# Metagenome sequencing and 98 microbial genomes from Juan de Fuca Ridge Flank subsurface fluids

Sean P. Jungbluth[1]#, Jan P. Amend[1,2,3], and Michael S. Rappé[4]#

[1]Center for Dark Energy Biosphere Investigations, University of Southern California, Los Angeles, CA

[2]Department of Earth Sciences, University of Southern California, Los Angeles, CA

[3]Department of Biological Sciences, University of Southern California, Los Angeles, CA

[4]Hawaii Institute of Marine Biology, SOEST, University of Hawaii, Kaneohe, HI

#Corresponding authors: jungbluth.sean@gmail.com or rappe@hawaii.edu

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

## ABSTRACT

The global deep subsurface biosphere is thought to be one of the largest reservoirs for microbial life on our planet. This study takes advantage of new sampling technologies and couples them with improvements to DNA sequencing and associated informatics tools to reconstruct the genomes of uncultivated Bacteria and Archaea from fluids collected deep within the subseafloor of the Juan de Fuca Ridge. Here, we generated two metagenomes from borehole observatories located 311 meters apart and, using binning tools, retrieved 98 genomes from metagenomes (GFMs) with completeness > 10%. Of the GFMs, 31 were estimated to be > 90% complete, while an additional 17 were > 70% complete. In most instances, estimated redundancy in the GFMs was < 10%. Phylogenomic analysis revealed 53 bacterial and 45 archaeal GFMs and nearly all were distantly related to known cultivates. In the GFMs, abundant bacteria included Chloroflexi, Nitrospirae, Acetothermia (OP1), EM3, Aminicenantes (OP8), Gammaproteobacteria, and Deltaproteobacteria and abundant archaea included Archaeoglobi, Bathyarchaeota (MCG), and Marine Benthic Group E (MBG-E). In this study, we identified the first near-complete genomes from archaeal and bacterial lineages THSCG, MBG-E, and EM3 and, based on the warm, subsurface and hydrothermally-associated from which these groups tend to be found, propose the names, Geothermarchaeota, Hydrothermarchaeota, and Hydrothermae, respectively. The data set presented here are the first description of Juan de Fuca igneous basement microbial GFMs reported and will provide a platform by which one can perform a higher level interrogation of the many uncultivated lineages presented herein.




PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

## BACKGROUND & SUMMARY

Beneath the sediments of the deep ocean, the subseafloor igneous basement presents a largely unexplored habitat that may play a crucial role in global biogeochemical cycling[1]. This system also provides a gradient of untapped environments for the discovery of novel microbial life. Because of extensive hydrothermal circulation, the porous uppermost igneous crust is likely quite suitable for microbial life[2]. Entrainment of deep seawater into young ridge flanks injects a variety of terminal electron acceptors into the deep ocean crust, establishing chemical gradients with the reducing deeper fluids, and thereby fueling redox-active elemental cycles[3]. The redox disequilibria and circulation of fluids through the permeable network of volcanic rock sustains a largely uncharacterized microbial community that potentially extends thousands of meters below the seafloor[4]. In such environments, temperatures are elevated, and energy and nutrient sources may be extremely limited, the combination of which provides unique challenges to microbial life.

CORK (circulation obviation retrofit kit) observatories have been used in recent years to collect warm, anoxic crustal fluids originating from boreholes drilled into 1.2 and 3.5 million-year-old ridge flank of the Juan de Fuca Ridge (JdFR)[5]. This young, hydrologically-active basaltic crustal environment is overlain by a thick (>100 m) blanket of sediment that serves to locally restrict fluid circulation in the ocean basement[6,7]. The sampling of raw basement fluids enabled by CORK observatories has demonstrated the presence of novel microbial lineages that are related to uncultivated candidate microbial phyla with unknown metabolic characteristics[8-11]. Here, we present the genomes from metagenomes (GFMs) of two pristine large-volume igneous basement fluid samples



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

collected from JdFR flank boreholes CORK observatories U1362A and U1362B (Figures 1A, 1B, and 1C).

Shotgun sequencing produced 503 and 705 megabase pairs (Mbp) of unassembled sequence data from individual borehole U1362A and U1362B samples (Table 1). The metagenomes were assembled separately into 137,575 and 212,307 scaffolds totaling 170 and 168 Mbp of sequence data from U1362A and U1362B, respectively (Tables 1 and 2). The maximum scaffold lengths constructed from U1362A and U1362B metagenome were, respectively, 541 and 1,137 Mbp (Table 2). The success of this assembly to generate long scaffolds that represent major, intact fractions of individual genomes provides a significant foundation for which to apply binning methods to piece together genomes from populations in the original samples.

Several methods were used to generate GFMs, which were then evaluated, further curated, and reduced to a set for additional characterization. Ultimately, analysis was performed on 98 GFMs that were over 200 Kbp in length, contained marker gene sets identified by CheckM, and were >10% complete (Tables 3 and S1).

Phylogenetic analysis of concatenated universally conserved marker gene alignments (Figures 2-5) and taxonomic identification of SSU rRNA genes (Table S2) allowed for the phylum-level identification of most of the 53 bacterial and 45 archaeal GFMs. The U1362A and U1362B borehole fluid GFMs were comprised of many of the same microbial lineages described previously using SSU rRNA sequencing[8,11]; including bacterial groups Chloroflexi (11), Nitrospirae (8), Acetothermia (OP1; 7), EM3 (5), Aminicenantes (OP8; 4), Gammaproteobacteria (4), and Deltaproteobacteria (4), and archaeal groups Archaeoglobi (21), Bathyarchaeota (MCG; 9), and Marine Benthic



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

Group E (MBG-E; 3) (Tables 4 and S1). In this study, we identified the first near-complete genomes from archaeal and bacterial lineages THSCG, MBG-E, and EM3 and, based on the warm, subsurface and hydrothermally-associated from which these groups tend to be found, propose the names, Geothermarchaeota, Hydrothermarchaeota, and Hydrothermae, respectively.

The 98 genomes described here were functionally annotated and deposited into the National Center for Biotechnology Information (NCBI) and Integrated Microbial Genomes (IMG) databases[12]. The genome data described here are the first GFMs described from the deep subseafloor volcanic basement environment and will be used to interrogate the functional underpinnings of individual microbial lineages within this remote and distinct ecosystem. Considering that genome binning methods cannot yield comprehensive segregation of all entities in complex samples[13], and that informatics tools are continuously improving, we recommend that anyone using these data verify the contents of these GFMs with the latest tools available.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

## METHODS

**Borehole fluid sampling.** Sample collection methods are described elsewhere[11]. Briefly, during R/V Atlantis cruise ATL18-07 (28 June 2011 – 14 July 2011) samples of basement crustal fluids were collected from CORK observatories located in 3.5 million-year-old ocean crust east of the Juan de Fuca spreading center. Basement fluids were collected from lateral CORKs (L-CORKs) at boreholes U1362A (47°45.6628'N, 127°45.6720'W) and U1362B (47°45.4997'N, 127°45.7312'W) via polytetrafluoroethylene (PTFE)-lined fluid delivery lines that extend to 200 (U1362A) and 30 (U1362B) meters sub-basement. Fluids were filtered in situ through Steripak-GP20 (Millipore, Billerica, MA, USA) polyethersulfone filter cartridges containing 0.22 µm pore-sized membranes using a mobile pumping system. Filtration rates were estimated at 1 L/min in laboratory trials, indicating that ~124 liters and ~70 liters were filtered from boreholes U1362A and U1362B, respectively.

**DNA extraction and metagenome sequencing.** Nucleic acids were extracted from borehole fluids using a modified phenol/chloroform lysis and purification method, and is described in detail elsewhere[11] (samples SSF21-22, SSF23-24). Library preparation, DNA sequencing, read quality-control, metagenome assembly, and gene prediction and annotation were conducted by the Department of Energy Joint Genome Institute as part of their Community Sequencing Program using previously described informatics workflows[12], which are described in detail elsewhere[14].



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

**Genome binning.** Assemblies from the U1362A and U1362B metagenomes were combined and used to generate GFMs. Four different genome binning approaches were used to identify the workflow that yielded the most favorable balance between maximizing genome completeness while minimizing contamination for these metagenomes: MaxBin[15], ESOM[16], MetaBAT[17], and CONCOCT[18].

Genome binning was performed using MaxBin version 2.1.1[15] with the 40 marker gene set universal among Bacteria and Archaea[19], minimum scaffold length of 2000 bp, and default parameters. Scaffold coverage from each metagenome was estimated using the quality-control filtered raw reads as input for mapping using Bowtie2 version 2.2.3[20] used within MaxBin.

Genome binning was also performed using a combination of tetranucleotide frequencies and differential coverage in emergent self-organizing maps (ESOM) [16]. Scaffold coverage was calculated using bbmap version 35.40 and the jgi_summarize_bam_contig_depths script from the MetaBAT pipeline[17]. Scripts downloaded from (http://github.com/tetramerFreqs/Binning) were used to calculate tetramer frequencies and create input files for ESOM. A robust Z-transformation was applied to the input data prior to generation of the ESOM. Scaffolds greater than or equal to 10 Kbp were cut into slices of 2000 bp prior to clustering. The number of epochs used for clustering was 20 and the dimensions of the ESOM were 400 x 430 (Figure 6).

Using MetaBAT version 0.26.3[17], genome binning was performed with the jgi_summarize_bam_contig_depths script and the same scaffold coverage map calculated using bbmap described above. Default parameters were used.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

Finally, genome binning was performed using CONCOCT[18] within the Anvi'o package, version 1.1.0[21]. The metagenomic workflow employed here is described online (merenlab.org/2015/05/02/anvio-tutorial), and included as input data the quality-filtered raw sequence reads from both metagenomes, as well as assemblies generated by the JGI. The scaffold coverage map was calculated using bbmap version 35.82. Scaffolds greater or equal to 2.5 Kbp were used for binning with CONCOCT.

**Comparison of genome binning methods and bin curation.** Completeness and redundancy of all GFMs created using the four binning methods were assessed using CheckM version 1.0.5[22]. Overall, GFMs generated with CONCOCT yielded the highest percent completeness for bins that were at least 50% complete (Table 3). Genome completeness was the primary criterion used in the selection of the binning method because the facilitated supervised binning via the "anvi-refine" function in Anvi'o works to remove contamination from a draft set of genome scaffolds. Manual refinements to the GFMs were executed in Anvi'o using differential coverage, tetranucleotide frequency, and marker gene content (i.e. completeness/redundancy). Bin splitting was assisted by the analysis of SSU rRNA genes identified using CheckM and inspected via the SILVA/SINA online aligner version 1.2.11[23] with the following parameters: minimum identity with query sequence, 0.8, and number of neighbors per query sequence, 3. When SSU rRNA genes of different taxonomic origin were found to conflict within a single bin, those bins were further scrutinized and split manually. In most instances where redundancy was > 50%, splitting bins into their U1362A and U1362B components resolved conflicts. Bins were split until no more SSU rRNA gene conflicts



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

were observed and all bins had been manually inspected and screened for outlying scaffolds. Four other marker gene sets[18,24-26] were used to compare completeness and redundancy within Anvi'o (Figure 7). A total of 252 GFMs were identified after curation with Anvi'o, and completeness and redundancy of the final GFMs was ultimately estimated with CheckM and the marker gene set of Wu and colleagues[19]. Of these, 98 were at least 10% complete (Tables 4 and S1), which was used as a minimum cutoff because the GFMs all contained marker genes that allowed them to be given phylogenetic identities using CheckM. The 98 GFMs included a total of 16,066 scaffolds and 154,610,017 bp.

**Phylogenomics and identification of genomes from metagenomes.** From all genomes described here with completeness > 10% and relevant GFMs and single-amplified genomes (SAGs) from the Integrated Microbial Genomes (IMG)[27], ggKbase, and National Center for Biotechnology Information (NCBI) GenBank databases, phylogenetically informative marker genes from were identified and extracted using the 'tree' command in CheckM. In CheckM, open reading frames were called using prodigal version 2.6.1[28] and a set of 43 lineage-specific marker genes, similar to the universal set used by PhyloSift[29], were identified and aligned using HMMER version 3.1b1[30]. The 61 GFMs with > 50% completeness were given taxonomic idendifications through analysis of a concatenated marker gene alignment (6988 amino acid positions) and placement in a phylogenomic tree with closest related GFMs and SAGs found in the NCBI, IMG, and ggKbase databases. The phylogeny was produced using FastTree (version 2.1.9[31]) with the WAG amino acid substitution model and 'fastest' mode.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

Bootstrap values reported by FastTree analysis indicate local support values. To leverage the taxonomic identifications assigned to the GFMs with > 50% completeness toward the identification of the 37 GFMs with completeness 10-50%, an additional phylogenetic analysis with only the 98 Juan de Fuca GFMs was performed in ARB[32] using RAxML version 7.7.2[33] with the PROTGAMMA rate distribution model and WAG amino acid substitution model. Bootstrapping was executred in ARB using the RAxML rapid bootstrap analysis algorithm[34] with 100 bootstraps. To further aid in identification of GFMs, SSU rRNA genes were extracted successfully from 49 genome bins using the "ssu_finder" command within CheckM and identified via the SILVA/SINA online aligner version 1.2.11 (Pruesse et al., 2012) with the version 123 database and the following parameters: minimum identity with query sequence, 0.8, and number of neighbors per query sequence, 3 (Table S2).

## DATA ACCESS

The U1362A and U1362B metagenome projects and raw sequencing reads are available via the IMG-M web portal under Taxon ID numbers 330002481 (U1362A) and 3300002532 (U1362B). Gold Analysis Project ID numbers are Ga0004278 (U1362A) and Ga0004277 (U1362B). Sample metadata can be accessed using the BioProject identifier PRJNA269163. The NCBI BioSamples used here are SAMN03166137 (U1362A) and SAMN03166138 (U1362B). FASTA files containing the contigs of all 98 genomes from metagenomes can be accessed at doi: 10.6084/m9.figshare.4269587.v1. A FASTA file containing 54 SSU rRNA genes with length >300 base pairs extracted from the 98 genomes from metagenomes can be accessed at doi:



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

10.6084/m9.figshare.4269593.v1. IMG/M-relevant files needed to isolate scaffold sets for all 98 genomes from metagenomes can be accessed at doi: 10.6084/m9.figshare.4269590.v1.

IMG/M annotations associated with the scaffolds of all 98 genomes from metagenomes can be accessed at doi: 10.6084/m9.figshare.4269581.v1.

## ACKNOWLEDGEMENTS

We thank the captain and crew, Andrew Fisher, Keir Becker, Geoff Wheat, and other members of the science teams on board R/V Atlantis cruise AT18-07. We also thank the pilots and crew of remote-operated vehicle *Jason II.* We are grateful to Huei-Ting Lin, Chih-Chiang Hsieh, Alberto Robador, Brian Glazer, and Jim Cowen for sampling, and Beth Orcutt and Ramunas Stepanauskas for facilitating metagenome sequencing. We thank Brian Foster, Alex Copeland, Tijana Glavina del Rio, and Susannah Tringe of the Department of Energy Joint Genome Institute for metagenome sequencing and assembly (Community Sequencing Award 987 to R. Stepanauskas). This study used samples and data provided by the Integrated Ocean Drilling Program.

## AUTHOR CONTRIBUTIONS

S.P.J. and M.S.R. designed the study. S.P.J. performed all analyses outside of those included in the standard operating procedure of the JGI, generated all figures and wrote the manuscript. All co-authors commented on and provided critical feedback for the final manuscript.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

**FUNDING INFORMATION**

This work, including the efforts of Sean Jungbluth and Michael Rappé, was funded by National Science Foundation grants MCB06-04014 and OCE-1260723. This work, including the efforts of all authors, was supported by the Center for Dark Energy Biosphere Investigations (C-DEBI), a National Science Foundation-funded Science and Technology Center of Excellence (OCE-0939564).

## REFERENCES


1 Schrenk, M. O., Huber, J. A. & Edwards, K. J. Microbial provinces in the subseafloor. *Ann Rev Mar Sci* **2**, 279-304, doi:10.1146/annurev-marine-120308-081000 (2010).

2 Baross, J. A., Wilcock, W. S. D., Kelley, D. S., DeLong, E. F. & Cary, S. C. in *The Subseafloor Biosphere at Mid-Ocean Ridges Geophysical Monograph* (eds W. S. D. Wilcock *et al.*) 1-11 (American Geophysical Union, 2004).

3 Edwards, K. J., Bach, W. & McCollom, T. M. Geomicrobiology in oceanography: microbe-mineral interactions at and below the seafloor. *Trends in Microbiology* **13**, 449-456, 10.1016/j.tim.2005.07.005 (2005).

4 Edwards, K. J., Fisher, A. T. & Wheat, C. G. The deep subsurface biosphere in igneous ocean crust: frontier habitats for microbiological exploration. *Frontiers in Microbiology* **3**, 8 (2012).

5 Wheat, C. G. *et al.* in *Proceedings of the Integrated Ocean Drilling Program* Vol. 327 (eds A. T. Fisher, T. Tsuji, K. Petronotis, & Expedition 327 Scientists) 1-36 (Integrated Ocean Drilling Program Management Internation, Inc., 2011).

6 Wheat, C. G. & Mottl, M. J. Hydrothermal circulation, Juan de Fuca Ridge eastern flank - factors controlling basement water composition. *Journal of Geophysical Research-Solid Earth* **99**, 3067-3080 (1994).

7 Cowen, J. P. The microbial biosphere of sediment-buried oceanic basement. *Res Microbiol* **155**, 497-506, doi:10.1016/j.resmic.2004.03.008 (2004).

8 Cowen, J. P. *et al.* Fluids from aging ocean crust that support microbial life. *Science* **299**, 120-123, doi:10.1126/science.1075653 (2003).

9 Jungbluth, S. P., Grote, J., Lin, H.-T., Cowen, J. P. & Rappé, M. S. Microbial diversity




PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

within basement fluids of the sediment-buried Juan de Fuca Ridge flank. *ISME Journal* **7**, 161-172 (2013).

10 Jungbluth, S. P., Lin, H.-T., Cowen, J. P., Glazer, B. T. & Rappé, M. S. Phylogenetic diversity of microorganisms in subseafloor crustal fluids from boreholes 1025C and 1026B along the Juan de Fuca Ridge flank. *Frontiers in Microbiology* **5**, 119, doi:10.2289/fmicb.2014.00119 (2014).

11 Jungbluth, S. P., Bowers, R., Lin, H.-T., Cowen, J. P. & Rappé, M. S. Novel microbial assemblages inhabiting crustal fluids within mid-ocean ridge flank subsurface basalt. *ISME Journal* **10**, 2033-2047 (2016).

12 Huntemann, M. *et al.* The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4). *Stand Genomic Sci* **11** (2016).

13 Nielsen, C. L. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnology* **32,** 822-828 (2014).

14 Jungbluth, S. P., Glavina del Rio, T., Tringe, S., Stepanauskas, R. & Rappé, M. S. Genomic characterization of a marine lineage ubiquitous throughout terrestrial and oceanic subsurface environments. *Submitted to PeerJ.*

15 Wu, W.-Y., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2** (2014).

16 Dick, G. J. *et al.* Community-wide analysis of microbial genome sequence signatures. *Genome Biol* **10**, R85, doi:10.1186/gb-2009-10-8-r85 (2009).



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

17 Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, doi:10.7717/peerj.1165 (2015).

18 Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**, 1144-1146, doi:10.1038/nmeth.3103 (2014).

19 Wu, D., Jospin, G. & Eisen, J. A. Systematic Identification of Gene Families for Use as "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. *PlOS One*, e77033, doi:10.1371/journal.pone.0077033 (2013).

20 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-U354, doi:10.1038/nmeth.1923 (2012).

21 Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3** (2015).

22 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 Res* **25**, 1043-1055, doi:10.1101/gr.186072.114 (2015).

23 Pruesse, E., Peplies, J. & Glockner, F. O. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823-1829, doi:10.1093/bioinformatics/bts252 (2012).

24 Creevey, C. J., Doerks, T., Fitzpatrick, D. A., Raes, J. & Bork, P. Universally Distributed Single-Copy Genes Indicate a Constant Rate of Horizontal Transfer. *PLoS One* **6**, e22099 (2011).



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

25 Dupont, C. L. *et al.* Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *Isme Journal* **6**, 1186-1199, doi:10.1038/ismej.2011.189 (2012).

26 Campbell, J. H. *et al.* UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc Natl Acad Sci U S A* **110**, 5540-5545 (2013).

27 Markowitz, V. M. *et al.* IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res* **42**, D568-573, doi:10.1093/nar/gkt919 (2014).

28 Hyatt, D., LoCascio, P. F., Hauser, L. J. & Uberbacher, E. C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**, 2223-2230, doi:10.1093/bioinformatics/bts429 (2012).

29 Darling, A. E. *et al.* PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**, e243, doi:10.7717/peerj.243 (2014).

30 Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195, doi:10.1371/journal.pcbi.1002195 (2011).

31 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490, doi:10.1371/journal.pone.0009490 (2010).

32 Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res* **32**, 1363-1371, doi:10.1093/nar/gkh293 (2004).

33 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690, doi:10.1093/bioinformatics/btl446 (2006).

34 Stamatakis, A., Hoover, P. & Rougemont, J. A rapid bootstrap algorithm for the RAxML



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

Web servers. *Syst Biol* **57**, 758-771, doi:10.1080/10635150802429642 (2008).

## FIGURE LEGENDS

Fig 1. (A) Bathymetric map of Juan de Fuca Ridge boreholes U1362A and U1362B with inset world map showing region location. (B) Schematic of CORK observatories at U1362A and U1362B. (C) Workflow used to process the basement crustal fluid samples and generate metagenomes and GFMs. This process included *in situ* filtration, extraction of nucleic acids via phenol-chloroform method, preparation of nucleic acids for Illumina sequencing, quality-filtering sequencing reads, assembling metagenome scaffolds, performing binning and associated quality-control and refinement, and gene annotation and phylogenetic analysis.

Fig 2. Phylogenomic relationships between archaeal genomes > 50% complete identified in CORK borehole fluid metagenomes and other closely related genomes retrieved from popular databases. The scale bar corresponds to 1.00 substitutions per amino acid position. Some groups are collapsed to enhance clarity and all groups with taxonomic identities are shown. The names of major lineages with GFMs found in Juan de Fuca Ridge basement fluids are indicated with the bold-face font. JdFR GFM prefixes are abbreviated from “JdFR” to “J” and labeled using red-colored text. Black (100%), gray (≥ 80%), and white (≥ 50%) circles indicate nodes with high local support values, from 1000 replicates.

Fig 3. Phylogenomic relationships between bacterial genomes > 50% complete identified in CORK borehole fluid metagenomes and other closely related genomes



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

retrieved from popular databases. JdFR GFM prefixes are labeled using green-colored text. Other information as in Figure 2.

Fig 4. Phylogenomic relationships between archaeal GFMs > 10% complete identified in metagenomes from deep subseafloor crustal fluids of boreholes U1362A and U1362B. Archaeal GFMs found in this study were used as the outgroup. The scale bar corresponds to 0.001 substitutions per amino acid position. Black (100%), gray (≥ 80%), and white (≥ 50%) circles indicate nodes with bootstrap support, from 100 replicates. All bins are abbreviated “J” for “JdFR”.

Fig 5. Phylogenomic relationships between bacterial GFMs > 10% complete identified in metagenomes from deep subseafloor crustal fluids of boreholes U1362A and U1362B. Bacterial GFMs found in this study were used as the outgroup. Other information as in Figure 4.

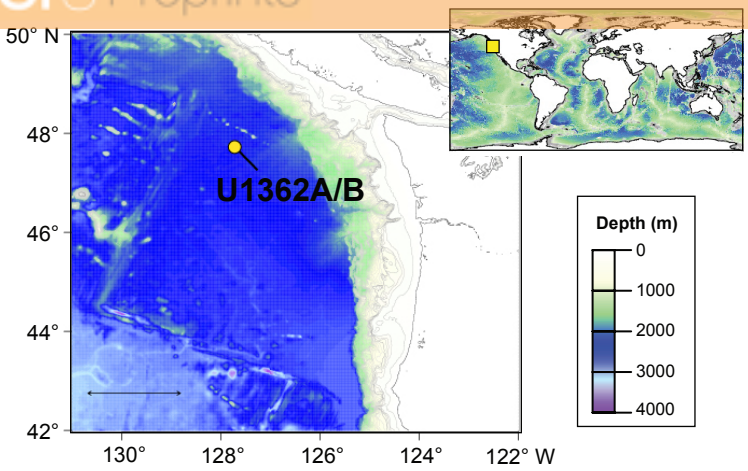
Fig 6. Assignment of contigs from CORK borehole fluid metagenomes using ESOM implemented with tetranucleotide frequencies and differential coverage. The ESOM is shown (A) before and (B) after identification of GFMs. Each point represents a contig and identified bins have a non-white color.

Fig 7. Overview of GFM completeness and redundancy and calculated average using five different marker gene sets. All bins are abbreviated “J” for “JdFR.”

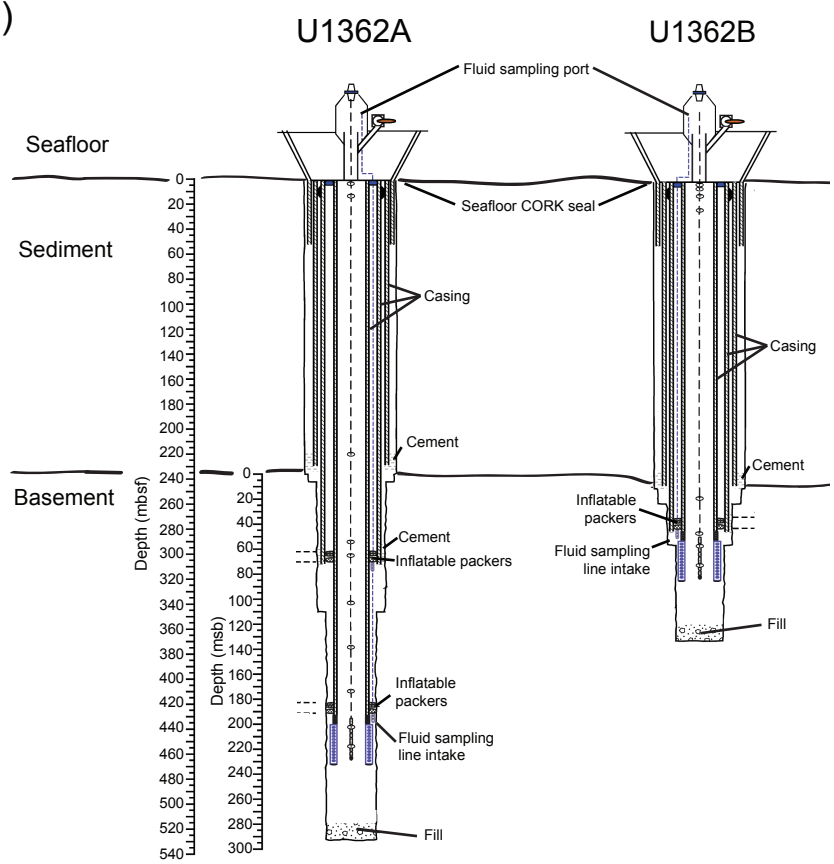


PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

A)




B)




C)

**Filtration**
*in situ*, 0.22 µm

↓

**DNA extraction**
phenol-chloroform method

↓

**Illumina Sequencing**
TruSeq library prep
HiSeq 2000 (2 x 150 bp)

↓

**Data pre-processing**
Quality-filtering and adapter trimming

↓

**Metagenome assemblies**
SOAPdenovo & Newbler, minimus 2

↓

**Genome binning**
Tetranucleotide and coverage-based CONCOCT

↓

**Genome quality control**
CheckM/Anvi'o

↓

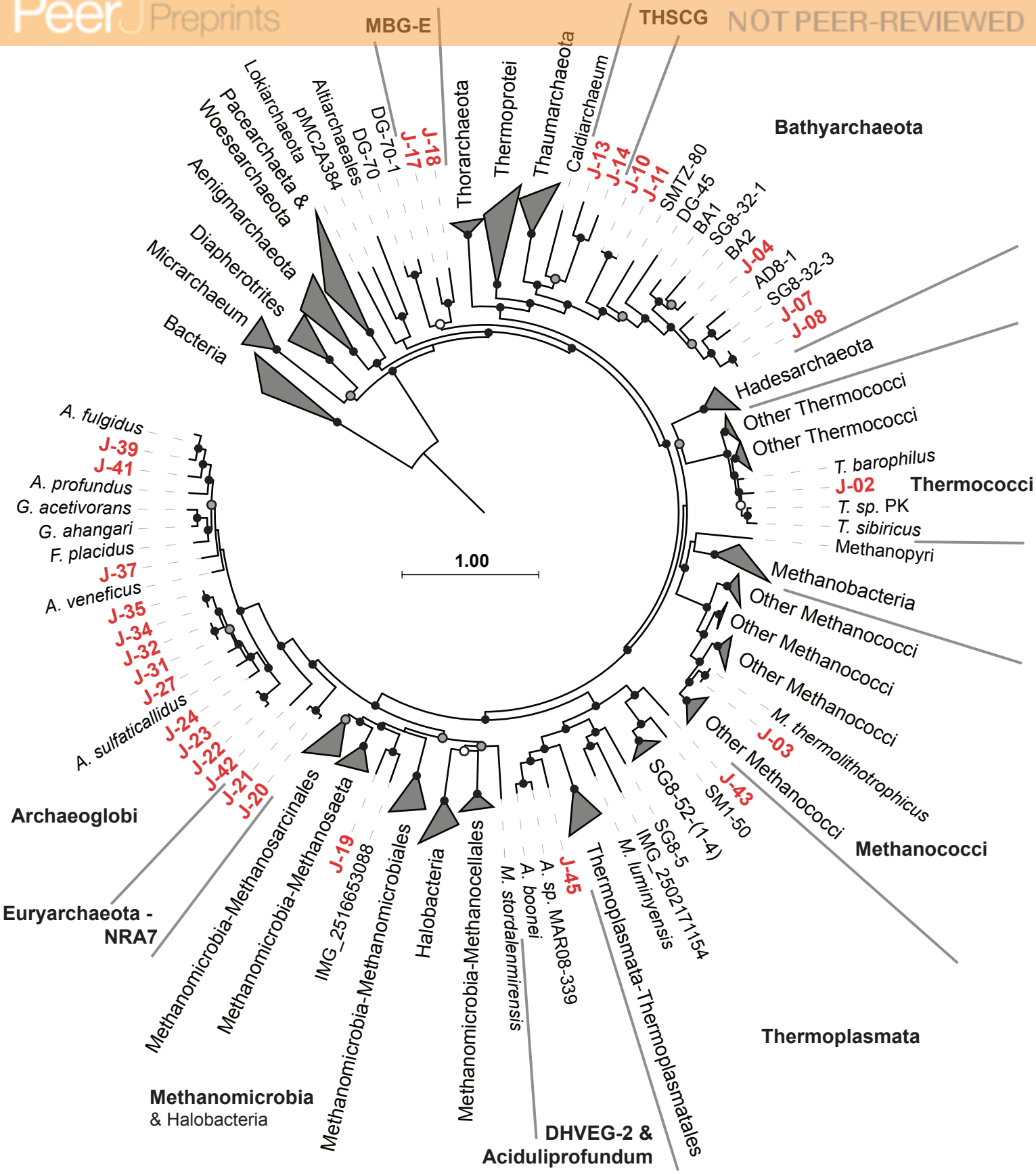
**Genome refinement**
Anvi'o

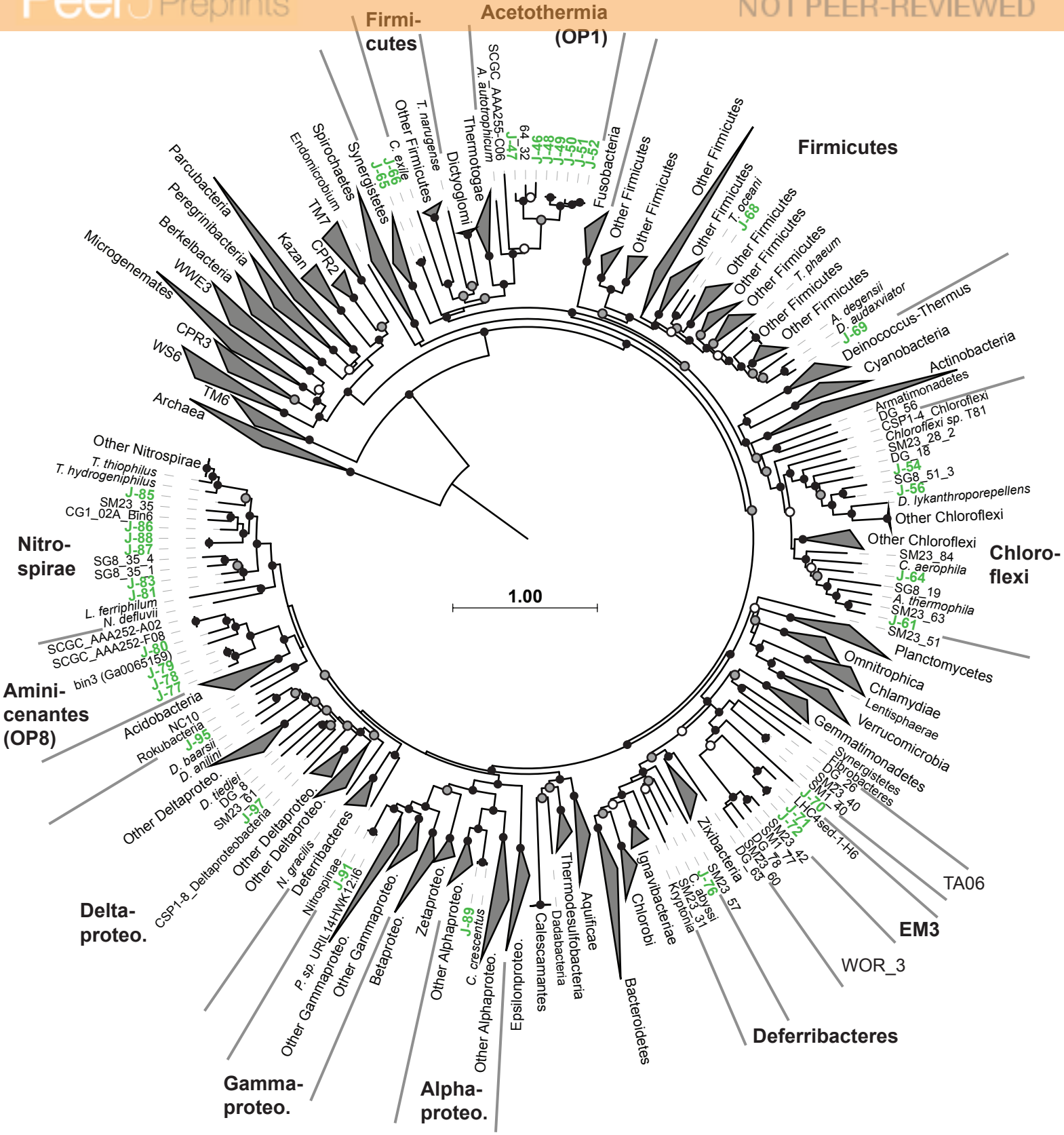
↓

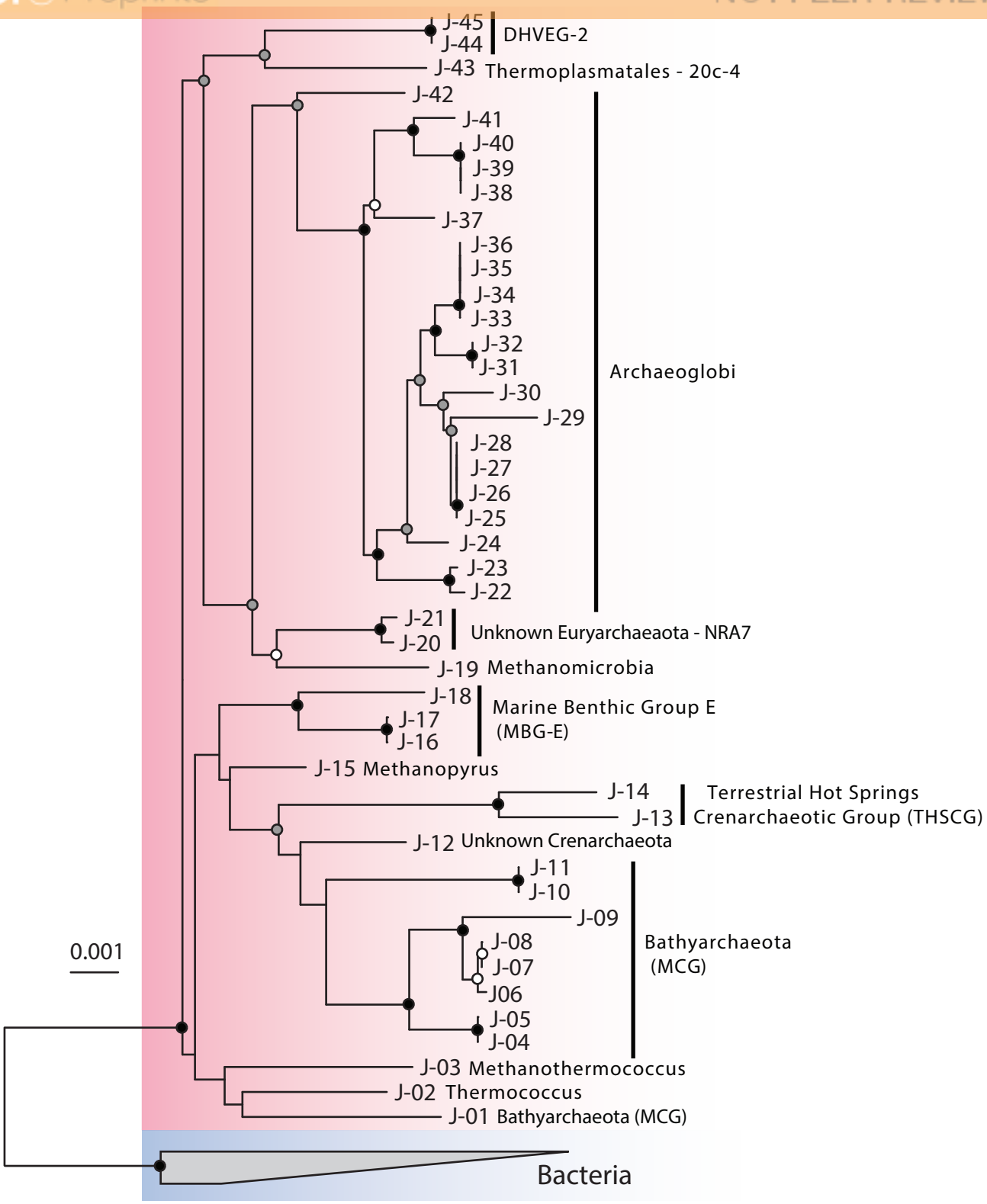
**Annotation**
IMG/ER

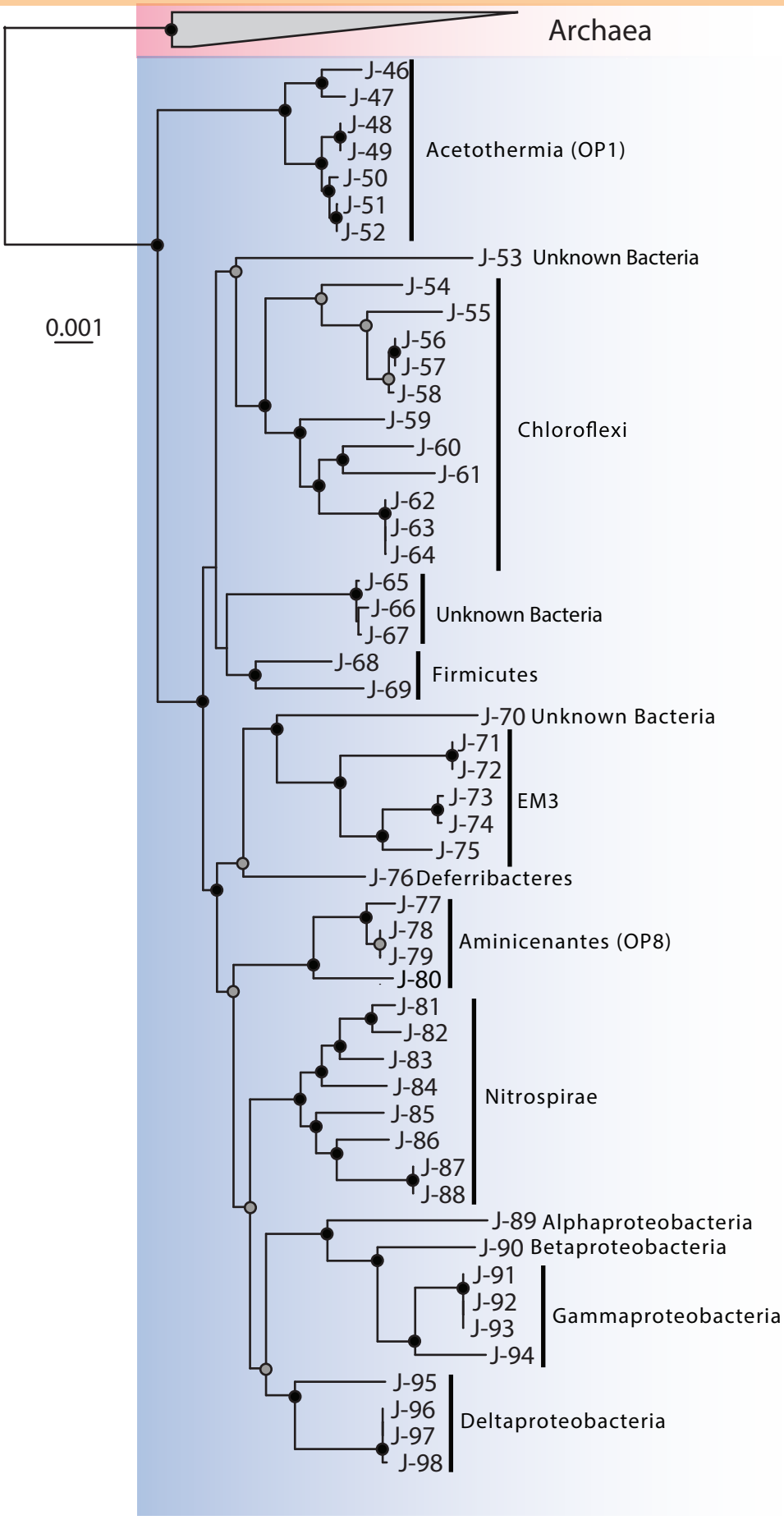
↓

**Phylogenome analysis**
Inferred from single-copy genes extracted with CheckM

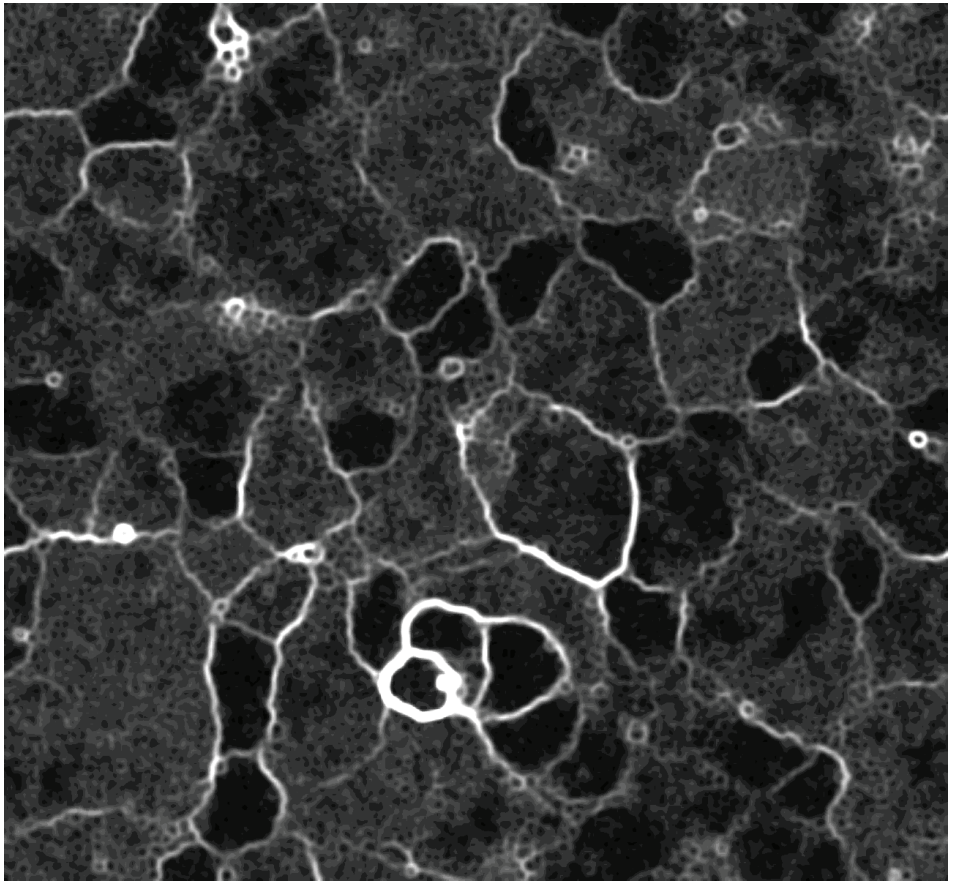




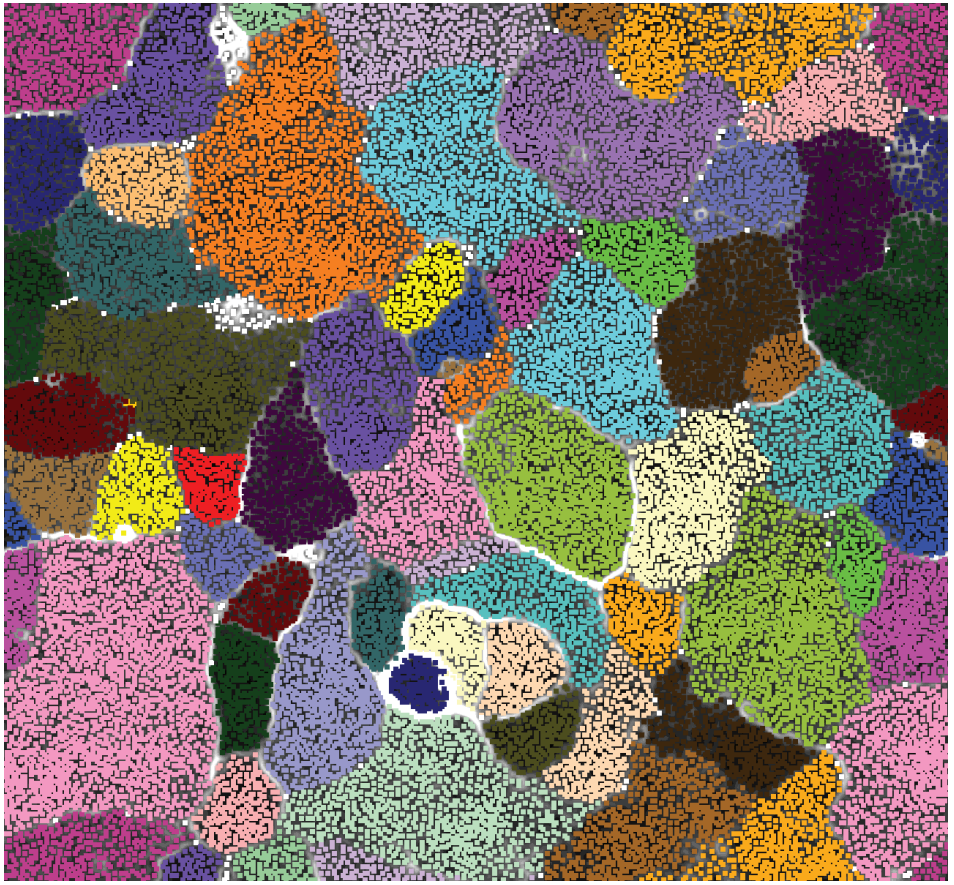




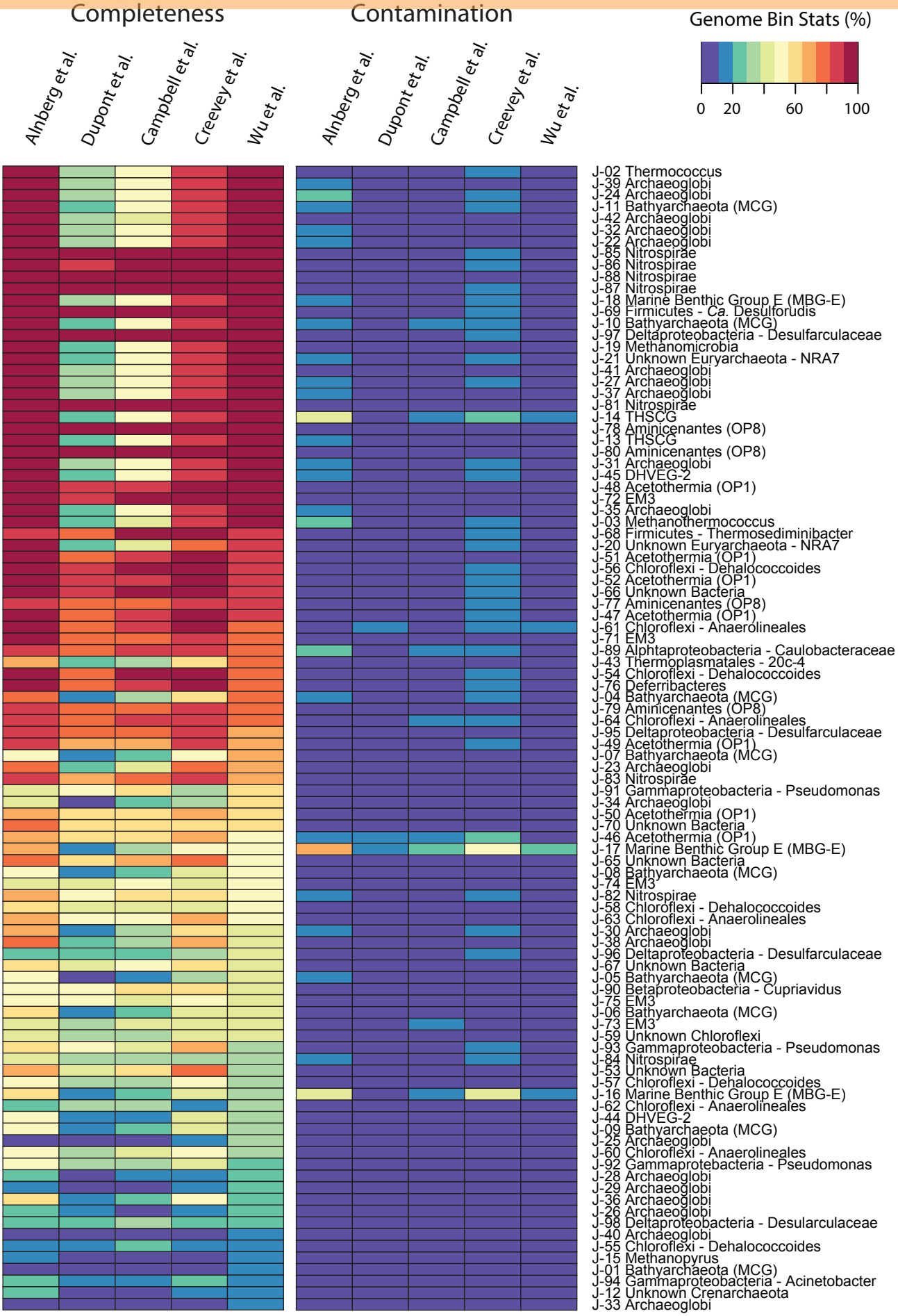
A)



B)



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:




PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2613v1 | CC BY 4.0 Open Access | rec: 1 Dec 2016, publ:

Table 1. Metagenome sequencing statistics reported in IMG

| | *U1362A* | | | *U1362B* | | |
|---|---|---|---|---|---|---|
| | no. assembled (% of assembled) | *no. unassembled (% of unassembled)* | *total (% of total)* | *no. assembled (% of assembled)* | *no. unassembled (% of unassembled)* | *total (% of total)* |
| **Number of sequences** | 137575 (8.08) | 1564185 (91.92) | 1701760 (100) | 212307 (7.60) | 2582305 (92.40) | 2794612 (100) |
| **Number of bases** | 169908118 (33.78) | 333077167 (66.22) | 502985285 (100) | 168044831 (23.83) | 537213224 (76.17) | 705258055 (100) |
| GC count | 82941377 (48.82) | 163998454 (49.24) | 246939831 (49.09) | 87552944 (52.10) | 270739112 (50.40) | 358292056 (50.80) |
| **Genes** | | | | | | |
| rRNA genes | 609 (0.22) | 1124 (0.08) | 1733 (0.10) | 682 (0.21) | 1219 (0.05) | 1901 (0.07) |
| *16S rRNA* | 198 (0.07) | 162 (0.01) | 360 (0.02) | 199 (0.06) | 191 (0.01) | 390 (0.01) |
| *23S rRNA* | 315 (0.12) | 617 (0.04) | 932 (0.05) | 359 (0.11) | 587 (0.02) | 946 (0.04) |
| Protein coding genes | 267511 (98.50) | 1489984 (99.63) | 1757495 (99.46) | 319764 (98.87) | 2344253 (99.37) | 2664017 (99.31) |
| *with Product Name* | 160006 (58.91) | 438495 (29.32) | 598501 (33.87) | 170964 (52.86) | 559698 (23.73) | 730662 (27.24) |
| with COG | 186319 (68.60) | 675287 (45.16) | 861606 (48.76) | 207169 (64.06) | 834581 (35.38) | 1041750 (38.84) |
| with Pfam | 172149 (63.38) | 519243 (34.72) | 691392 (39.13) | 187717 (58.04) | 647505 (27.45) | 835222 (31.14) |
| with KO | 131624 (48.46) | 604486 (40.42) | 736110 (41.66) | 151186 (46.75) | 773722 (32.80) | 924908 (34.48) |
| with Enzyme (EC) | 73927 (27.22) | 356052 (23.81) | 429979 (24.33) | 83086 (25.69) | 440214 (18.66) | 523300 (19.51) |
| with MetaCyc | 52288 (19.25) | 244997 (16.38) | 297285 (16.82) | 58809 (18.18) | 301799 (12.79) | 360608 (13.44) |
| with KEGG | 78361 (28.85) | 365246 (24.42) | 443607 (25.10) | 88171 (27.26) | 455581 (19.31) | 543752 (20.27) |

Table 2. Metagenome scaffold length statistics

| *Minimum scaffold length* | *U1362A* | | *U1362B* | |
|---|---|---|---|---|
| | *Num. of Scaffolds*[a] | *Total Scaffold Length*[a] | *Num. of Scaffolds*[a] | *Total Scaffold Length*[a] |
| All | 137575 | 169908118 | 212307 | 168044831 |
| 1 kb | 25958 | 122371000 | 22179 | 94767619 |
| 2.5 kb | 10118 | 98145686 | 7817 | 72903412 |
| 5 kb | 4544 | 78915922 | 3232 | 57281039 |
| 10 kb | 1933 | 60882353 | 1339 | 44376823 |
| 25 kb | 615 | 41195243 | 435 | 30631998 |
| 50 kb | 273 | 29394283 | 191 | 22129275 |
| 100 kb | 105 | 18147775 | 72 | 13983109 |
| 250 kb | 15 | 5160259 | 11 | 5597623 |
| 500 kb | 1 | 540961 | 3 | 2801775 |
| 1 mb | 0 | 0 | 1 | 1136825 |

[a]Numbers listed are the cumulative sum of all scaffolds equal to or above the scaffold length

Table 3. Genome binning method summary

| *Method* | *Num Bins* | *Num Bins >10% Complete* | *Num Bins >50% Complete* | *Avg. Completeness (%)*[a] | *Avg. Contamination (%)*[a] |
|---|---|---|---|---|---|
| CONCOCT | 66 | 56 | 46 | 90.9 | 50.8 |
| ESOM | 60 | 54 | 49 | 90.4 | 71.5 |
| MaxBin2 | 75 | 66 | 51 | 85.7 | 42.9 |
| MetaBAT | 69 | 64 | 45 | 87.7 | 9.7 |
| **CONCOCT (post manual curation in Anvi'o)** | **252** | **98** | **61** | **84.4** | **3.3** |

[a]Average calculated for bins >50% completeness

Table 4. Summary of genomes from metagenomes (GFMs)

| Bin | Taxonomy | Size (Kbp) | Contigs | Genes | N50 | %GC |
|---|---|---|---|---|---|---|
| JdFR-01 | Bathyarchaeota (MCG) | 200 | 53 | 262 | 3811 | 38.2 |
| JdFR-02 | Thermococcus | 2362 | 44 | 2708 | 135395 | 38.4 |
| JdFR-03 | Methanothermococcus | 1476 | 164 | 1639 | 11655 | 33.0 |
| JdFR-04 | Bathyarchaeota (MCG) | 1068 | 204 | 1345 | 5722 | 41.9 |
| JdFR-05 | Bathyarchaeota (MCG) | 433 | 101 | 572 | 4308 | 42.4 |
| JdFR-06 | Bathyarchaeota (MCG) | 735 | 72 | 920 | 24679 | 38.1 |
| JdFR-07 | Bathyarchaeota (MCG) | 908 | 43 | 1061 | 33770 | 39.1 |
| JdFR-08 | Bathyarchaeota (MCG) | 919 | 45 | 1079 | 28674 | 38.6 |
| JdFR-09 | Bathyarchaeota (MCG) | 499 | 16 | 621 | 39478 | 39.1 |
| JdFR-10 | Bathyarchaeota (MCG) | 1818 | 57 | 2051 | 94071 | 51.4 |
| JdFR-11 | Bathyarchaeota (MCG) | 1726 | 12 | 1932 | 251681 | 52.0 |
| JdFR-12 | Unknown Crenarch | 301 | 65 | 372 | 3767 | 36.1 |
| JdFR-13 | THSCG | 1672 | 4 | 1780 | 488929 | 37.6 |
| JdFR-14 | THSCG | 1635 | 6 | 1722 | 462362 | 41.9 |
| JdFR-15 | Methanopyrus | 208 | 54 | 273 | 3688 | 36.0 |
| JdFR-16 | MBG-E | 1353 | 241 | 1714 | 6267 | 50.4 |
| JdFR-17 | MBG-E | 2178 | 344 | 2766 | 7687 | 50.1 |
| JdFR-18 | MBG-E | 2062 | 22 | 2328 | 149032 | 39.1 |
| JdFR-19 | Methanomicrobia | 1289 | 27 | 1594 | 78121 | 43.2 |
| JdFR-20 | Unknown Euryarch - NRA7 | 1610 | 209 | 2109 | 8881 | 41.1 |
| JdFR-21 | Unknown Euryarch - NRA7 | 1417 | 22 | 1724 | 102423 | 41.2 |
| JdFR-22 | Archaeoglobi | 2063 | 130 | 2360 | 20363 | 39.7 |
| JdFR-23 | Archaeoglobi | 1629 | 101 | 1924 | 25841 | 40.0 |
| JdFR-24 | Archaeoglobi | 2702 | 154 | 3011 | 48758 | 38.2 |
| JdFR-25 | Archaeoglobi | 885 | 31 | 1007 | 65531 | 39.9 |
| JdFR-26 | Archaeoglobi | 645 | 21 | 713 | 141267 | 40.2 |
| JdFR-27 | Archaeoglobi | 2360 | 67 | 2680 | 82469 | 40.6 |
| JdFR-28 | Archaeoglobi | 752 | 14 | 829 | 97698 | 40.4 |
| JdFR-29 | Archaeoglobi | 738 | 185 | 967 | 3748 | 44.9 |
| JdFR-30 | Archaeoglobi | 1100 | 55 | 1276 | 33048 | 39.5 |
| JdFR-31 | Archaeoglobi | 2351 | 103 | 2752 | 43068 | 42.1 |
| JdFR-32 | Archaeoglobi | 1967 | 120 | 2225 | 24104 | 41.8 |
| JdFR-33 | Archaeoglobi | 479 | 89 | 593 | 5918 | 40.3 |
| JdFR-34 | Archaeoglobi | 1088 | 88 | 1246 | 18271 | 41.3 |
| JdFR-35 | Archaeoglobi | 1703 | 167 | 2029 | 14968 | 41.2 |
| JdFR-36 | Archaeoglobi | 581 | 75 | 732 | 9486 | 40.6 |
| JdFR-37 | Archaeoglobi | 1972 | 52 | 2248 | 64398 | 44.7 |
| JdFR-38 | Archaeoglobi | 1025 | 99 | 1208 | 13482 | 44.5 |
| JdFR-39 | Archaeoglobi | 2279 | 93 | 2743 | 55279 | 43.8 |
| JdFR-40 | Archaeoglobi | 530 | 76 | 671 | 8743 | 42.8 |
| JdFR-41 | Archaeoglobi | 1752 | 103 | 1921 | 25992 | 42.3 |
| JdFR-42 | Archaeoglobi | 2149 | 42 | 2479 | 70809 | 40.3 |
| JdFR-43 | Thermoplasmatales - 20c-4 | 1231 | 219 | 1425 | 6292 | 37.5 |
| JdFR-44 | DHVEG-2 | 519420 | 139 | 527 | 3618 | 55.3 |
| JdFR-45 | DHVEG-2 | 1249758 | 161 | 1454 | 9500 | 57.9 |
| JdFR-46 | Acetothermia (OP1) | 1088071 | 254 | 1259 | 4279 | 65.9 |
| JdFR-47 | Acetothermia (OP1) | 1622290 | 230 | 1811 | 8291 | 63.7 |
| JdFR-48 | Acetothermia (OP1) | 1893129 | 62 | 1992 | 56879 | 61.1 |
| JdFR-49 | Acetothermia (OP1) | 1698648 | 312 | 2003 | 5976 | 59.9 |
| JdFR-50 | Acetothermia (OP1) | 1292279 | 255 | 1516 | 5492 | 62.2 |
| JdFR-51 | Acetothermia (OP1) | 1763844 | 185 | 1919 | 13057 | 63.0 |
| JdFR-52 | Acetothermia (OP1) | 1710598 | 155 | 1870 | 17111 | 62.7 |
| JdFR-53 | Unknown Bacteria | 346303 | 76 | 434 | 4902 | 38.4 |

| | | | | | | |
|---|---|---|---|---|---|---|
| JdFR-54 | Chloroflexi - Dehalococcoides | 1691297 | 67 | 1743 | 54623 | 59.9 |
| JdFR-55 | Chloroflexi - Dehalococcoides | 545773 | 132 | 599 | 4081 | 62.8 |
| JdFR-56 | Chloroflexi - Dehalococcoides | 1798153 | 95 | 1964 | 28676 | 57.1 |
| JdFR-57 | Chloroflexi - Dehalococcoides | 776292 | 207 | 951 | 3716 | 57.6 |
| JdFR-58 | Chloroflexi - Dehalococcoides | 908200 | 203 | 1152 | 4219 | 57.4 |
| JdFR-59 | Unknown Chloroflexi | 2039159 | 527 | 2194 | 3822 | 61.0 |
| JdFR-60 | Chloroflexi - Anaerolineales | 863183 | 239 | 992 | 3541 | 64.5 |
| JdFR-61 | Chloroflexi - Anaerolineales | 2906759 | 380 | 2909 | 8915 | 64.2 |
| JdFR-62 | Chloroflexi - Anaerolineales | 906207 | 183 | 1068 | 5571 | 52.3 |
| JdFR-63 | Chloroflexi - Anaerolineales | 1318067 | 282 | 1521 | 5066 | 52.4 |
| JdFR-64 | Chloroflexi - Anaerolineales | 2358376 | 468 | 2680 | 5560 | 52.6 |
| JdFR-65 | Unknown Bacteria | 936826 | 193 | 1101 | 5093 | 42.2 |
| JdFR-66 | Unknown Bacteria | 1838550 | 224 | 2056 | 11163 | 40.4 |
| JdFR-67 | Unknown Bacteria | 873684 | 172 | 978 | 5273 | 41.0 |
| JdFR-68 | Firmicutes - Thermosediminibacter | 2975652 | 458 | 3396 | 7798 | 39.7 |
| JdFR-69 | Firmicutes – "*Can*. Desulforudis" | 1779991 | 36 | 1865 | 111790 | 61.1 |
| JdFR-70 | Unknown Bacteria | 921100 | 214 | 596 | 4521 | 35.8 |
| JdFR-71 | EM3 | 1702039 | 18 | 1729 | 142680 | 34.8 |
| JdFR-72 | EM3 | 2059798 | 19 | 2026 | 176033 | 34.7 |
| JdFR-73 | EM3 | 1305425 | 257 | 1371 | 5573 | 32.8 |
| JdFR-74 | EM3 | 1800622 | 267 | 1861 | 8536 | 32.9 |
| JdFR-75 | EM3 | 789045 | 171 | 979 | 4417 | 32.1 |
| JdFR-76 | Deferribacteres | 3099908 | 557 | 3087 | 5995 | 52.3 |
| JdFR-77 | Aminicenantes (OP8) | 2326237 | 200 | 2380 | 16386 | 30.5 |
| JdFR-78 | Aminicenantes (OP8) | 2530072 | 42 | 2463 | 114879 | 32.5 |
| JdFR-79 | Aminicenantes (OP8) | 2046026 | 32 | 1995 | 94220 | 32.7 |
| JdFR-80 | Aminicenantes (OP8) | 2914703 | 113 | 2726 | 57497 | 44.5 |
| JdFR-81 | Nitrospirae | 2050278 | 95 | 2117 | 39421 | 48.0 |
| JdFR-82 | Nitrospirae | 735991 | 154 | 875 | 5174 | 43.8 |
| JdFR-83 | Nitrospirae | 1165085 | 177 | 1310 | 8103 | 42.1 |
| JdFR-84 | Nitrospirae | 1526138 | 342 | 1833 | 4439 | 39.9 |
| JdFR-85 | Nitrospirae | 2325903 | 47 | 2399 | 78042 | 41.4 |
| JdFR-86 | Nitrospirae | 2103466 | 25 | 2166 | 124076 | 45.2 |
| JdFR-87 | Nitrospirae | 1861077 | 58 | 2005 | 51364 | 62.5 |
| JdFR-88 | Nitrospirae | 1858353 | 22 | 1983 | 131863 | 62.8 |
| JdFR-89 | Caulobacteraceae (Alphaprot.) | 4055294 | 650 | 4284 | 7375 | 67.8 |
| JdFR-90 | Cupriavidus (Betaprot.) | 2666033 | 703 | 3105 | 3723 | 62.8 |
| JdFR-91 | Pseudomonas (Gammaprot.) | 3770010 | 358 | 3583 | 15126 | 60.9 |
| JdFR-92 | Pseudomonas (Gammaprot.) | 1611073 | 457 | 1865 | 3440 | 59.3 |
| JdFR-93 | Pseudomonas (Gammaprot.) | 2230228 | 225 | 2227 | 13223 | 59.8 |
| JdFR-94 | Acinetobacter (Gammaprot.) | 368932 | 117 | 478 | 3101 | 37.9 |
| JdFR-95 | Desulfarculaceae (Deltaprot.) | 2807742 | 499 | 2872 | 6282 | 68.3 |
| JdFR-96 | Desulfarculaceae (Deltaprot.) | 1392833 | 312 | 1551 | 4552 | 58.3 |
| JdFR-97 | Desulfarculaceae (Deltaprot.) | 4102562 | 127 | 3814 | 54652 | 57.0 |
| JdFR-98 | Desulfarculaceae (Deltaprot.) | 932589 | 224 | 1057 | 3945 | 57.7 |