**The Global Coral Microbiome Project dataset.** Coral microbiome DNA sequences were selected from samples collected by the Global Coral Microbiome Project (GCMP) as described in Pollock et al (1), but including additional locations outside of Australia in Panama, Saudi Arabia, Columbia, Singapore, and Réunion that were not described in that manuscript. Briefly, samples were collected from water, sediment, and corals from 457 coral colonies, then DNA was extracted using the MoBio Powersoil DNA Isolation Kit and processed by the Earth Microbiome Project at the CMI center. PCR was run on the V4 region of the 16S rRNA gene using 515f/806r primers (5'-GTGTGCCAGCMGCCGCGGTAA-3') and sequenced using Illumina HiSeq with 125bp paired-end reads. Sequences were downloaded from the Earth Microbiome Project via QIITA project ID 10895 (specifically prep id 3439). In QIITA, these sequences were processed using standard EMP workflows: fastq files were demultiplexed using 12bp Golay codes with the QIIME 1.9.1 split_libraries script (default parameters), trimmed to 100nt, and then subjected to quality control with deblur 1.1.0 (default parameters).

**Analytical steps and code availability.** We downloaded the Earth Microbiome Project GCMP 'all.biom' (CRC32 id: 8817b8b8) and 'all.seqs.fa' (CRC32 id: ac925c85) from QIITA and reimported the sequences into QIIME2 2020.11 (2) and filtered to only samples collected from coral mucus, tissue, or skeleton. All analytical steps for the results described in this manuscript, including relevant parameters and explanations, are available at https://zenodo.org/record/4551201. Additionally, a live copy of the code is maintained on GitHub as part of the GCMP Global Disease Project at https://github.com/zaneveld/GCMP_Global_Disease/tree/master/analysis/organelle_removal.

**Construction of reference taxonomies and taxonomic assignment.** Reference taxonomies were constructed using Greengenes 13_8 (3), and SILVA 132 (4) as base references. Sequences and annotations from the Metaxa2 BLAST database (5) were extracted with a

BLAST utility ("blastdbcmd -entry all -db blast -out metaxa2.fasta"). Mitochondrial sequences were selected from the Metaxa2 database and manually annotated as Bacteria / Proteobacteria / Alphaproteobacteria / Rickettsiales / Mitochondria in the style of each base reference (SILVA 132 or greengenes_13_8). These supplementary mitochondrial sequences were then added to each base reference to create the expanded reference taxonomies. GCMP sequences from coral mucus, tissue, or skeleton samples were annotated with each reference taxonomy using the q2 feature-classifier classify-consensus-vsearch method (6).

**Investigation of differentially classified sequences.** Sequences labeled differently by an expanded taxonomy compared to the associated base taxonomy were separately investigated with Blast2GO(13) using the public QBlast NR database (blastn-short; e-value $10^{-3}$; word size 7; hsp length cutoff 33, low complexity filter was on). Blast2GO outputs contained scientific names which were expanded into full lineages by querying the NCBI Taxonomy database (14) via Biopython's Entrez module (15).

**Alpha and beta diversity metrics**. Non-phylogenetic alpha (observed_features, dominance, simpson_e) and beta diversity metrics (bray_curtis) were generated for each version of the results (SILVA, Greengenes, SILVA + Metaxa2, and Greengenes + Metaxa2) in QIIME2 using the QIIME2 Python API. Alpha and beta diversity were then compared across coral families using the column 'taxonomy_string_to_family' in the GCMP metadata. In this analysis, we accounted for the effects of improved identification of mitochondria on rarefaction analysis. Because expanding the taxonomies results in additional annotations of mitochondria, which are then removed, the expanded taxonomies may cause more samples to fall below a given rarefaction threshold (e.g. here a minimum of 1000 reads/sample). To allow for fair comparison of taxonomies despite this issue, we compare only samples that survive rarefaction under each paired taxonomic scheme (e.g. Greengenes and Greegenes + Metaxa2 or SILVA and SILVA +

Metaxa2). This is a conservative choice, as not applying this correction would result in a larger (possibly artifactual) observed difference between the base and expanded taxonomies due to sample size differences. This analysis was repeated for GCMP samples derived from coral mucus, tissue, and skeleton separately.

**Application to chronic Montipora White Syndrome microbiomes.** As a further demonstration of the effects of coral mitochondria removal, we used the expanded SILVA taxonomy as the reference for annotations of a dataset of whole crushed coral microbiomes from a study of chronic *Montipora* White Syndrome (cMWS; Brown *et al.*, in preparation).

**Evaluation of accuracy of extended taxonomic references.** To test the effect of these additional reference sequences on classification of non-mitochondrial sequences, we annotated several mock communities of known composition with each of the four reference taxonomies using the QIIME2 quality-control evaluate-composition method (6). To test for non-inferiority of the expanded taxonomies, assignments created with the base taxonomies were used as the "expected" taxa and assignments created with the expanded taxonomies were used as the "observed" taxa. We compared assignments of mock communities 12-16 (7–9) and 18-22 (9–11) from Mockrobiota (12) which were chosen for compatibility with QIIME2.

**Comparison of annotation of Aquarickettsiales sequences.** In order to test whether expanded taxonomies might misannotate key coral symbionts that are relatively close taxonomic relatives of mitochondria, 38 Aquarickettsiales 16S sequences from previous coral microbiome studies (16,17) were imported into QIIME2 and annotated with the q2 feature-classifier classify-consensus-vsearch method (6) using each of the reference taxonomies.

**References**


1. Pollock FJ, McMinds R, Smith S, Bourne DG, Willis BL, Medina M, et al. Coral-associated bacteria demonstrate phylosymbiosis and cophylogeny. Nat Commun. 2018 Nov 22;9(1):1–13.
2. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, et al. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. PeerJ Prepr. 2018 Oct;6:e27295v1.
3. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol. 2006 Jul 1;72(7):5069–72.
4. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013 Jan 1;41(D1):D590–6.
5. Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, et al. metaxa2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. Mol Ecol Resour. 2015;15(6):1403–14.
6. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018 May 17;6(1):90.
7. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016 Jul;13(7):581–3.
8. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. Appl Environ Microbiol. 2013 Sep 1;79(17):5112–20.

9. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 2015 Mar 31;43(6):e37–e37.
10. Tourlousse DM, Yoshiike S, Ohashi A, Matsukura S, Noda N, Sekiguchi Y. Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. Nucleic Acids Res. 2017 Feb 28;45(4):e23–e23.
11. Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, et al. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat Biotechnol. 2016 Sep;34(9):942–9.
12. Bokulich NA, Rideout JR, Mercurio WG, Shiffer A, Wolfe B, Maurice CF, et al. mockrobiota: a Public Resource for Microbiome Bioinformatics Benchmarking. mSystems [Internet]. 2016 Oct 25 [cited 2020 Jun 26];1(5). Available from: https://msystems.asm.org/content/1/5/e00062-16
13. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008 Jun 1;36(10):3420–35.
14. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database J Biol Databases Curation. 2020 01;2020.
15. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 Jun 1;25(11):1422–3.
16. Klinges JG, Rosales SM, McMinds R, Shaver EC, Shantz AA, Peters EC, et al. Phylogenetic, genomic, and biogeographic characterization of a novel and ubiquitous marine invertebrate-associated Rickettsiales parasite, Candidatus *Aquarickettsia rohweri*, gen. nov., sp. nov. ISME J. 2019 Dec;13(12):2938–53.

17. Klinges G, Maher RL, Thurber RLV, Muller EM. Parasitic ‘Candidatus *Aquarickettsia rohweri*’ is a marker of disease susceptibility in *Acropora cervicornis* but is lost during thermal stress. Environ Microbiol. 2020;22(12):5341–55.