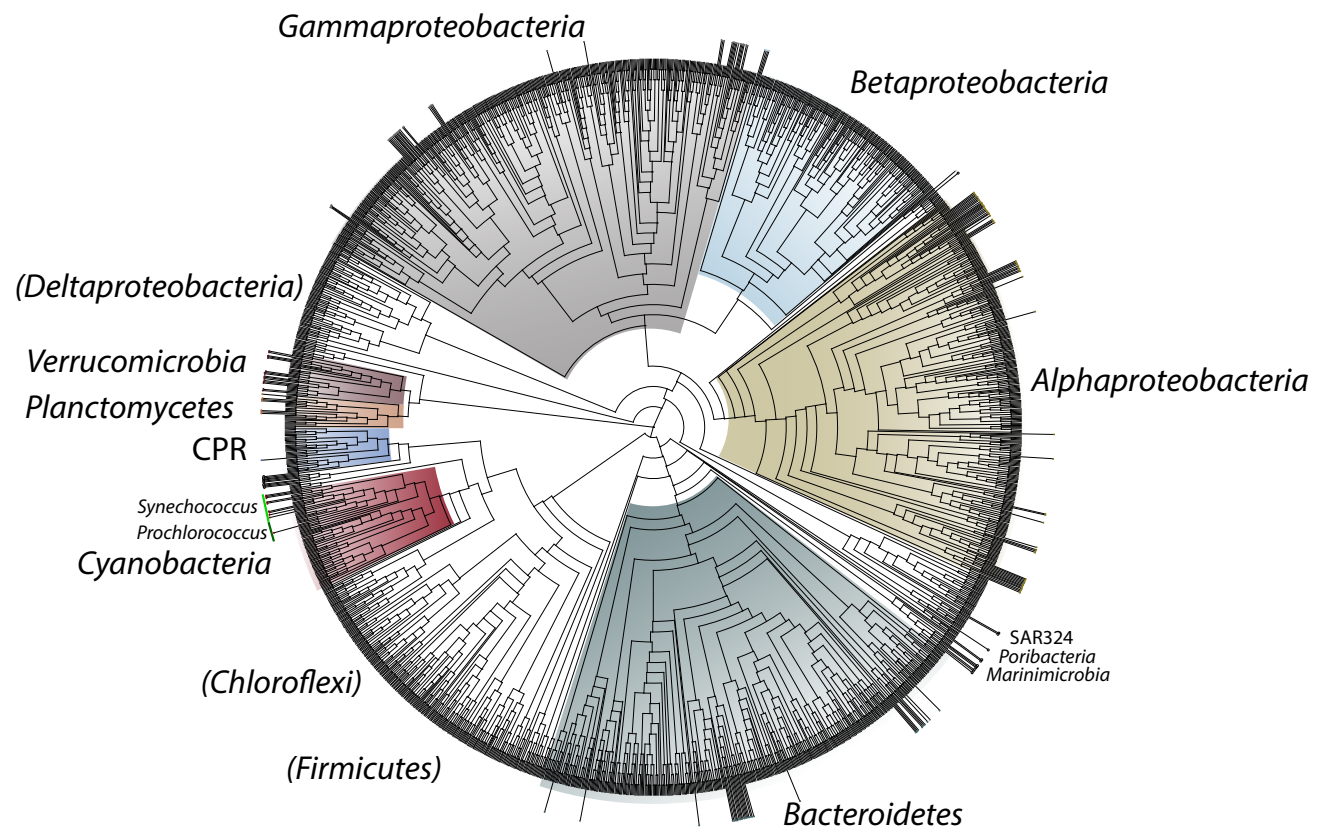


Figure 4. Cladogram of a FastTree approximate maximum-likelihood phylogenetic tree constructed using 16 syntenic, single-copy marker genes for 193 high-quality genomes and 1,729 reference genomes. Leaves denoting the position of the TMED genomes have been indicated by extending beyond the edge of the tree. Sequence alignment is available in Supplemental Information S4. Phylogenetic tree newick file is available in Supplemental Information S5.

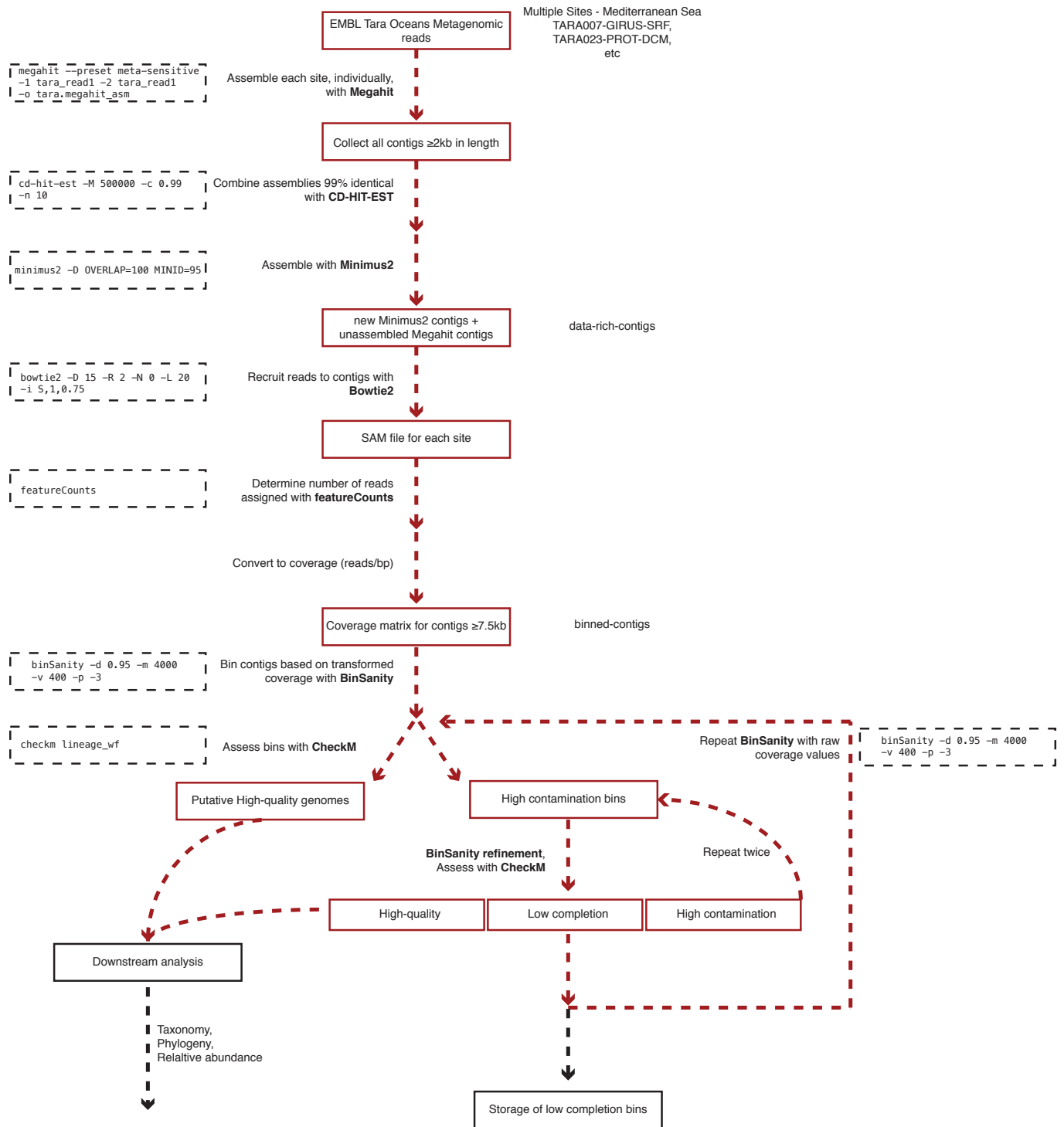


Figure 2. Workflow used to process Tara Oceans Mediterranean Sea metagenomic datasets. Black hash boxes, program or tool used with parameters.

Figure 3. FastTree approximate maximum-likelihood phylogenetic tree constructed with 37 and 785 16S rRNA genes from putative high-quality genomes and references, respectively.

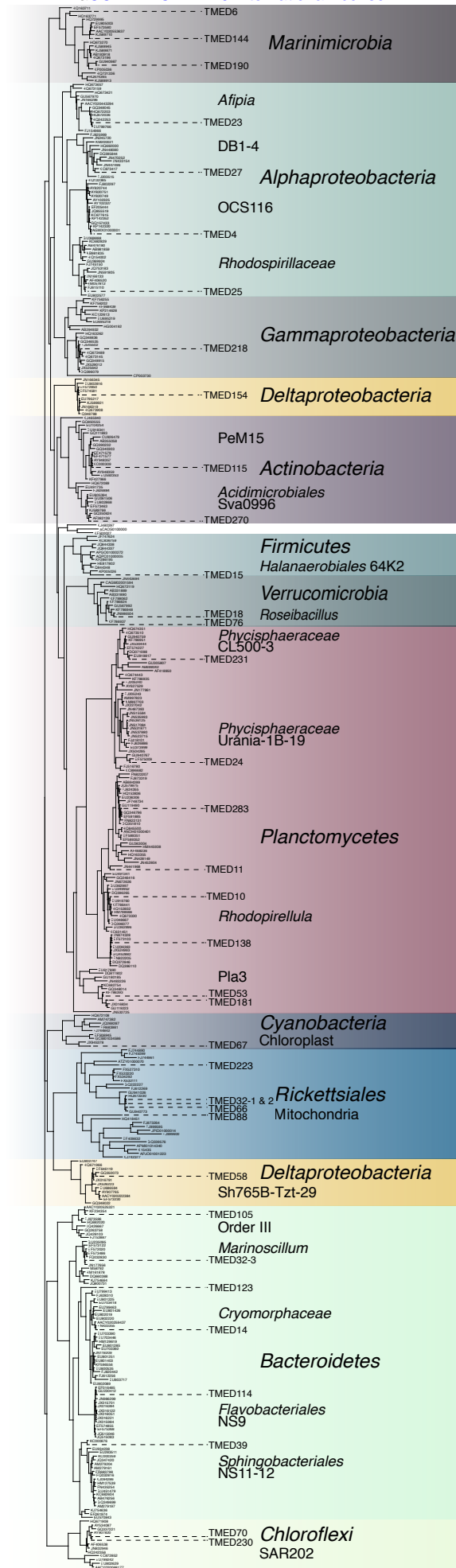


Table 1. Statistics for Megahit assemblies, recruitment to data-rich-contigs, and relative abundance of high-quality genome results for each sample

TARA Sample Site	Size Fraction (Girus, Bacteria, or Protist)	Depth (Surface or DCM*)	No. of reads	No. of initial Megahit assembly	N50 ^a (bp; initial Megahit assembly)	Longest initial Megahit assembly (bp)	Recruitment (% data-rich-contigs)	Relative abundance ^g of high-quality genomes (%)	Relative abundance ^g of ten most abundant genomes (%)
TARA007	Girus	DCM	178,519,830	1,318,470	828	220,754	72.84	14.64	6.35
TARA007	Girus	Surface	221,166,612	1,308,847	861	211,946	81.74	14.83	6.12
TARA007	Protist	DCM	744,458,992	4,667,618	654	188,635	19.45	8.60	3.18
TARA007	Protist	Surface	265,432,098	2,590,120	564	18,444	25.58	1.57	0.61
TARA009	Girus	DCM	416,553,274	2,796,841	831	1,643,839	69.48	14.16	6.32
TARA009	Girus	Surface	489,617,426	1,787,467	929	1,142,851	68.85	12.29	4.76
TARA009	Protist	DCM	329,036,110	1,938,636	613	95,724	22.07	13.35	4.20
TARA009	Protist	Surface	370,813,078	1,700,350	588	292,050	22.53	15.97	6.17
TARA018	Bacteria	DCM	408,021,182	2,520,645	840	1,573,060	76.22	11.49	3.18
TARA018	Bacteria	Surface	414,976,308	2,604,031	816	2,086,508	75.80	11.03	3.02
TARA023	Bacteria	DCM	147,400,552	1,273,576	830	213,456	76.08	13.29	4.09
TARA023	Bacteria	Surface	149,566,010	1,237,617	825	134,179	75.98	13.82	4.01
TARA023	Protist	DCM	508,610,652	2,707,801	734	336,689	28.23	25.07	7.83
TARA023	Protist	Surface	397,044,232	2,246,571	593	397,140	23.00	25.16	10.31
TARA025	Bacteria	DCM	386,627,816	2,516,865	806	388,546	69.77	14.55	5.35
TARA025	Bacteria	Surface	457,560,422	2,326,838	857	330,773	75.57	10.99	3.18
TARA030	Bacteria	DCM	346,837,034	1,968,945	1097	508,775	80.16	10.31	2.57
TARA030	Bacteria	Surface	478,785,582	1,639,697	1194	204,976	77.70	7.26	2.64
TARA030	Protist	DCM	426,896,616	1,620,343	616	478,892	15.12	17.83	5.13
TARA030	Protist	Surface	430,029,974	1,838,588	628	287,782	22.36	17.60	6.73

*DCM - deep chlorophyll maximum

^aN50 - length of DNA sequence above which 50% of the total is contained

^grelative abundance - determined using the reads recruited data-rich-contigs

Table 2. Assembly statistics at various steps during processing

Contig Grouping	No. of contigs	N50*	Total sequence (bp)
Megahit assemblies 200-499bp	24,999,285	n.d.	9,293,098,676
Megahit assemblies 500-1,999bp	16,103,221	n.d.	13,382,057,993
Megahit assemblies ≥ 2 kb	1,517,360	4,658	6,691,877,664
Megahit assemblies ≥ 2 kb (post-CD-HIT-EST)	1,126,975	4,520	4,894,479,496
Minimus2 contigs	158,414	15,394	1,727,079,865
Minimus2 + unassembled Megahit contigs ≥ 2 kb (data-rich-contigs)	660,937	5,466	3,612,405,904
Minimus2 + unassembled Megahit contigs ≥ 7.5 kb (binned-contigs)	95,506	20,556	1,725,063,313

*N50 - length of DNA sequence above which 50% of the total is contained