

# SCIENTIFIC DATA

**OPEN**

**SUBJECT CATEGORIES**

- » Water microbiology
- » Marine biology
- » Genome informatics

Received: 20 April 2016

Accepted: 25 May 2016

Published: 5 July 2016

## Data Descriptor: A catalogue of 136 microbial draft genomes from Red Sea metagenomes

Mohamed F. Haroon<sup>1</sup>, Luke R. Thompson<sup>1,2</sup>, Donovan H. Parks<sup>3</sup>, Philip Hugenholtz<sup>3,4</sup> & Ulrich Stingl<sup>1</sup>

Earth is expected to continue warming and the Red Sea is a model environment for understanding the effects of global warming on ocean microbiomes due to its unusually high temperature, salinity and solar irradiance. However, most microbial diversity analyses of the Red Sea have been limited to cultured representatives and single marker gene analyses, hence neglecting the substantial uncultured majority. Here, we report 136 microbial genomes (completion minus contamination is  $\geq 50\%$ ) assembled from 45 metagenomes from eight stations spanning the Red Sea and taken from multiple depths between 10 to 500 m. Phylogenomic analysis showed that most of the retrieved genomes belong to seven different phyla of known marine microbes, but more than half representing currently uncultured species. The open-access data presented here is the largest number of Red Sea representative microbial genomes reported in a single study and will help facilitate future studies in understanding the physiology of these microorganisms and how they have adapted to the relatively harsh conditions of the Red Sea.

<b>Design Type(s)</b>	observation design • organism identification objective
<b>Measurement Type(s)</b>	DNA sequence data
<b>Technology Type(s)</b>	metagenomics analysis
<b>Factor Type(s)</b>	
<b>Sample Characteristic(s)</b>	sea water • Red Sea • Gulf of Aden

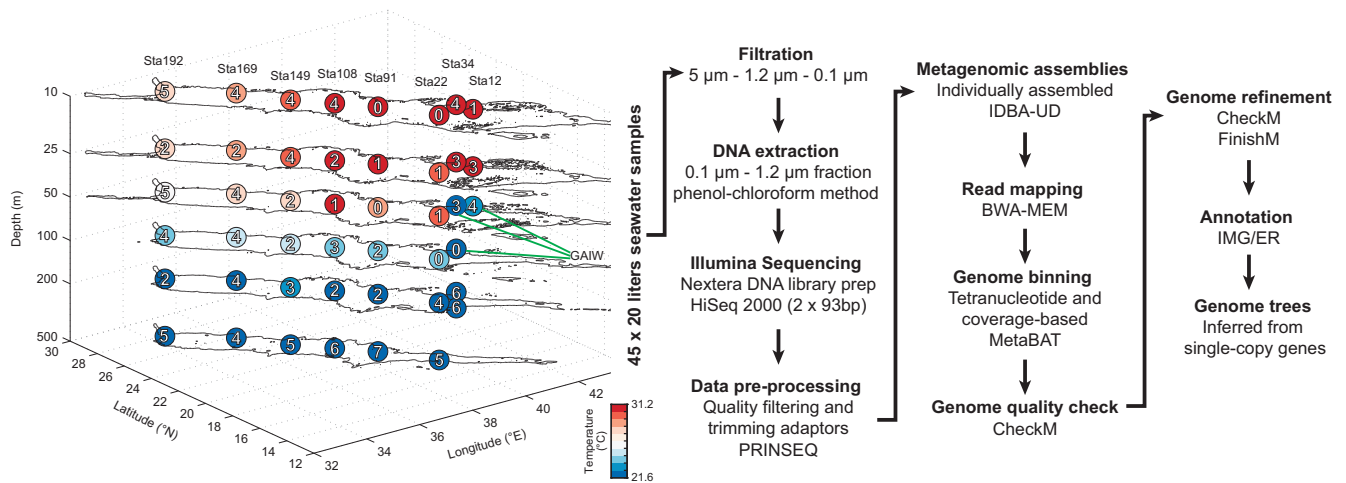
<sup>1</sup>Red Sea Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia. <sup>2</sup>Department of Pediatrics, University of California, San Diego, California 92037, USA. <sup>3</sup>Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, Queensland 4072, Australia. <sup>4</sup>Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072, Australia. Correspondence and requests for materials should be addressed to M.F.H. (email: fauziharoon@gmail.com) or to U.S. (email: ulistingl@gmail.com).

## Background & Summary

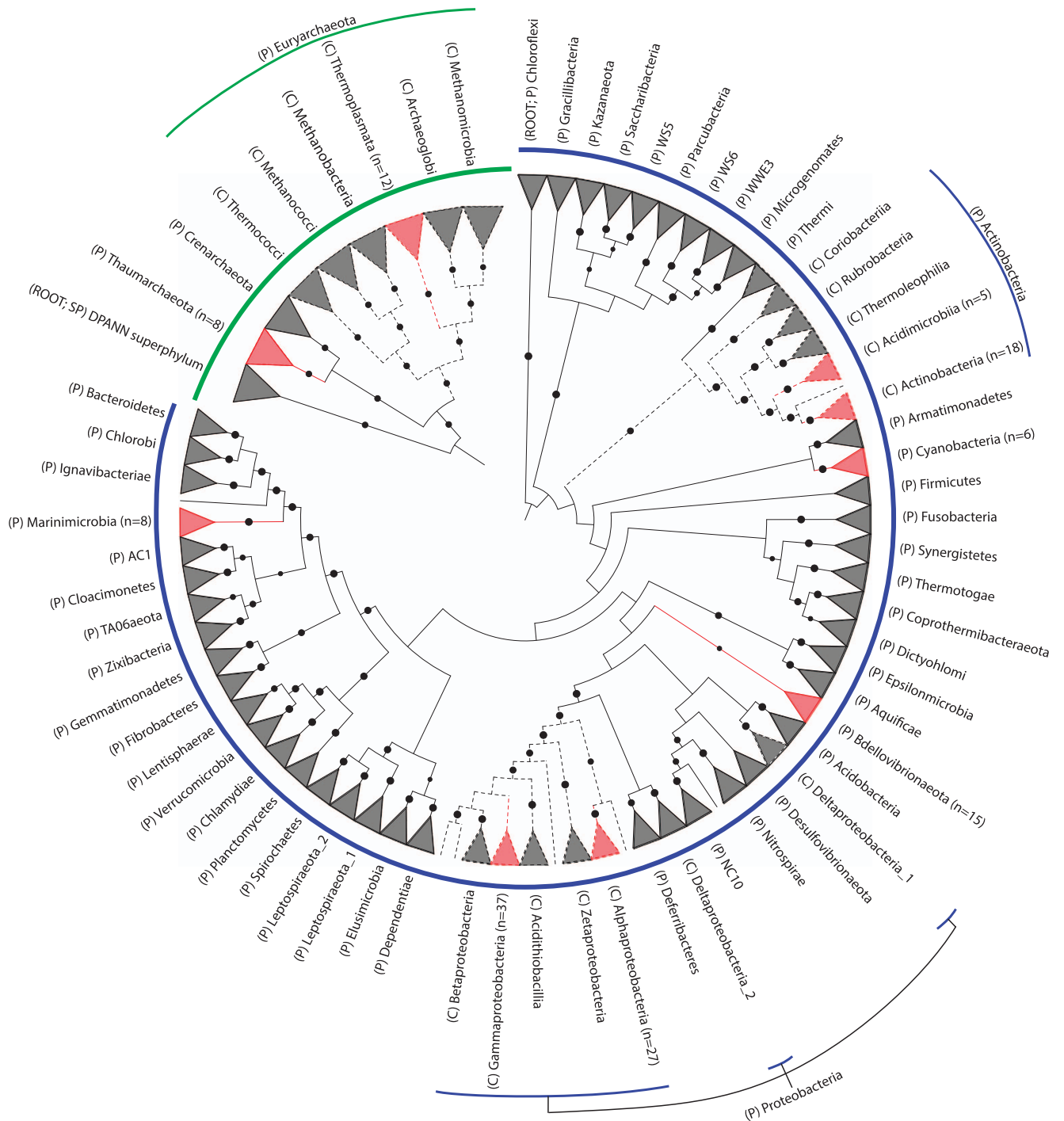
The Red Sea is an ideal marine environment to study microbial adaptation to physical conditions atypical of global oceans: high temperature, high salinity, and high irradiance. In late summer 2011, we undertook the King Abdullah University of Science and Technology (KAUST) Red Sea Expedition (KRSE2011) in the eastern Red Sea in order to map its diversity along environmental gradients that occur with changes in latitude, longitude, and depth<sup>1</sup>. This time of year is not only when temperatures and evaporation (and hence salinity) are highest, but also when a foreign water mass called the Gulf of Aden Intermediate Water (GAIW) intrudes into the Red Sea<sup>1,2</sup> (Fig. 1). The GAIW brings nutrient-rich water to the Red Sea, providing nitrogen, phosphorus, and other elements to this otherwise oligotrophic sea, and is likely to introduce important microbial diversity.

Insights into the taxonomic, evolutionary, and functional diversity of the Red Sea have largely been based on studies of pure cultures<sup>3–5</sup> and single marker genes such as the 16S rRNA<sup>6,7</sup>, or internal transcribed spacer<sup>8</sup>. Recently, investigations of microbial ecology have steered towards whole genome-based culture-independent methods notably single-cell genomics and metagenomics<sup>9,10</sup>. Single-cell genomics is an exciting field that recovers complete and partial single cell genomes from complex environments, albeit the need of specialised equipment, high cost and relatively low throughput<sup>11–13</sup>. Metagenomics is paving the way forward by harnessing the recent wave of sequencing technology and bioinformatics advancements to recover genomes of individual populations or populations of closely related organisms<sup>14–16</sup>. Application of these methods has resulted in the recovery of numerous genomes of uncultivated microorganisms that have provided surprising insights into the diversity and function of microbial communities<sup>10,14,17–19</sup>.

During the KRSE2011, eight stations were sampled along a cruise track from south to north, capturing gradients in temperature, salinity, oxygen, and nutrients, including the unique GAIW water mass (Fig. 1 and Table 1 (available online only)). At each station, samples were collected from the surface to mesopelagic depths (10, 25, 50, 100, 200, and 500 m), except for stations 12 and 34, which had depths shallower than 500 m (Fig. 1 and Table 1 (available online only)), in order to capture a greater variation in environmental parameters and microbial diversity. Here, we successfully reconstructed 136 genomes from 45 individually assembled metagenomes (Figs 1 and 2, Tables 1 and 2 (available online only), Data Citation 1) by differential read coverage and tetranucleotide frequency methods. Of these, 43 were ‘near-complete’ with an estimated completion minus contamination of  $\geq 90\%$ , while the other 93 draft genomes had completion minus contamination of  $\geq 50\%$  (Table 2 (available online only)).



**Figure 1.** Experimental workflow for this study. The circles superimposed on the Red Sea 3D map shows the sampling points during the King Abdullah University of Science and Technology Red Sea Expedition 2011. The green lines represent the three Gulf of Aden Intermediate Water (GAIW) sampling points. The numbers within the circles represent the number of genomes recovered from each of the sample. Colors represent the high (dark red) to low (dark blue) water temperature. A total of 45 samples of 20 l each were collected and filtered through a series of filters. For this study, DNA extraction was performed on the small microbial fractions (between 0.1 to 1.2  $\mu\text{m}$ ). Extracted DNA was sequenced on the Illumina HiSeq 2,000 generating paired-end reads ( $2 \times 93$  bp). Reads from each metagenome were cleaned and assembled individually. Genomes were binned based on tetranucleotide and coverage-based method, refined and quality checked. All 136 genomes were annotated by IMG/ER and taxonomically assigned based on genome trees inferred from single-copy genes.



**Figure 2.** Phylogenetic trees for the archaeal (green lines; top left) and bacterial (blue lines; bottom right) domains based on 122 and 120 single-copy marker genes, respectively. The clades represented by the triangles are collapsed at the phylum (P) level except for phyla containing genomes from this study which are expanded at the class (C) level and highlighted in red. Certain phyla have genome representatives only at the phylum level (Thaumarchaeota, Marinimicrobia, Cyanobacteria, and Bdellovibrionaeota). Numbers in parentheses indicate the count of recovered genomes from a particular taxonomic level. Dashed lines indicate nodes for class level. Robustness of the tree is indicated by black circles (size of circles scaled from 80 to 100% bootstrap support values). Trees were inferred independently. The archaeal tree was rooted with the DPANN superphylum<sup>9</sup> while the bacterial tree was ‘arbitrarily’ rooted with the phylum Chloroflexi<sup>42</sup> but should be treated as unrooted.

To our knowledge, this is the largest number of microbial genomes from the Red Sea to be reported in a single study.

Phylogenomic analysis based on sets of single-copy marker genes universal to either the bacterial or archaeal domain showed that the 136 genomes encompassed seven phyla across these domains: Thaumarchaeota, Euryarchaeota, Actinobacteria, Cyanobacteria, Bdellovibrionaeota, Proteobacteria, and Marinimicrobia (Fig. 2 and Table 2 (available online only)). As expected, most of the recovered genomes were affiliated with known marine microorganisms such as phototrophic *Prochlorococcus*<sup>20,21</sup> and *Synechococcus*<sup>22,23</sup>; representative of clades first discovered in the Sargasso Sea (SAR86, SAR116, SAR324 and SAR406)<sup>24–26</sup>; common marine bacteria in tropical biomes such as *Alteromonas macleodii*<sup>27</sup>; an ammonia oxidizing thaumarchaeon from the genus *Nitrosopelagicus*<sup>28</sup>; euryarchaeotal Marine Group II organisms reported to be abundant in surface waters<sup>29</sup>; members of the *Alpha*- and *Gamma*-proteobacteria such as *Aeromicrobium*, *Erythrobacter*, *Maritimibacter*, *Idiomarina*, *Marinobacter*, *Candidatus Thioglobus* (SUP05 cluster) and several unclassified *Gammaproteobacteria*, consistent with the high relative abundance of these two groups in the recent Tara Oceans survey<sup>30</sup>. Additionally, actinobacterial *Acidiimicrobia* and *Nocardioidea* genomes thought to be responsible for secondary metabolite production in marine ecosystems<sup>31</sup> were recovered from the metagenomes. An important strength of this dataset is the recovery of multiple, closely-related genomes from different stations or depths in the Red Sea (Data Citation 2). When complemented with physicochemical data<sup>1</sup>, genome plasticity between these organisms to confer fitness under varying conditions can be investigated in future studies.

To allow easy access to the genomes, all 136 genomes were functionally annotated and deposited into the National Centre for Biotechnology Information (NCBI) and Integrated Microbial Genomes (IMG) databases<sup>32</sup>. The wealth of metagenomic and genomic data described here greatly expands the repertoire of microbial genomic information from the Red Sea which might help to better understand the effects of global warming to ocean microbiomes. These datasets will also strengthen studies to better understand the drivers of marine nutrient cycling, help approaches for bioprospecting for novel thermo- and halo-philic enzymes, and allow for a better understanding of microbial adaptation strategies against high temperature, salinity and solar irradiance.

## Methods

### Metagenomic sequencing and assembly

Seawater samples were collected from eight stations and from different depths (10, 25, 50, 100, 200, and 500 m; locations are shown in Fig. 1) during summer as part of KRSE2011 (ref. 1). Genomic DNA was extracted from the 0.1–1.2  $\mu\text{m}$  size fraction using an established phenol-chloroform extraction protocol<sup>1,33</sup>. Paired-end libraries (2  $\times$  100 bp) were prepared using Nextera DNA Library Prep Kit (Illumina) and sequenced on a HiSeq 2000 (Illumina). Reads were quality checked and trimmed using PRINSEQ v0.20.4 (ref. 34) generating read lengths of  $\sim$ 93 bp and a total of  $\sim$ 10 million reads per sample with median insert sizes ranging from 183–366 bp<sup>1</sup> (Data Citation 1). Trimmed metagenome reads were individually assembled (Table 1 (available online only)) using IDBA-UD v1.1.1 (ref. 35) using the ‘--pre-correction’ option. To obtain coverage profile of contigs from each metagenomic assembly, the trimmed reads were mapped back to contigs using BWA v0.7.12 (ref. 36) with the bwa-mem algorithm.

### Genome binning, refinement, and annotation

For each metagenome, genome bins were recovered based on tetranucleotide frequencies and read coverage using MetaBAT v0.26.1 (ref. 37) with default parameters. The completeness and contamination of the bins were assessed using CheckM v1.0.3 (ref. 38) using the lineage-specific workflow (Table 2 (available online only)). Bins were further refined using the CheckM ‘merge’ and ‘outliers’ commands which merge bins with complementary sets of marker genes to improve completeness and remove contigs from bins which appear to be outliers relative to reference GC and tetranucleotide distributions in order to reduce contamination<sup>38</sup>. The FinishM v0.0.7 (<https://github.com/wwood/finishm>) ‘roundup’ workflow which comprise of ‘wander’ and ‘gapfill’ modes was used to scaffold contigs together and fill gaps within individual bins. The ‘wander’ mode uses a de Bruijn graph (kmer length of 51 bp and coverage cutoff of 5) to determine contig ends which are connected while the ‘gapfill’ mode align the reads to regions of ambiguous nucleotides and replaces them with the appropriate nucleotides. Genome bins that passed the quality filter of completion minus contamination of  $\geq$ 50% were submitted to IMG/ER<sup>32</sup> for gene calling and functional annotation.

### Genome tree construction

The archaeal and bacterial genome trees (Fig. 2) were inferred from the concatenation of 122 and 120 proteins, respectively, identified as being present in  $\geq$ 90% of the genomes in their respective domains and, when present, single-copy in  $\geq$ 95% of genomes (Supplementary Tables 1 and 2). These marker genes were aligned using HMMER v3.1b1 (ref. 39) and the tree inference from the concatenated alignment with FastTree v2.1.7 (ref. 40) under the WAG+GAMMA models (Data Citation 2). Support values were determined using 100 non-parametric bootstrap replicates<sup>41</sup>. The archaeal tree was rooted with the DPANN (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanohaloarchaeota, and Nanoarchaeota) superphylum in concordance with a recent large-scale phylogenomic study<sup>9</sup> while the bacterial tree was ‘arbitrarily’ rooted with the phylum Chloroflexi<sup>42</sup> but should be treated as unrooted. The trees were visualized in ARB<sup>43</sup>, annotated by iTOL<sup>44</sup> and edited in Illustrator CC 2014 (Adobe).

### Code availability

All versions of third-party software and scripts used in this study are described and referenced accordingly in the Methods sub-sections for ease of access and reproducibility.

### Data Records

The raw Illumina sequencing paired-end reads (Table 1 (available online only)), 45 assembled metagenome sequences (Table 1 (available online only)) and 136 assembled genome sequences (Table 2 (available online only)), generated from the KAUST Red Sea Expedition 2011, are available from NCBI databases (Data Citation 1). The genome trees and associated fasta amino acid alignment files are available from Figshare (Data Citation 2).

### Technical Validation

To validate the completeness and contamination of the genomes, we accessed the number of marker genes present in all bacterial and archaeal genomes using CheckM<sup>38</sup>. The genomes were also manually cleaned from vector contamination by comparing against the UniVec core database (<ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/>).

### Usage Notes

The annotated genome assemblies can be downloaded and accessed via the Integrated Microbial Genomes (IMG) system (<https://img.jgi.doe.gov/cgi-bin/m/main.cgi>). The IMG genome IDs are provided in Table 2 (available online only).

### References

1. Thompson, L. R. *et al.* Metagenomic covariation along densely sampled environmental gradients in the Red Sea. *bioRxiv* doi:10.1101/055012 (2016).
2. Churchill, J. H., Bower, A. S., McCorkle, D. C. & Abualnaja, Y. The transport of nutrient-rich Indian Ocean water through the Red Sea and into coastal reef systems. *Journal of Marine Research* **72**, 165–181 (2014).
3. Sagar, S. *et al.* Cytotoxic and apoptotic evaluations of marine bacteria isolated from brine-seawater interface of the Red Sea. *BMC Complementary and Alternative Medicine* **13**, 1–8 (2013).
4. Jimenez-Infante, F. *et al.* Genomic differentiation among two strains of the PS1 clade isolated from geographically separated marine habitats. *FEMS microbiology ecology* **89**, 181–197 (2014).
5. Zhang, G., Haroon, M. F., Zhang, R., Hikmawan, T. & Stingl, U. Draft Genome Sequence of *Pseudoalteromonas* sp. Strain XI10 Isolated from the Brine-Seawater Interface of Erba Deep in the Red Sea. *Genome Announcements* **4**, e00109–16 (2016).
6. Fuller, N. J. *et al.* Clade-specific 16S ribosomal DNA oligonucleotides reveal the predominance of a single marine *Synechococcus* clade throughout a stratified water column in the Red Sea. *Applied and Environmental Microbiology* **69**, 2430–2443 (2003).
7. Qian, P.-Y. *et al.* Vertical stratification of microbial communities in the Red Sea revealed by 16S rDNA pyrosequencing. *The ISME journal* **5**, 507–518 (2011).
8. Ngugi, D. K. & Stingl, U. Combined analyses of the ITS loci and the corresponding 16S rRNA genes reveal high micro- and macrodiversity of SAR11 populations in the Red Sea. *PLoS ONE* **7**, e50274 (2012).
9. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
10. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
11. Grötzinger, S. W. *et al.* Mining a database of single amplified genomes from Red Sea brine pool extremophiles—improving reliability of gene function prediction using a profile and pattern matching algorithm (PPMA). *Frontiers in Microbiology* **5**, 134 (2014).
12. Clingenpeel, S., Clum, A., Schwientek, P., Rinke, C. & Woyke, T. Reconstructing each cell's genome within complex microbial communities - dream or reality? *Frontiers in Microbiology* **5** (2015).
13. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).
14. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology* **31**, 533–538 (2013).
15. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotech.* **32**, 822–828 (2014).
16. Sangwan, N., Xia, F. & Gilbert, J. A. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**, 1–11 (2016).
17. Haroon, M. F. *et al.* Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. *Nature* **500**, 567–570 (2013).
18. Soo, R. M. *et al.* An expanded genomic representation of the phylum Cyanobacteria. *Genome biology and evolution* **6**, 1031–1045 (2014).
19. Evans, P. N. *et al.* Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* **350**, 434–438 (2015).
20. Moore, L. R., Rocab, G. & Chisholm, S. W. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393**, 464–467 (1998).
21. Partensky, F., Hess, W. & Vaulot, D. *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiology and molecular biology reviews* **63**, 106–127 (1999).
22. Moore, L. R., Goericke, R. & Chisholm, S. W. Comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Marine ecology progress series. Oldendorf* **116**, 259–275 (1995).
23. Palenik, B. *et al.* The genome of a motile marine *Synechococcus*. *Nature* **424**, 1037–1042 (2003).
24. Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**, 60–63 (1990).
25. Britschgi, T. B. & Giovannoni, S. J. Phylogenetic analysis of a natural marine bacterioplankton population by rRNA gene cloning and sequencing. *Applied and Environmental Microbiology* **57**, 1707–1713 (1991).
26. Haroon, M. F., Thompson, L. R. & Stingl, U. Draft genome sequence of uncultured SAR324 bacterium lautmerh10, binned from a Red Sea metagenome. *Genome Announcements* **4**, e01711–e01715 (2016).

27. Ivars-Martinez, E. *et al.* Comparative genomics of two ecotypes of the marine planktonic copiotroph *Alteromonas macleodii* suggests alternative lifestyles associated with different kinds of particulate organic matter. *The ISME journal* **2**, 1194–1212 (2008).
28. Santoro, A. E. *et al.* Genomic and proteomic characterization of ‘*Candidatus Nitrosopelagicus brevis*’: An ammonia-oxidizing archaeon from the open ocean. *Proceedings of the National Academy of Sciences* **112**, 1173–1178 (2015).
29. DeLong, E. F. Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences* **89**, 5685–5689 (1992).
30. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
31. Bull, A. T. & Stach, J. E. Marine actinobacteria: new opportunities for natural product search and discovery. *Trends in microbiology* **15**, 491–499 (2007).
32. Markowitz, V. M. *et al.* IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research* **40**, D115–D122 (2012).
33. Rusch, D. *et al.* The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biology* **5**, e77 (2007).
34. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
35. Peng, Y., Leung, H. C., Yiu, S.-M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
36. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
37. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
38. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**, 1043–1055 (2015).
39. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Computational Biology* **7**, e1002195 (2011).
40. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
41. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
42. Dagan, T., Roettger, M., Bryant, D. & Martin, W. Genome Networks Root the Tree of Life between Prokaryotic Domains. *Genome Biology and Evolution* **2**, 379–392 (2010).
43. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Research* **32**, 1363–1371 (2004).
44. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research* **39**, W475–W478 (2011).

## Data Citations

1. Haroon, M. F. *National Center for Biotechnology Information (NCBI) BioProject database* PRJNA289734 (2015).
2. Haroon, M. F. *Figshare* <https://dx.doi.org/10.6084/m9.figshare.3362899.v1> (2016).

## Acknowledgements

We acknowledge the people who were involved in the KAUST Red Sea Expedition 2011 and those that helped to generate the data, include, but are not limited to, those named here: Matt Cahill, Mamoon Rashid, Vinu Manikandan, David Ngugi and Ahmed Shibl. This work was supported by King Abdullah University of Science and Technology (KAUST), Saudi Basic Industries Corporation (SABIC) fellowship to L.R.T., and SABIC presidential chair to U.S.

## Author Contributions

M.F.H. trimmed and assembled the metagenomes, binned, refined and annotated the genomes, submitted all sequences to databases, made figures and tables, and wrote the manuscript. L.R.T. organized the cruise, collected seawater samples, extracted DNA and helped to write the manuscript. D.H.P. helped write the manuscript and constructed the genome trees. P.H. constructed the genome trees. U.S. planned the study, organized the cruise and helped write the manuscript.

## Additional Information

Tables 1 and 2 are only available in the online version of this paper.

Supplementary Information accompanies this paper at <http://www.nature.com/sdata>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Haroon, M. F. *et al.* A catalogue of 136 microbial draft genomes from Red Sea metagenomes. *Sci. Data* **3**:160050 doi: 10.1038/sdata.2016.50 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>

Metadata associated with this Data Descriptor is available at <http://www.nature.com/sdata/> and is released under the CC0 waiver to maximize reuse.