The Organelle in the Room: Under-annotated Mitochondrial Reads Bias Coral Microbiome Analysis

Dylan Sonett[1], Tanya Brown[1], Johan Bengtsson-Palme[2,3], Jacqueline L. Padilla-Gamiño[4], Jesse R. Zaneveld[1]

[1] Division of Biological Sciences, School of STEM, University of Washington Bothell, Bothell, WA, USA

[2] Department of Infectious Diseases, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Sweden

[3] Centre for Antibiotic Resistance Research (CARe) at University of Gothenburg, Sweden

[4] School of Aquatic and Fisheries Sciences, University of Washington, Seattle, WA, USA

Corresponding Author:

Dr. Jesse Zaneveld [1]

Division of Biological Sciences, Bothell, School of Science, Technology, Engineering, and Mathematics, University of Washington, UWBB 249, Bothell, WA, 98011, USA

Email address: zaneveld@uw.edu

Competing Interests: The authors declare no competing interests.

Abstract

The microbiomes of tropical corals are actively studied using 16S rRNA gene amplicons to understand microbial roles in coral health, metabolism, and disease resistance. However, primers targeting bacterial and archaeal 16S rRNA genes may also amplify organelle rRNA genes from the coral, associated microbial eukaryotes, and encrusting organisms. In this manuscript, we demonstrate that standard workflows for annotating microbial taxonomy severely under-annotate mitochondrial sequences in 1272 coral microbiomes from the Earth Microbiome Project. This issue prevents annotation of >95% of reads in some samples and persists when using either Greengenes or SILVA taxonomies. Worse, mitochondrial under-annotation varies between species and across anatomy, biasing comparisons of α- and β-diversity. By supplementing existing taxonomies with diverse mitochondrial rRNA sequences, we resolve ~97% of unique unclassified sequences as mitochondrial, without increasing misannotation in mock communities. We recommend using these extended taxonomies for coral microbiome analysis and encourage vigilance regarding similar issues in other hosts.

Introduction

Corals are animals that exist in intimate symbiosis with a wide variety of microscopic symbionts including microbial eukaryotes, bacteria, archaea, viruses, and phages (1–3). Collectively, these coral associates and the cnidarian host animal are known as the coral holobiont. The interactions of coral-associated microbes with their host and one another are complex, but of great interest due to their potential to modulate coral disease susceptibility (e.g. (1)) and responses to environmental stressors. Marker gene studies using small-subunit ribosomal RNA gene amplicons (SSU rRNA) have played a key role in describing the coral holobiont. Yet there are older symbioses that complicate these studies.

Traces of the evolutionary history of organelles as formerly free-living bacteria can be found in their genomes. For example, animal mitochondria carry their own small subunit rRNA gene, known as the 12S rRNA gene. These 12S rRNA genes are often amplified by the same PCR primers used for 16S rRNA analysis of bacteria and archaea. This can pose problems for microbiome studies (2,3) if organelle SSU rRNA gene sequences are not removed *in silico*, or excluded using special laboratory procedures like peptide nucleic acid clamps (2) or CRISPR-Cas9 cleavage (3). Because laboratory methods are relatively laborious and taxon-specific, a common approach is to identify and filter out organelle rRNA sequences *in silico* using standard taxonomy annotation pipelines such as the naive-Bayesian RDP classifier (4), alignment-based algorithms such as USEARCH (5) and VSEARCH (6), or machine learning approaches (7). If this process is accurate and unbiased across categories of samples, then removal of mitochondrial SSU rRNA sequences reduces effective sequencing depth but does not otherwise compromise microbiome analysis.

Application of *in silico* methods to coral microbiome libraries typically does identify some mitochondrial SSU rRNA gene sequences. However, variation in coral mitochondrial 12S rRNA genes has long been known to exist. Indeed the deeply divergent ‘robust’ and ‘complex’ clades of the phylogenetic tree of scleractinian corals were initially named and characterized based on their ‘short’ or ‘long’ 12S rRNA PCR products (8,9). The existing literature does not establish whether known taxon-specific variation in coral mitochondrial 12S rRNA gene sequences might impede their *in silico* removal from coral microbiome SSU rRNA marker gene libraries in a host-specific manner.

In this manuscript we report widespread, severe, and host-specific under-annotation of mitochondrial sequences in short-read coral microbiome SSU rRNA amplicon libraries; demonstrate that differences in mitochondrial under-annotation between coral taxa bias comparisons of microbiome diversity across coral families; and propose a simple extension to existing microbial taxonomies that appears to mostly resolve this issue.

SILVA and Greengenes under-annotate mitochondrial ribosomal RNAs

The Global Coral Microbiome Project (GCMP) dataset includes a collection of 1 272 16S rRNA gene amplicon libraries from the mucus, tissue, and skeleton of phylogenetically diverse coral taxa sequenced on Illumina HiSeq using the Earth Microbiome Project protocol ((10); Supplemental Table 1a). During analysis of this dataset we noticed many sequences which were not annotated by standard workflows (e.g. ‘Unknown’ annotations). Indeed ‘Unknown’ sequences represented 38% of total reads according to vsearch annotation with SILVA and 41% using Greengenes. Such ‘Unknown’ sequences accounted for >95% of microbial relative abundance in 51 coral samples using SILVA or 59 coral samples using Greengenes. Many of these unannotated sequences appeared to be mitochondrial in origin based on *ad hoc* BLAST searches. Troublingly, BLAST even identified possible cryptic mitochondrial sequences in samples where other coral mitochondrial 12S rRNA genes were successfully annotated using vsearch. We reasoned that apparent under-annotation of mitochondrial SSU rRNA sequences in coral microbiomes could be explained by some combination of species-specific mitochondrial 12S rRNA gene length and sequence variation (e.g. (8,9)); coral heteroplasmy (11)); mitochondrial sequences from encrusting or ingested organisms (12) and incomplete representation of mitochondrial sequences from diverse hosts in SILVA (13) and Greengenes (14).

Expanding taxonomic references improves detection of mitochondrial RNA genes

To address this problem more formally, we developed a workflow for expanding the SILVA 132 (13) and Greengenes 13_8 (14) reference taxonomies with diverse mitochondrial sequences from the Metaxa2 (15) project (Supplementary Methods). We then used vsearch classification to re-annotate GCMP sequences with one of several standard taxonomies: SILVA 132, Greengenes 13_8, or the expanded versions of the same, which we refer to as silva_metaxa2 and Greengenes_metaxa2. We expected that if the mitochondrial references in existing taxonomies were already sufficiently diverse, then adding additional references would not alter taxonomic annotations. Conversely, if these existing taxonomies lacked representation of diverse mitochondrial SSU rRNA gene sequences, adding those sequences might ameliorate mitochondrial under-annotation.

Taxonomic annotations of the GCMP dataset using the expanded silva_metaxa2 or Greengenes_metaxa2 taxonomies had 97% fewer ‘unannotated’ sequences, and roughly proportional increases in annotated mitochondria (Fig. 1a). This resolved more than 99% of fully unannotated sequences as coral mitochondria. This suggests that many sequences of unknown taxonomy in coral microbiomes are divergent under-annotated mitochondrial 12S rRNA genes.

When coral mucus, tissue and skeleton were analyzed separately using the expanded taxonomies, samples from coral tissue - which we expect to be richest in coral mitochondria - had higher proportions of under-annotated coral mitochondria (Fig. 1b). Finally, BLAST searches of all sequences that were differentially annotated when using the expanded taxonomies identified the most commonly differentially annotated sequences (among those with BLAST hits) as cnidarian mitochondria (26%), primarily from coral families Pocilloporidae, Merulinidae, Poritidae, Acroporidae and Lobophylliidae (Supplementary Tables S2a and S2b).

Mitochondrial under-annotation biases diversity estimates

We next explored how under-annotation of coral mitochondria might influence statistical analysis of coral microbiomes. Coral mitochondrial under-annotation was strongly biased across coral families (Supplementary Fig. 1). Failure to annotate these mitochondria was sufficient to alter the outcome of cross-family comparisons of microbial richness and evenness in the GCMP dataset (Supplementary Table S1e). For example, using standard SILVA or Greengenes taxonomies, coral families appear to exhibit significant differences in microbiome richness and evenness in mucus, tissue, and skeleton (e.g. for the ‘observed species’ metric, Kruskal-Wallis p = 0.003 with standard SILVA). Yet improved annotation of mitochondria renders cross-family differences in mucus microbiome richness and evenness not significant (p = 0.078 with SILVA + Metaxa2; full results in Supplementary Table S1e). At the same time, improved annotation of mitochondria increased the significance of cross-family differences in the microbiome richness of tissue or skeleton by up to 5 orders of magnitude, and evenness by up to 10 orders of magnitude (Supplementary Table S1e). Although the significance of β-diversity comparisons of microbiome differences between coral families did not change when mitochondrial under-annotations were resolved, the effect size of these differences (i.e. Kruskal-Wallis H statistics) were cut to only 58% of their prior values when using the expanded taxonomies. This suggests that cryptic coral mitochondria can dramatically alter estimates of cross-family differences in coral microbiomes.

Longer read lengths reduce but do not eliminate under-annotation

To test whether mitochondrial under-annotation was peculiar to short-read Illumina HiSeq libraries used in the Earth Microbiome Project, we also reanalyzed microbiomes from corals affected by chronic *Montipora* White Syndrome (cMWS; Brown *et al.*, in revision). Brown *et al.,* used Ion Torrent sequencing, and had longer read lengths than the Earth Microbiome Project. In the Brown *et al.* dataset, we found fewer unclassified sequences, but still observed a 16-fold increase in mitochondrial annotations (~31 million vs. 1.9 million) with the silva_metaxa2 expanded reference set relative to the standard SILVA reference (Supplemental Data Table S3; Supplementary Methods). Thus, mitochondrial under-annotation does not appear to be unique to Earth Microbiome Project protocols. Together, these findings suggested that inclusion of diverse mitochondrial reference sequences greatly increased annotation of mitochondrial 12S rRNA sequences in coral microbiomes.

Expanding reference taxonomies does not lead to mitochondrial over-annotation

One potential concern with expanding reference taxonomies is that it might increase mis-annotation of certain bacteria as mitochondria. We tested whether increased mitochondrial annotations might lead to false positives by applying our expanded taxonomies to mock communities of known composition that did not contain mitochondria. In all tested mock communities from the mockrobiota project (16), expanding the mitochondrial reference set did not increase false positive annotations of mitochondria, and did not affect overall accuracy (maximum change in f-measure < 10^{-5}; Supplementary Figure S2).

Recent studies have identified bacteria in the genus *Aquarickettsia* as of particular interest as mediators of coral health (1,17). However, as *Aquarickettsia* and other members of Midichloreaceae are - in relative terms - somewhat closely related to mitochondria, we wanted to test whether our expanded taxonomies might increase mis-annotation of these important coral symbionts as mitochondria. Reassuringly, annotations for all coral- or placozoan-associated *Aquarickettsia* from recent studies (1,17) did not change using the updated taxonomies (Supplementary Table S2d).

Conclusion

These results demonstrate that extension of the SILVA or Greengenes taxonomies with diverse mitochondrial sequences can improve taxonomic annotations of coral mitochondria. While we explore mitochondrial misannotation in corals, it seems likely that similar issues may occur in any sufficiently diverse and deeply divergent set of eukaryotic hosts (e.g. marine sponges (18)). To address this issue, we recommend that investigators studying host-associated microbiomes ensure that diverse host mitochondrial reference sequences are included in their reference database by either using the pre-calculated QIIME2-compatible taxonomies supplied here, or, in the future, by updating the most recent SILVA or Greengenes taxonomies with diverse mitochondrial sequences.

Acknowledgements

J.R.Z. and T.B. are supported by NSF IOS CAREER grant 1942647. TB and JLPG are supported by NSF IOS grant 1655682. A static copy of code and relevant data files for this project can be found at https://zenodo.org/record/4551201. Additionally, a live copy 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.

The authors thank Nicholas Bokulich and Greg Caporaso for suggestions regarding software to compare mock communities, Grace Klinges for sharing *Aquarickettsia* sequences, Elizabeth Traylor for initial exploration of this issue, and Ayo Akinrinade for comments that improved the manuscript.

Competing Interests

The authors declare no conflicts of interest.

Figures

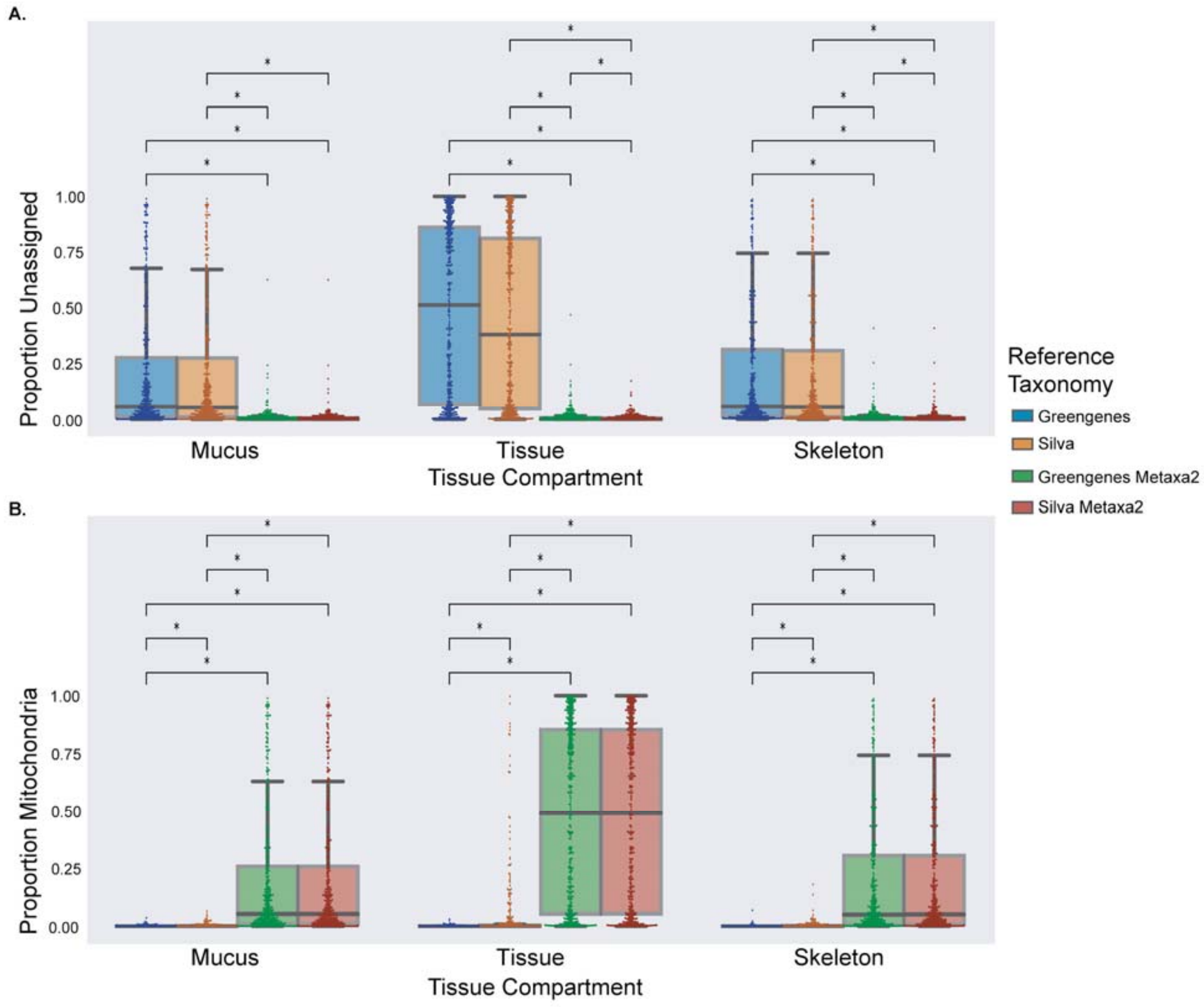
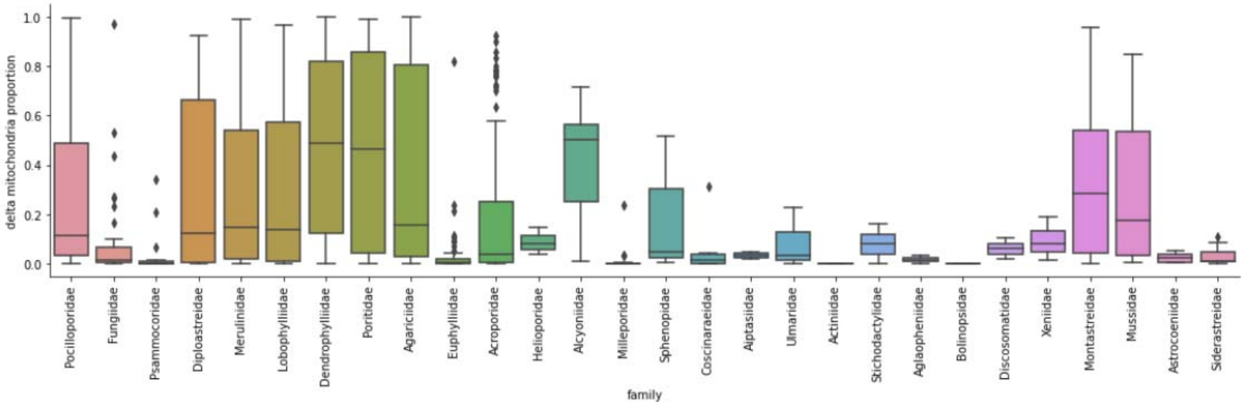


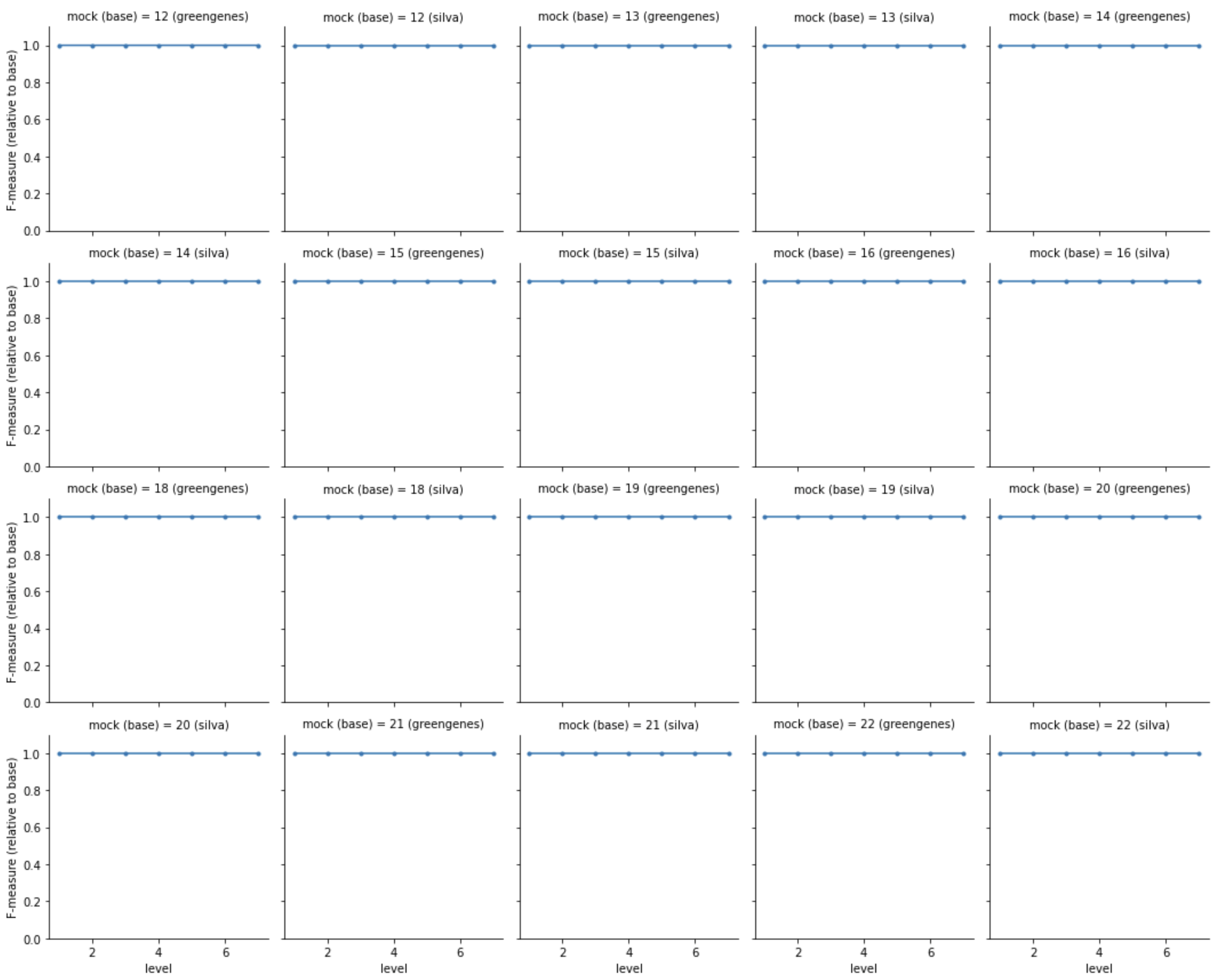
Fig. 1. Expanding the reference set dramatically increases coral mitochondrial annotations. Results generated by running the mitochondrial removal pipeline across 1 272 coral microbiome SSU rRNA gene amplicon libraries from the Global Coral Microbiome Project (Supplementary Table 1b), showing either **a.** the proportion of sequences that were not classified at the domain level (Supplementary Table 1c) or **b.** the proportion of sequences annotated as mitochondria (Supplementary Table 1d). We annotated each coral tissue compartment, and display results for coral mucus, tissue and skeleton samples in separate columns. In both panels, colors reflect the taxonomic annotation scheme used (Greengenes

Metaxa2 and Silva Metaxa2 contain supplemental mitochondrial sequences). Asterisks (*) indicate Bonferroni-corrected pairwise p-values below 0.05 in Kruskal-Wallis tests. In all cases, adding additional mitochondrial references significantly reduced the number of ‘Unknown’ sequences and significantly increased the number annotated as coral mitochondria relative to using the base taxonomy. This may indicate that coral-associated mitochondria are sufficiently unique in SSU rRNA sequence that standard microbial taxonomy workflows using either Greengenes_13_8 or SILVA miss a large proportion of them.



Supplementary Fig. S1. Coral mitochondrial sequences under-annotated by SILVA 132 or Greengenes_13_8 reference taxonomies are differentially represented across coral families. These differences are plotted above. The y-axis represents the difference in the relative abundance of mitochondria between SILVA 132 and our expanded silva_metaxa2 reference. Similar results were found for Greengenes annotations (raw data available in Supplementary Table S1b). As the only difference between these resources is the addition of supplementary reference sequences, we interpret this difference as a measure of the minimal amount of mitochondrial misannotation when using standard taxonomies. For both SILVA and Greengenes, inter-family differences in mitochondrial misannotation were significant overall when compared using Kruskal-Wallis tests (Greengenes: H = 211.5, p = $1.6*10^{-30}$; SILVA 132: H

= 228.6, $p = 8.1*10^{-34}$). Additionally, 49 coral family pairs significantly differed in their degree of mitochondrial misannotation using SILVA 132, while 43 differed using Greengenes_13_8 (Kruskall-Wallis test, Bonferroni-corrected $p < 0.05$; Supplementary Table S1b). Such differences may systematically bias comparisons of α- or β-diversity between coral families.



Supplementary Fig. S2. Effects of the expanded database on the accuracy of non-mitochondrial taxonomic annotations. We expected that the expanded taxonomies developed in this paper (Greengenes_metaxa2 or silva_metaxa2) should have identical results as the standard taxonomies when applied to communities lacking mitochondria. We tested this expectation using 10 mock communities from the mockrobiota resource (16). Each panel compares the accuracy of mock community annotations using the supplemented taxonomies (greegenes_metaxa2 or silva_metaxa2) vs. the base taxonomies (Greengenes or silva) at each taxonomic level (x-axis) for a single mock community. As there are expected to be no mitochondria in these mock communities, the standard taxonomies are treated as correct, and any deviations are treated as errors. The F-measure (the harmonic mean of precision and recall) is plotted on the y-axis of each plot. In all cases, annotations were identical or virtually identical between taxonomies (maximum F-measure difference <10^{-5}, Supplementary Table S2c).

Supplementary Data Tables

Supplementary Data Table 1. Results of testing annotation accuracy for the Global Coral Microbiome Project (GCMP) dataset. **a.** Metadata for the GCMP project samples used in this study (e.g. coral species, temperature, depth, etc). **b.** Annotation statistics by sample and taxonomy. Sample data and annotation results by reference taxonomy. This table includes counts for the number of ASVs that could not be annotated and the number annotated as mitochondria for each taxonomic resource. **c.** Comparison of proportions of Unknown ASVs assigned by taxonomy. This table holds the results of pairwise Kruskal-Wallis tests comparing compartment-specific proportions of sequences which were not identified at the domain level across different reference taxonomies. **d.** Comparison of proportions of mitochondrial ASVs assigned by reference taxonomy. This table holds the results of pairwise Kruskal-Wallis tests comparing compartment-specific proportions of sequences identified as mitochondria based on different reference taxonomies. **e.** Comparison of α- and β-diversity results by annotation scheme. Differences in α-diversity between coral families based on mitochondrial annotation method. Statistics reflect the results of either Kruskal-Wallis tests for several α-diversity metrics across coral families. **f.** Comparison of β-diversity results by annotation scheme. Statistics reflect the results of either Kruskal-Wallis tests for several β-diversity metrics across coral families.

Supplementary Data Table 2. Performance of reference taxonomies. **a.** NCBI BLAST lineages of differentially annotated sequences (SILVA). NCBI Taxonomy lineages of BLASTed sequences annotated differently by the silva_metaxa2 and SILVA reference taxonomies. **b.** NCBI BLAST lineages of differentially annotated sequences (Greengenes). NCBI Taxonomy lineages of BLASTed sequences annotated differently by the Greengenes_metaxa2 and Greengenes reference taxonomies. **c.** Mock community accuracy comparisons. Accuracy statistics of annotations of Mockrobiota mock communities generated by comparing extended reference taxonomies to their base taxonomy using the qiime2 evaluate-taxonomy method. Base reference annotations were considered perfectly accurate, for purposes of comparison. **d.** Annotations of Aquarickettsiales sequences by reference taxonomy. Full annotations for 38 Aquarickettsiales sequences from (1,17) using each reference taxonomy.

Supplementary Data Table 3. Results of testing annotation accuracy for the Brown *et al*. chronic Montipora White Syndrome (cMWS) dataset. This table reports the per-sample sequence counts of non-organelle sequences from Brown *et al.* when annotated with either silva or silva_metaxa2.

References

1. 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.
2. Fitzpatrick CR, Lu-Irving P, Copeland J, Guttman DS, Wang PW, Baltrus DA, et al. Chloroplast sequence variation and the efficacy of peptide nucleic acids for blocking host amplification in plant microbiome studies. Microbiome. 2018 Aug 18;6(1):144.

3. Song L, Xie K. Engineering CRISPR/Cas9 to mitigate abundant host contamination for 16S rRNA gene-based amplicon sequencing. Microbiome [Internet]. 2020 Jun 3 [cited 2020 Jul 15];8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7268715/

4. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ Microbiol. 2007 Aug;73(16):5261–7.

5. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010 Oct 1;26(19):2460–1.

6. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016 Oct 18;4:e2584.

7. 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.

8. Romano SL, Palumbi SR. Evolution of Scleractinian Corals Inferred from Molecular Systematics. Science. 1996 Feb 2;271(5249):640–2.

9. Chen CA, Wallace CC, Wolstenholme J. Analysis of the mitochondrial 12S rRNA gene supports a two-clade hypothesis of the evolutionary history of scleractinian corals. Mol Phylogenet Evol. 2002 May;23(2):137–49.

10. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017 Nov;551(7681):457–63.

11. Emblem Å, Karlsen BO, Evertsen J, Miller DJ, Moum T, Johansen SD. Mitogenome polymorphism in a single branch sample revealed by SOLiD deep sequencing of the *Lophelia pertusa* coral genome. Gene. 2012 Sep 15;506(2):344–9.

12. Leal MC, Ferrier□Pagès C, Calado R, Thompson ME, Frischer ME, Nejstgaard JC. Coral feeding on microalgae assessed with molecular trophic markers. Mol Ecol.

2014;23(15):3870–6.

13. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res. 2014 Jan 1;42(Database issue):D643–8.

14. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. 2012 Mar;6(3):610–8.

15. 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.

16. 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

17. 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.

18. Thomas T, Moitinho-Silva L, Lurgi M, Björk JR, Easson C, Astudillo-García C, et al. Diversity, structure and convergent evolution of the global sponge microbiome. Nat Commun. 2016 Jun 16;7:ncomms11870.