ZetaHunter Documentation

Sean M. McAllister, Ryan M. Moore, and Clara S. Chan

University of Delaware
Contact: zetahunter.help@gmail.com

Please site: doi:xxxxx/xxxxx

Summary

The Zetaproteobacteria are a poorly resolved taxonomic group, primarily lumped into a single classification by most common classification tools. This low-resolution taxonomy results in the inability to describe or discern ecological roles and niches. Therefore, we designed the ZetaHunter program for the reproducible assignment of small subunit ribosomal RNA (16S rRNA) gene sequences to a curated database of previously-described Zetaproteobacterial operational taxonomic units (ZOTUs). A stable classification tool will allow researchers to assess ZOTU distribution in the context of prior research. Furthermore, ZetaHunter can be used with any curated 16S rRNA sequence database, such as in an example given for the Omnitrophica (OP3) candidate phylum.

I. Introduction

Taxonomic groups with limited cultivated representatives are often lumped into a single taxonomic label, which includes multiple distinct ecological units. This makes it difficult to discuss these groups in an ecological context and is particularly problematic within the extensive uncultivated microbial groups in the Candidate Phyla Radiation (CPR) as well as other poorly represented taxa [1]. One class with few cultivated representatives is the Zetaproteobacteria, from which only three distinct genera have been isolated [2, 3, 4]. Researchers using standard 16S rRNA classification tools (SILVA, RDP, Greengenes) are shown the classification to the nearest defined taxonomy, which only exists for genus *Mariprofundus* and is found to represent a limited proportion (3.8%) of Zetaproteobacteria sequences found in the environment [5]. As a result of this taxonomy, researchers can only accurately classify the group at the class level.

OTU comparisons are often difficult to reproduce between researchers and over multiple studies. As a result, OTU definitions are usually abandoned. However, using a dataset with a defined, high-resolution taxonomy, those definitions can be preserved and reproduced, allowing for ecological comparisons across studies. Here, we introduce the ZetaHunter program, which can reproducibly assign novel 16S rRNA sequences to any curated taxonomic framework. The program was designed for and comes with a curated database to identify the members of the Zetaproteobacteria, though it may be used with any curated 16S rRNA database. The use of this program will allow researchers to explicitly discuss ZOTU abundance, significance, and niche preference between studies for the Zetaproteobacteria and other microbes with poorly defined taxonomy. Indeed, ZetaHunter has already been used by researchers for classifying Zetaproteobacteria isolates [3, 6], classifying Zetaproteobacteria within 16S rRNA surveys [7, 8], classifying Zetaproteobacteria ecotypes determined from minimum entropy decomposition [9, 10], and in a review of Zetaproteobacteria physiology, ecology, and genomics [5]. Below we discuss

the ZetaHunter pipeline, curated database, and some usage examples.

II. ZetaHunter program

The Zetaproteobacteria are represented by only a few cultured isolates. Since the Zetaproteobacteria lack cultured representatives from a majority of the sequenced biodiversity, a reproducible taxonomic classification method is needed to compare diversity across samples and studies. For this purpose, we introduce the ZetaHunter program. ZetaHunter is a command line program (written in Ruby) designed to assign user-supplied 16S rRNA gene sequences to OTUs defined by a reference sequence database. ZetaHunter can be used on the Linux, MacOSX, and Windows platforms through the use of a Docker container, or through installation from source (Linux and MacOSX only). By default, ZetaHunter uses a curated database of Zetaproteobacteria 16S rRNA genes (described below). Zetaproteobacteria OTU definitions are the same as those suggested by McAllister et al. [11] at 97% identity, preserving the order of the ZOTU numbering for ease in comparisons across studies. Only 28 ZOTUs were discovered in this initial publication, meaning that numbered ZOTUs from ZOTU29 to ZOTU59 (as of database v.3) are newly discovered and have been added to the ZetaHunter database with each update to the program. Input fasta files for ZetaHunter must contain sequences aligned by the SINA aligner [12]. After input, these alignments are masked using the same 1,282 bp mask used in McAllister et al. [11] to obtain reproducible ZOTU calls through closed reference OTU binning. User sequences that represent novel ZOTUs are *de novo* binned into "NewZetaOTUs", numbered by abundance. Care should be taken in interpreting these novel OTUs, particularly if they represent only a single sequence. Singletons may be real low-abundance Zetaproteobacteria, or they may be the result of sequencing error, chimeric reads, or poor sequence alignment [13].

The **default pipeline of ZetaHunter** (Figure 1) takes SINA-aligned 16S rRNA gene sequences and processes them in this order:

1. Sequences are checked for chimeras using mothur's uchime algorithm [14, 15].
2. SortmeRNA is used to cluster new sequences to the reference database based on genetic distance (closed reference binning; [16]).
3. The remaining sequences are clustered into novel OTUs by average neighbor OTU clustering with mothur [14].
4. Summary, biom (table showing OTU abundance per sample), and OTU network files are exported for use.

This pipeline is an implementation of open-reference OTU picking similar to the one found in QIIME [17].

To use ZetaHunter for a **custom database** (example given in Section V), all that is needed are:

1. a set of SINA-aligned 16S rRNA gene sequences to be used as the database for the closed reference OTU picking step
2. a sequence mask with asterisks at each alignment column that should be used for taxonomic assignment (i.e. unambiguously aligned base positions trimmed to the length of the user's references)
3. a tab-delimited file assigning each sequence in the new database to a particular taxonomic group at the similarity threshold desired by the user [This can be determined for the reference database through clustering with mothur (97% similarity recommended; [14]).]

ZetaHunter is available for download at <https://github.com/mooreryan/ZetaHunter>.

III. The Zetaproteobacteria curated database

The backbone of the ZetaHunter program is the curated database, which assigns sequences

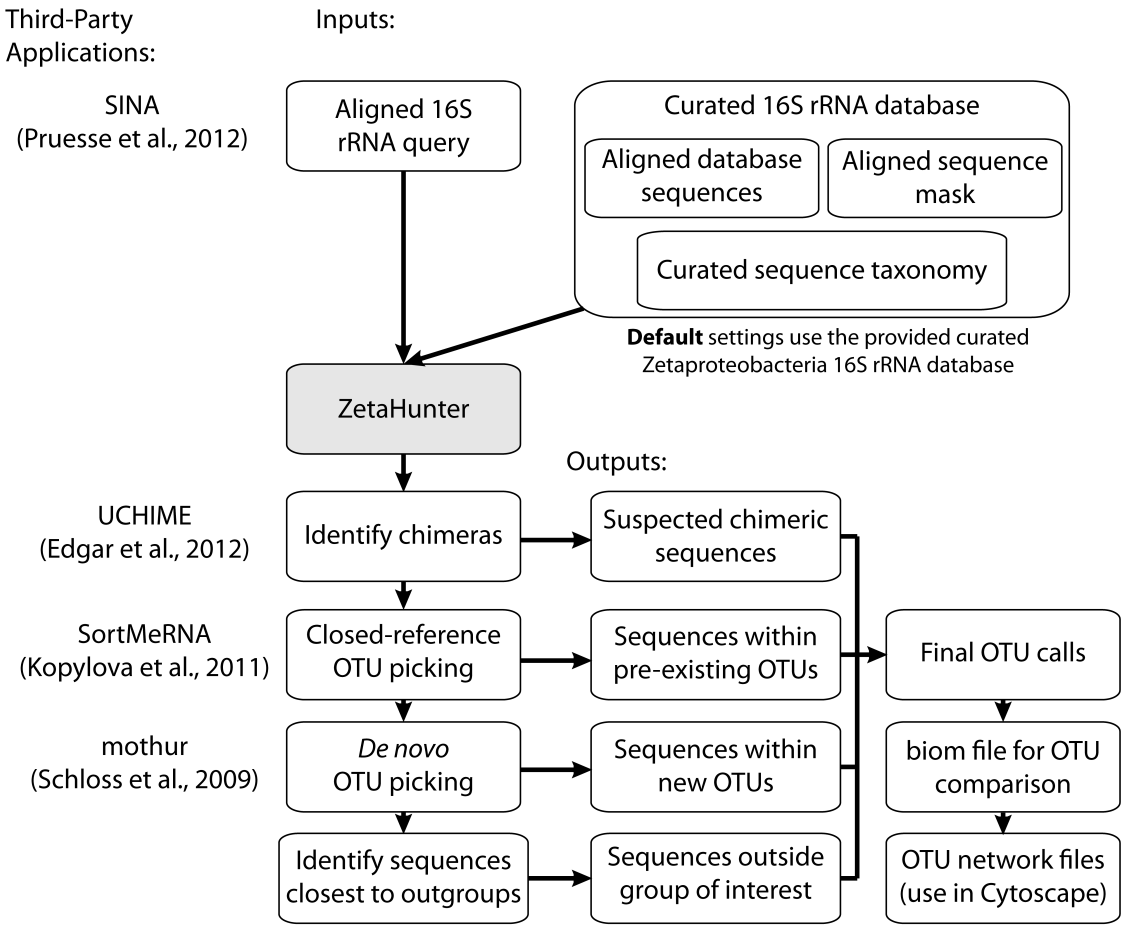


Figure 1: Flow chart showing the ZetaHunter pipeline. Third-party tools used in the pipeline are indicated to the left.

to a given taxonomic structure. That structure is based on the initial 28 ZOTUs, numbered by abundance by McAllister et al. [11], as well as novel full-length sequences from environmental clones and isolates. Each update of the database assigns numbers to new ZOTUs based on their relative abundance within that update, to preserve ZOTU numbering across uses of ZetaHunter. In 2018, we have more than doubled the initial estimates of biodiversity, with 59 ZOTUs determined using all near-full-length, non-chimeric, Zetaproteobacteria 16S rRNA gene sequences derived from arb SILVA (current database from release 128) and Zetaproteobacteria genomes from JGI's Integrated Microbial Genomes (IMG). See Supplemental Table 1 for the sequences found in the current ZetaHunter database (v.3). Notes or warnings for sequences within the ZetaHunter database will be maintained in ZetaHunter/assets/zh_db_v#_warning.txt.

IV. ZetaHunter use examples

ZetaHunter options:

```
-i, --inaln=<s+>    Input alignment(s)
-o, --outdir=<s>  Directory for output
-t, --threads=<i> Number of processors
                  to use (default: 2)

-d, --db-otu-info=<s>     Database OTU
                        info file name
    (default: /home/ZetaHunter/assets/
    db_otu_info.txt)

-m, --mask=<s>         Fasta file with
                              the mask
    (default: /home/ZetaHunter/assets/
    mask.fa.gz)

-b, --db-seqs=<s>      Fasta file with
                       aligned DB seqs
    (default: /home/ZetaHunter/assets/
    db_seqs.fa.gz)

-r, --mothur=<s> The mothur executable
    (default: /home/ZetaHunter/bin/
    linux/mothur)

-s, --sortmerna=<s>      The SortMeRNA
                            executable
    (default: /home/ZetaHunter/bin/
    linux/sortmerna)

-n, --indexdb-rna=<s>    The SortMeRNA
                    idnexdb_rna execu.
    (default: /home/ZetaHunter/bin/
    linux/indexdb_rna)

-c, --cluster-method=<s>        Either
         furthest, average, or nearest
    (default: average)

-u, --otu-percent=<i>   OTU similarity
                            percentage
    (default: 97)

-k, --check-chimeras,    Flag to check
    --no-check-chimeras       chimeras
    (default: true)

-a, --base=<s>    Base name for output
                                files
    (default: ZH_2017_11_28_15_36)

-e, --debug   Debug mode, don't delete
                 tmp files or clean up
              the working dir (the out
                    dir will be empty)

-v, --version   Print version and exit
-h, --help           Show this message
```

Generic run using Docker with four threads:

```
run_zeta_hunter -i Zetas_aligned.fasta
    -o ZH_out -t 4
```

Generic run using source installation:

```
ruby ~/software/ZetaHunter/zeta_hunter.
    rb -i Zetas_aligned.fasta -o ZH_out
    -t 4
```

Generic run without checking for chimeras:

```
run_zeta_hunter -i Zetas_aligned.fasta
    -o ZH_out -t 4 --no-check-chimeras
```

Generic run assigning alternative OTU clustering similarity:

```
run_zeta_hunter -i Zetas_aligned.fasta
    -o ZH_out -t 4 -u 98
```

Run using multiple infiles. Each infile is assigned a unique number, which is tracked by ZetaHunter. This is particularly useful with the biom format (compares ZOTU abundance across different samples) and Cytoscape network files (each edge can be associated with a particular sample). Example:

```
run_zeta_hunter -i ./in/*aligned.fasta
    -o ZH_MultiSample_out -t 4
```

ZetaHunter can also be run on non-Zetaproteobacteria sequences. The curated database developed for the poorly characterized clade is fed into the command line to replace the Zetaproteobacterial database defaults. Example:

```
run_zeta_hunter -i IN -o OUT -d
    db_otu_info -m mask.fasta.gz -b
    db_seqs_aligned.fasta.gz
```

V. Example using ZetaHunter to classify other poorly characterized microbes

The ZetaHunter program can be used with any curated database. This curation entails assigning reference sequences (preferably near-full length) to the desired taxonomy. Here we show an application with the candidate phylum OP3, recently named Omnitrophica [18]. For this phylum, Glöckner et al. [19] have defined a taxonomy based on five stable phylogenetic

divisions, which exist at a higher taxonomic level than the OTU. The average genetic distance within each division was 84.3% similarity, with a minimum similarity of 80%. Using the sequences from Glöckner et al. [19], we created an OP3 16S taxonomic database consisting of 104 near-full length sequences masked to 1,237 bp. After developing this curated database, we validated the ZetaHunter results using data from an intertidal mixing zone of a beach aquifer where the OP3 lineage was dominant (average read length 475 bp; [20]).

In McAllister et al. [20], 1,397 sequences were manually assigned to Omnitrophica divisions based on their position within a maximum likelihood phylogenetic tree. Using ZetaHunter at an 80% similarity threshold, we were able to assign these sequences to division level classifications in 4.6 min, with 88.0% accuracy to the original division assignments. For comparison, McAllister et al. [20] reported Omnitrophica divisions II and III dominated the beach aquifer at 53% and 31%, respectively, compared to ZetaHunter results at 56% and 28% (Figure 2). Though OTU comparisons at 97% are advised, this example highlights the flexibility of ZetaHunter to operate at any taxonomic level.

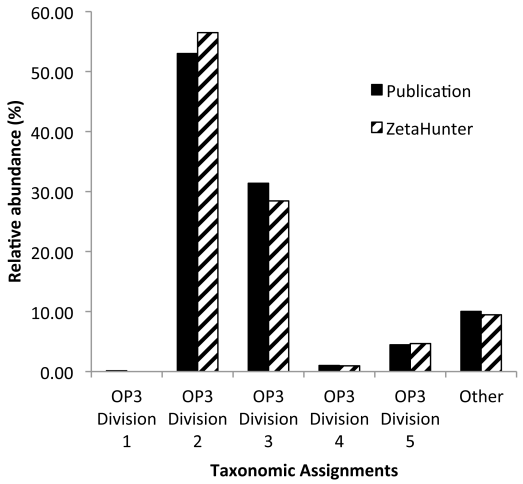


Figure 2: Comparison of taxonomic assignments between the manual curation of McAllister et al. [20] based on phylogeny and automated assignment by ZetaHunter.

VI. ZetaHunter output files

Example of the files and file structure found in the ZetaHunter output directory:

```
ZH_output/
├── biom
|   └── ZH_yyyy_mm_dd_hh_mm.biom.txt
├── cytoscape
|   ├── ZH_yyyy_mm_dd_hh_mm.cytoscape_network_edges.txt
|   └── ZH_yyyy_mm_dd_hh_mm.cytoscape_node_table.txt
├── dangerous_seqs
|   ├── chimera_details
|   |   ├── infile_aligned.ref.uchime.accnos
|   |   └── infile_aligned.ref.uchime.chimeras
|   └── ZH_yyyy_mm_dd_hh_mm.dangerous_seqs.chimeras.txt
├── log
|   ├── ZH_yyyy_mm_dd_hh_mm.log.mothur.txt
|   └── ZH_yyyy_mm_dd_hh_mm.log.zh.txt
├── misc
|   ├── ZH_yyyy_mm_dd_hh_mm.all_sortmerna_db_hits.txt
|   ├── ZH_yyyy_mm_dd_hh_mm.closest_db_seqs.txt
|   └── ZH_yyyy_mm_dd_hh_mm.sample_id_to_fname.txt
└── otu_calls
    ├── ZH_yyyy_mm_dd_hh_mm.otu_calls.closed_ref.txt
    ├── ZH_yyyy_mm_dd_hh_mm.otu_calls.denovo.txt
    └── ZH_yyyy_mm_dd_hh_mm.otu_calls.final.txt

7 directories, 14 files
```

Final ZOTU calls for each sequence can be found in the file: ZH_output/otu_calls/otu_calls.final.txt.
In this file, every sequence is listed with its sample number (S#), ZOTU call, percent of maximum entropy, percent of masked bases, and any flags. Sample numbers can be associated with the original file name using the sample_id_to_fname.txt file. Closed reference ZOTUs are indicated as "ZetaOtu#". *De novo* classified OTUs are indicated as "NewZetaOtu#", and are ordered by abundance. The ZetaHunter database includes 23 out group sequences, with representatives from each class of the Proteobacteria, and one sequence each from Thermotogae and Aquificae. If a sequence is within 97% identity to an out group sequence in the ZetaHunter database, the OTU call is to that out group, and the flag "OG_GTE_97" is applied (**O**ut **G**roup **G**reater **T**han or **E**qual to **97** percent identity). If a

Table 1: Example otu_calls.final.txt file (original tab-delimited).

#SeqID	Sample	OTU	PercEntropy	PercMaskedBases	Flag
Sequence1	S2	ZetaOtu4	99.4	99.9	0
Sequence4	S11	ZetaOtu1	100.0	100.0	0
Sequence3	S9	ZH_GOLD_v3_Aquifexpyr	98.2	99.0	OG_GTE_97
Sequence5	S10	NewZetaOtu1	99.3	99.6	OG_LT_97
Sequence6	S4	NewZetaOtu2	99.9	99.8	0
Sequence7	S1	NewZetaOtu2	99.9	99.8	0
Sequence8	S6	NewZetaOtu3	59.0	66.1	FRAGMENT
Sequence9	S6	NewZetaOtu3	59.0	66.1	FRAGMENT
Sequence10	S12	NewZetaOtu4	99.8	99.9	SINGLETON
Sequence11	S6	NewZetaOtu5	58.2	56.4	FRAGMENT|SINGLETON
Sequence12	S5	NewZetaOtu6	99.7	99.8	CHIMERA|OG_LT_97|SINGLETON

Table 2: Example closest_db_seqs.txt file (original tab-delimited).

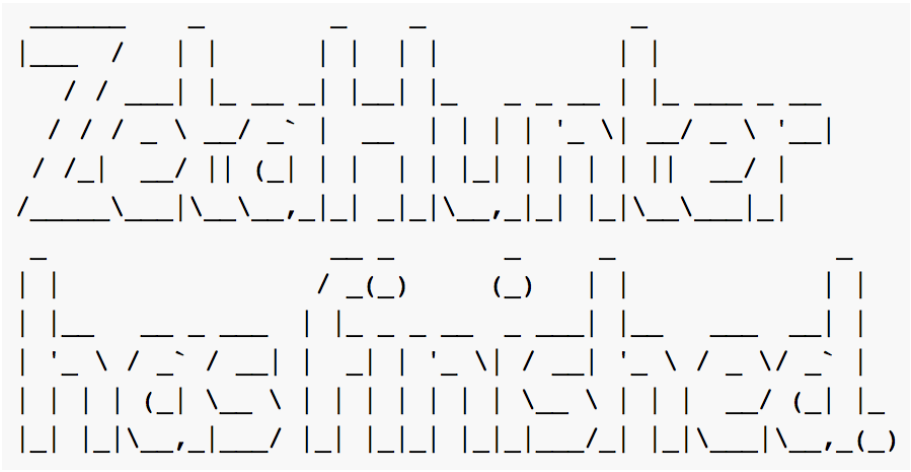
#SeqID	Sample	OTU	PercEntropy	PercMaskedBases	Hit	PID	QCov
Sequence1	S2	ZetaOtu4	99.4	99.9	ZH_GOLD_v3_KY417846_1	98.4	100.0
Sequence4	S11	ZetaOtu1	100.0	100.0	ZH_GOLD_v3_2525846989	99.1	100.0
Sequence3	S9	OG	98.2	99.0	ZH_GOLD_v3_Aquifexpyr	99.2	99.9
Sequence5	S10	OG	99.3	99.6	ZH_GOLD_v3_Gallionefe	95.9	100.0
Sequence6	S4	ZetaOtu51	99.9	99.8	ZH_GOLD_v3_LC086662_1_1504	96.6	100.0
Sequence7	S1	ZetaOtu51	99.9	99.8	ZH_GOLD_v3_LC086662_1_1504	95.8	100.0
Sequence8	S6	ZetaOtu22	59.0	66.1	ZH_GOLD_v3_EU491311_1_150	96.4	100.0
Sequence9	S6	ZetaOtu22	59.0	66.1	ZH_GOLD_v3_EU491311_1_1500	96.5	100.0
Sequence10	S12	ZetaOtu36	99.8	99.9	ZH_GOLD_v3_2572241983	96.4	100.0
Sequence11	S6	ZetaOtu54	58.2	56.4	ZH_GOLD_v3_2695406179	96.3	100.0
Sequence12	S5	OG	99.7	99.8	ZH_GOLD_v3_Gallionefe	91.5	96.2

sequence is closest to an outgroup, but not within a 97% OTU, then it is still classified as a NewZetaOtu, but is given the flag "OG_LT_97" (LT = **L**ess **T**han). These sequences are likely non-Zetaproteobacteria, yet could be novel Zetaproteobacteria sequences at the base of the phylogenetic tree. Sequences covering less than 75% of the Zetaproteobacteria 16S rRNA entropy (i.e. 75% of the base positions that contain information that separate ZOTUs) are given the "FRAGMENT" flag. Singletons and chimeras (as defined by UCHIME) are also flagged. Flags can be cumulative. An example of otu_calls.final.txt is shown in Table 1.

Another important out file is ZH_output/misc/closest_db_seqs.txt. From this file, the closest database hits for novel NewZetaOtus can be seen. The closest hit determines the OTU, yet the percent identity (PID) determines whether that OTU designation receives the final call (i.e. PID greater than OTU cutoff) or the sequence is passed to *de novo* clustering. An example of this file is shown in Table 2.

The standardized biom format is useful for comparing ZOTU composition across multiple samples. ZH_output/biom/biom.txt shows ZOTU abundance within each sample, and can be used to easily create bar charts to view ZOTU abundance. This file is converted automatically by ZetaHunter into node and edge files for input into an OTU network in Cytoscape. These files can show the connectedness of ZOTUs within and between a user's samples. Further control on which NewZetaOtus are shown within this OTU network can be obtained through filtering. The supplied script (ZetaHunter/bin/biom_to_cytoscape.rb) filters out NewZetaOtus below a minimum number of sequences, defined by the user. Usage (default min_otu_size = 1):

```
biom_to_cytoscape.rb biom.txt
    min_otu_size
```



Acknowledgements

We thank Dave Emerson for providing 16S rRNA sequences for the ZetaHunter database from novel Zetaproteobacteria isolates. D Emerson and B Chiu are thanked for their comments on early versions of this manuscript. We also thank K Hager, H Fullerton, J Scott, and J Vander Roost for testing and troubleshooting ZetaHunter using muliple platforms and datasets. We also acknowledge the ARB/SILVA group for their work with stable 16S rRNA alignments using the SINA tool; this tool is key to ZetaHunter's function. This work was funded by NSF grant OCE-1155290 (to CSC). The authors declare no conflict of interest.

References

[1] Hug, LA, BJ Baker, K Anantharaman, CT Brown, AJ Probst, CJ Castelle, CN Butterfield, AW Hernsdorf, Y Amano, K Ise, Y Suzuki, N Dudek, DA Relman, KM Finstad, R Amundson, BC Thomas, and JF Banfield. 2016. A new view of the tree of life. *Nat. Microbiol.*, 1:16048. doi:10.1038/nmicrobiol.2016.48

[2] Emerson, D, and CL Moyer. 2002. Neutrophilic Fe-oxidizing Bacteria are abundant at the Loihi Seamount hydrothermal vents and play a major role in Fe oxide deposition. *Appl. Environ. Microbiol.*, 68:3085–3093. doi:10.1128/AEM.68.6.3085-3093.2002

[3] Mori, JF, JJ Scott, KW Hager, CL Moyer, K Küsel, and D Emerson. 2017. Physiological and ecological implications of an iron- or hydrogen-oxidizing member of the Zetaproteobacteria, *Ghiorsea bivora*, gen. nov., sp. nov. *ISME J.*, 11:2624–2636. doi:10.1038/ismej.2017.132

[4] Beam, JP, JJ Scott, SM McAllister, CS Chan, J McManus, FJR Meysman, and D Emerson. 2017. A biological source of marine sedimentary iron oxides. *bioRxiv*, 108621. doi:10.1101/108621

[5] McAllister, SM, RM Moore, A Gartman, GW Luther III, and CS Chan. In prep. The marine Fe(II)-oxidizing Zetaproteobacteria.

[6] Chiu, BK, S Kato, SM McAllister, EK Field, and CS Chan. 2017. Novel pelagic iron-oxidizing Zetaproteobacteria from the Chesapeake Bay oxic-anoxic transition zone. *Front. Microbiol.*, 8:1280. doi:10.3389/fmicb.2017.01280

[7] Hager, KW, H Fullerton, DA Butterfield, and CL Moyer. 2017. Community structure of lithotrophically-driven hydrothermal microbial mats from the Mariana Arc and Back-Arc. *Front. Microbiol.*, 8:1578. doi:10.3389/fmicb.2017.01578

[8] Vander Roost, J, IH Thorseth, and H Dahle. 2017. Microbial analysis of Zetaproteobacteria and co-colonizers of iron mats in the Troll Wall Vent Field, Arctic Mid-Ocean Ridge. *PLoS ONE*, 12:e0185008. doi:10.1371/journal.pone.0185008

[9] Scott, JJ, BT Glazer, and D Emerson. 2017. Bringing microbial diversity into focus: high-resolution analysis of iron mats from the Lō'ihi Seamount. *Environ. Microbiol.*, 19:301–316. doi:10.1111/1462-2920.13607

[10] Emerson, D, JJ Scott, A Leavitt, E Fleming, and C Moyer. 2017. *In situ* estimates of iron-oxidation and accretion rates for iron-oxidizing bacterial mats at Lō'ihi Seamount. *Deep Sea Res. Pt. I*, 126:31–39. doi:10.1016/j.dsr.2017.05.011

[11] McAllister, SM, RE Davis, JM McBeth, BM Tebo, D Emerson, and CL Moyer. 2011. Biodiversity and emerging biogeography of the neutrophilic iron-oxidizing Zetaproteobacteria. *Appl. Environ. Microbiol.*, 77:5445–5457. doi:10.1128/AEM.00533-11

[12] Pruesse, E, J Peplies, and FO Glöckner. 2012. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, 28:1823–1829. doi:10.1093/bioinformatics/bts252

[13] Schloss, PD, D Gevers, and SL Westcott. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One*, 6:e27310. doi:10.1371/journal.pone.0027310

[14] Schloss PD, SL Westcott, T Ryabin, JR Hall, M Hartmann, EB Hollister, RA Lesniewski, BB Oakley, DH Parks, CJ Robinson, JW Sahl, B Stres, GG Thallinger, DJ van Horn, and CF Weber. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, 75:7537–7541. doi:10.1128/AEM.01541-09

[15] Edgar, RC, BJ Haas, JC Clemente, C Quince, and R Knight. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27:2194–2200. doi:10.1093/bioinformatics/btr381

[16] Kopylova, E, L Noé, and H Touzet. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, 28:3211–3217. doi:10.1093/bioinformatics/bts611

[17] Navas-Molina, JA, JM Peralta-Sánchez, A González, PJ McMurdie, Y Vázquez-Baeza, Z Xu, LK Ursell, C Lauber, H Zhou, SJ Song, J Huntley, GL Ackermann, D Berg-Lyons, S Holmes, JG Caporaso, and R Knight. 2013. Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol.*, 531:371–444. doi:10.1016/B978-0-12-407863-5.00019-8

[18] Rinke, C, P Schwientek, A Sczyrba, NN Ivanova, IJ Anderson, J-F Cheng, A Darling, S Malfatti, BK Swan, EA Gies, JA Dodsworth, BP Hedlund, G Tsiamis, SM Sievert, W-T Liu, JA Eisen, SJ Hallam, NC Kyrpides, R Stepanauskas, EM Rubin, P Hugenholtz, and T Woyke. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499:431–437. doi:10.1038/nature12352

[19] Glöckner, J, M Kube, PM Shrestha, M Weber, FO Glöckner, R Reinhardt, and W Liesack. 2010. Phylogenetic diversity and metagenomics of candidate division OP3. *Environ. Microbiol.*, 12:1218–1229. doi:10.1111/j.1462-2920.2010.02164.x

[20] McAllister, SM, JM Barnett, JW Heiss, AJ Findlay, DJ MacDonald, CL Dow, GW Luter III, HA Michael, and CS Chan. 2015. Dynamic hydrologic and biogeochemical processes drive microbially enhanced iron and sulfur cycling within the intertidal mixing zone of a beach aquifer. *Limnol. Oceanogr.*, 60:329–345. doi:10.1002/lno.10029