# Reconstruction of metagenome-assembled genomes from aquaria

Cassandra L. Ettinger[a,b,c,#], Jordan Bryan[d], Sima Tokajian[e], Guillaume Jospin[a,f], David Coil[a], Jonathan A. Eisen[a,b,g]

[a]Genome Center, University of California, Davis, CA, USA
[b]Department of Evolution and Ecology, University of California, Davis, CA, USA
[c]Department of Microbiology & Plant Pathology, University of California, Riverside, CA, USA
[d]College of Agriculture and Life Sciences, Cornell University, Ithaca, NY, USA
[e]Department of Natural Sciences, Lebanese American University, Byblos, Lebanon
[f]AnimalBiome, Oakland, CA, USA
[g]Department of Medical Microbiology and Immunology, University of California, Davis, CA, USA

[#]**Address correspondence to:** Cassandra L. Ettinger, cassande@ucr.edu

## Abstract

We report 11 metagenome-assembled genomes (MAGs) reconstructed from freshwater and saltwater aquaria including representatives of Polynucleobacter, Anaerolinae, Roseobacter, Flavobacteriia, Octadecabacter, Mycobacterium and Candidate Phyla Radiation (CPR) members. These MAGs can serve as a resource for aquatic research and elucidating the role of CPR taxa in the built environment.

**Keywords:** aquarium, metagenomics, metagenome assembled genomes, candidate phyla radiation

## Announcement

Microbial communities play critical roles in aquaria health. Aquaria support complex multi-trophic interactions between fish, invertebrates, plants and microbial communities that occur in an enclosed built environment. Understanding the genomics of aquaria microbial communities is critical to understanding the health of other enclosed aquatic systems.

Samples were collected prior to the start of an undergraduate research project that investigated microbial community assembly of multiple aquaria in the Fall of 2012 at the University of California, Davis (1). Tropical tank sediment (n=3), cold reef tank sediment (n=1), freshwater tank sediment (n=3), cold reef tank water (n=3), freshwater wipes (n=3) and freshwater tank water (n=3) were collected and processed for DNA extraction as described in Bik et al (1). Libraries were made using a Nextera XT DNA library sample preparation kit (Illumina, Inc.) and were sequenced on an Illumina MiSeq (paired end, 150 bp reads).

All reads were co-assembled using MEGAHIT (2) v.1.0.6. Metagenome-assembled genomes (MAGs) were generated using anvi'o v. 2.3.2 (3). First, a contig database was produced using 'anvi-gen-contigs-database' and open reading frames identified with Prodigal (4) v.2.6.2. We then used 'anvi-run-hmms' to run HMMER v3.1b2 (5) to identify bacterial (6) and archeal (7) single-copy genes. Contig taxonomy was inferred using Kaiju v.1.5.0 (8) with the NCBI BLAST non-redundant protein database including fungi and microbial eukaryotes v.2016-09-18. Reads were mapped using Bowtie2 v.2.2.8 (9) and samtools v.1.4.1 (10). Using 'anvi-profile' and 'anvi-merge', contigs > 2.5 kbp were mapped to samples and then profiles were combined. On average, 780,565 reads per sample mapped to the contig database with the majority of mapped reads from cold reef tank water (57.3%) and freshwater tank water (41.7%). Contigs were clustered using 'anvi-cluster-with-concoct' to automatically bin MAGs (11). MAG completeness and contamination was assessed in anvi'o using 'anvi-summarize' and confirmed with CheckM v.1.0.7 (12). Phylosift v. 1.0.1 (13) was used to place MAGs in a phylogenetic context to provide additional information about  taxonomic assignments. Candidate Phyla Radiation (CPR) genomes were identified with 'anvi-script-gen-cpr-classifier' and 'anvi-script-predict-cpr-genomes' using the Brown et al (14) and Cambell et al (6) databases. CPR genome completion was then estimated for 43 single copy marker genes (14). The CPR are putatively a diverse group of uncultured bacterial lineages with poorly understood metabolic functions known mostly from metagenomic sequencing work. Representatives of CPR have been previously found in a wide range of aquatic habitats including bioreactors, ocean, lakes, groundwater and waterways (15–21).

We report two high-quality draft MAGs >90% completion and four medium-quality draft MAGs with >70% completion (Table 1). Additionally, we report five draft MAGs that were identified as potential CPR genomes with >90% completion (Table 2). These metagenome-assembled genomes will enable deeper insights into the ecology of aquaria microbial communities and also into the possible functional roles of understudied lineages (e.g. CPR members) in the built environment.

**Data availability.** The raw sequencing reads, co-assembly and individual MAGs were deposited at DDBJ/ENA/GenBank under BioProject accession number PRJNA728121. Contigs identified as possible contaminants or adaptors by NCBI's Contamination Screen were subsequently trimmed or removed from the co-assembly or individual MAGs during deposition.

## Acknowledgements

Illumina sequencing was performed at the DNA Technologies Core facility in the Genome Center at the University of California, Davis, Davis, California (UCD). We acknowledge the contributions of multiple people for help in various stages of the project, especially Thein Mai and Matt Wein for access to and assistance with the aquaria used in this study. Funding for this study was provided by the Alfred P. Sloan Foundation through their program in the "Microbiology of the Built Environment".

## References

1. Bik HM, Alexiev A, Aulakh SK, Bharadwaj L, Flanagan J, Haggerty JM, Hird SM, Jospin G, Lang JM, Sauder LA, Neufeld JD, Shaver A, Sethi A, Eisen JA, Coil DA. 2019. Microbial Community Succession and Nutrient Cycling Responses following Perturbations of Experimental Saltwater Aquaria. mSphere 4.

2. Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W. 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 102:3–11.

3. Eren AM, Murat Eren A, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319.

4. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.

5. Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput Biol 7:e1002195.

6. Campbell JH, O'Donoghue P, Campbell AG, Schwientek P, Sczyrba A, Woyke T, Soll D, Podar M. 2013. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. Proceedings of the National Academy of Sciences 110:5540–5545.

7. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu W-T, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437.

8. Menzel P, Ng KL, Krogh A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 7:11257.

9. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

10. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

11. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ,

Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nat Methods 11:1144–1146.

12. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055.

13. Darling AE, Jospin G, Lowe E, Matsen FA 4th, Bik HM, Eisen JA. 2014. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ 2:e243.

14. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523:208–211.

15. Cabello-Yeves PJ, Zemskaya TI, Zakharenko AS, Sakirko MV, Ivanov VG, Ghai R, Rodriguez-Valera F. 2020. Microbiome of the deep Lake Baikal, a unique oxic bathypelagic habitat. Limnology and Oceanography.

16. Ruuskanen MO, Colby G, St. Pierre KA, St. Louis VL, Aris-Brosou S, Poulain AJ. 2020. Microbial genomes retrieved from High Arctic lake sediments encode for adaptation to cold and oligotrophic environments. Limnology and Oceanography.

17. Thrash JC, Seitz KW, Baker BJ, Temperton B, Gillies LE, Rabalais NN, Henrissat B, Mason OU. 2017. Metabolic Roles of Uncultivated Bacterioplankton Lineages in the Northern Gulf of Mexico “Dead Zone.” MBio 8.

18. Zhao Y, Liu S, Jiang B, Feng Y, Zhu T, Tao H, Tang X, Liu S. 2018. Genome-Centered

Metagenomics Analysis Reveals the Symbiotic Organisms Possessing Ability to Cross-Feed with Anammox Bacteria in Anammox Consortia. Environ Sci Technol 52:11285–11296.

19. Vigneron A, Cruaud P, Langlois V, Lovejoy C, Culley AI, Vincent WF. 2020. Ultra-small and abundant: Candidate phyla radiation bacteria are potential catalysts of carbon transformation in a thermokarst lake ecosystem. Limnology and Oceanography Letters.

20. Danczak RE, Johnston MD, Kenah C, Slattery M, Wrighton KC, Wilkins MJ. 2017. Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. Microbiome 5:112.

21. León-Zayas R, Peoples L, Biddle JF, Podell S, Novotny M, Cameron J, Lasken RS, Bartlett DH. 2017. The metabolic potential of the single cell genomes obtained from the Challenger Deep, Mariana Trench within the candidate superphylum Parcubacteria (OD1). Environ Microbiol 19:2769–2784.

22. Delmont TO, Quince C, Shaiber A, Esen OC, Lee STM, Lucker S, Murat Eren A. 2017. Nitrogen-Fixing Populations Of Planctomycetes And Proteobacteria Are Abundant In The Surface Ocean.

23. Galperin MY, Makarova KS, Wolf YI, Koonin EV. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res 43:D261–9.

24. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60.

25. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu W-T, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Genome Standards Consortium, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35:725–731.

**Table 1.** Genomic feature summary for metagenome-assembled genomes identified from aquaria.

| **Bin ID** | **Draft quality** | **Putative taxonomy** | **Genome Size (Mb)** | **%GC** | **CheckM % Completion** | **CheckM % Redundancy** | **anvi'o % Completion** | **anvi'o % Redundancy** |
|---|---|---|---|---|---|---|---|---|
| **AQU-01** | High | *Polynucleobacter* sp. | 1.68 | 45.38 | 97.69 | 0.16 | 98.56 | 0.72 |
| **AQU-02** | High | Anaerolineae sp. | 5.35 | 53.21 | 91.36 | 1.73 | 95.68 | 2.88 |
| **AQU-03** | Medium | *Roseobacter* sp. | 2.69 | 60.69 | 88.80 | 0.63 | 75.54 | 2.16 |
| **AQU-04** | Medium | Flavobacteriia sp. | 1.85 | 41.36 | 86.62 | 0.07 | 92.81 | 1.44 |
| **AQU-05** | Medium | *Octadecabacter* sp. | 2.61 | 55.83 | 84.33 | 1.96 | 76.98 | 1.44 |
| **AQU-06** | Medium | *Mycobacterium* sp. | 3.35 | 66.59 | 70.94 | 1.74 | 71.94 | 2.16 |

**Table 2.** Genomic feature summary for Candidate Phyla Radiation (CPR) metagenome-assembled genomes identified from aquaria.

| **Bin ID** | **Putative taxonomy** | **Genome Size (Mb)** | **%GC** | **CPR % Completion**[a] |
|---|---|---|---|---|
| **AQU-07** | Candidatus Shapirobacteria sp. | 0.86 | 35.61 | 93.02 |
| **AQU-08** | Candidatus Kerfeldbacteria sp. | 1.07 | 46.20 | 93.02 |
| **AQU-09** | Candidatus Uhrbacteria sp. | 1.12 | 51.31 | 93.02 |
| **AQU-10** | Candidatus Moranbacteria sp. | 0.89 | 43.56 | 93.02 |
| **AQU-11** | Candidatus Saccharibacteria sp. | 1.09 | 47.69 | 90.70 |

[a] Completion estimates were generated using 43 single copy markers for CPR following Brown et al (14).