AN ABSTRACT OF THE DISSERTATION OF

Ryan McMinds for the degree of Doctor of Philosophy in Microbiology presented on December 14, 2018

Title: The Evolution of Scleractinian Microbial Symbionts

Abstract approved: _____
Rebecca L. Vega Thurber

Abstract

The modern world has presented many threats to the health and stability of ecosystems worldwide. One of the most biodiverse ecosystems, coral reefs, faces particularly strong pressures, and is already declining rapidly in complexity and area. Although the stressors that affect reefs are diverse, ranging from nutrient pollution to overfishing, invasive species to climate change, the impact of many of these stressors is ultimately mediated through interactions between the coral animal and its microbial associates, or microbiome. Some such interactions are readily apparent and have been studied for decades. For instance, coral bleaching, which is caused in part by increases in water temperature due to climate change, has devastated large swaths of reefs in recent years. The visual 'bleaching' that characterizes this phenomenon is the result of a breakdown in the symbiosis of the coral with photosynthetic algae of the family Symbiodiniaceae that normally live within its tissue. These algae provide the coral with essential energy, nutrients, and other services, but under temperature stress, they are expelled from the transparent tissue, leaving the white underlying coral skeleton visible and the animal without its food source.

However, other interactions between the coral and its microbiome are less well-defined. Many coral diseases, for instance, may be caused by the opportunistic overgrowth of fungi and bacteria in nutrient-rich water or under conditions of general stress. Even less clear is how bacteria may act as mutualists in the coral 'holobiont'. Other cnidarians have been shown to require developmental stimulation from particular bacterial species, and

non-Symbiodiniaceae microbes have also been hypothesized to act as nutritional symbionts or as defense against other, pathogenic microbes.

Scleractinian corals are diverse; having been evolving for more than 450 million years and including over 1,500 species. Because of this, they are likely to have many different modes of interaction with their microbiomes. To begin to better understand the similarities and differences among the microbiomes of corals, I conducted during my PhD the Global Coral Microbiome Project (GCMP), which sampled thousands of coral colonies from dozens of phylogenetically diverse species. Through the course of my work, I identified the similarities and differences between the microbiomes of these many species, showed the importance of considering shared evolutionary history in the analysis of such datasets, and developed new, rigorous methods of microbiome analysis that separate the effects of distinct evolutionary and ecological processes.

The Evolution of Scleractinian Microbial Symbionts

by

Ryan McMinds

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented December 14, 2018
Commencement June 2019

Doctor of Philosophy dissertation of Ryan McMinds presented on December 14, 2018.

APPROVED:

_____
Major Professor, representing Microbiology

_____
Head of the Department of Microbiology

_____
Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

_____
Ryan McMinds, Author

ACKNOWLEDGEMENTS

Over the course of my time at OSU, I have met an extraordinary number of amazing scientists and non-scientists around the world who have helped me with fieldwork, paperwork, and life. In particular, thank you to Justin Smith and David Baker for accompanying to various locations around the world, for the many meaningful discussions about our work, and for doing such a great job communicating our science to the public. To all of those that I've met at conferences and during my travels, thank you so much!

I owe an unpayable debt to my parents, Dan and Sandy McMinds, who instilled in me a love for nature, an insatiable curiosity, and the support I've needed for my entire life. I have been incredibly lucky to be able to pursue my dreams the way I have, and I owe so much of it to them. I am also extremely grateful for my brother, Justin McMinds, for pushing me (intentionally or not) to be more competitive, and for being a model of hard work, humility, generosity, and passion.

To my wife Brianna McMinds: I love you so much. Thank you for being so supportive over the last few years, as the stress from this dissertation grew and grew. I really could not have done any of this without you. I look forward to our future, and our future science/art collaborations. Thank you also to my in-laws, who have also been so welcoming and supportive over the last few years.

I am infinitely grateful to my scientific mentors in both the recent and less-recent past. Since I was little, my interest and skills in genetics and biology have been shaped by the guidance of diverse teachers, advisors and peers. My teachers before college fostered a love for learning that has only grown over time. I became confident with my skills and my career trajectory during my first science job with Dr. David Gent and Nancy Adair. While studying as an undergraduate at the University of Miami, my mentor Dr. Alex Wilson helped me become more rigorous and more precise, and Dr. Dan Price guided the development of my skills in the laboratory. I learned how to mix science with art and public communication during a summer of fieldwork with Dr. Nathan Dappen and will forever draw on those skills as I continue in my career. All of my professors, including

ACKNOWLEDGEMENTS

my anthropology, history, and other non-biologist professors, have helped shape the way I approach my work, and I thank you all.

To my PhD advisor Dr. Rebecca Vega Thurber: thank you so much for all that you have done for me. You have given me innumerable opportunities to diversify my scientific knowledge, collaborate with amazing scientists, and travel to incredible locations. You have fostered an incredible family of lab-mates and helped all of us grow as people as well as scientists. Thank you so much for allowing me the freedom to explore new concepts and techniques, for pushing me to spend time with meaningful outreach activities, and for continually building my scientific confidence. I will always look to you to find the big picture when I feel I am getting lost in the minutia.

Since joining the Vega Thurber lab, I have been supported by numerous amazing colleagues. I am very grateful for the friendship, scientific minds, and technical support of our various former and current postdocs, graduate students, and undergraduate techs. These include but are certainly not limited to: Dr. Jerome Payet, Dr. Jesse Zaneveld, Dr. Nitzan Soffer, Dr. Rory Welsh, Dr. Stephanie Rosales, Grace Klinges, Becca Maher, Adriana Messyasz, and Emily Schmeltzer. You really are like family, and I will truly miss you all.

I would also like to extend a huge thanks to the funding sources that allowed me to do this work. Oregon State University's Provost Fellowship helped get me started in the lab, NSF's East Asia and Pacific Summer Institutes program (EAPSI) got my global project rolling, NSF's Dimensions of Biodiversity program funded the rest of the project, and the Department of Microbiology's Middlekauf Graduate Student Fellowship and Joan Countryman Suit Fellowship picked up the slack near the end of my tenure. I recognize that all these funds, including tax-funded federal grants, ultimately come from people whose lives are not likely to be directly improved by the projects they are supporting, and I deeply appreciate their commitment to the zig-zagging, stop-and-go, imperfect and seemingly impossible progress of knowledge called science.

CONTRIBUTION OF AUTHORS

Ryan McMinds independently wrote the general abstract, general introduction, and general conclusion. He was a principle author for all other chapters. The contributions of co-authors to each chapter are listed in designated sections within each chapter.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

General Introduction

Corals and coral reefs

Coral reefs are one of the most biodiverse ecosystems on Earth. They are home to around one third of the Earth's fish diversity despite covering less than 0.1% of the Earth's surface (Bowen *et al.* 2013). Corals themselves, animals within the phylum Cnidaria, are largely responsible for the existence and complex structure of reefs. As corals grow, they lay down layers of calcium carbonate and serve as ecosystem engineers. The rate of corals' growth has been fast enough on average that they can compensate for the erosion and subsidence of islands and continental margins, creating shallow-water fringing habitats and even isolated atolls where there would otherwise be only deep water (Darwin 1889, Grigg 1982). When reefs form near land they also provide essential services to nearby human populations, including food resources and physical protection from storm surge.

Unfortunately, coral ecosystems are highly threatened by anthropogenic impacts. Climate change, pollution, disease, and other stressors are all contributing to a rapid decline in the complexity and cover of coral reefs worldwide (Pandolfi *et al.* 2003, Van Oppen & Gates 2006, Lesser *et al*. 2007, Barnosky *et al.* 2011, Hughes *et al.* 2017a, Hughes *et al*. 2018). In the face of this, there is an urgent need to learn more about the forces that contribute to corals' resilience and resistance to stressors. However, Scleractinian corals comprise over 1,500 different species that have been evolving for more than 450 million years (Huang & Roy 2015, Figure 0.1). This length of time is longer than that since the origin of mammals (Groussin *et al*. 2017), and similar to the time since the origin of all terrestrial plants (Werner *et al.* 2014, Figure 0.1). Thus it might be that we should expect an extremely diverse array mechanisms by which each species interacts with its environment, and it may be difficult to find characteristics that apply to all or even most species. On the other hand, a comparison to other lineages of only evolutionary time may not be fair. For instance, because modern terrestrial plant species number at almost 400,000 (Paton *et al.* 2008), they represent far more diversity than is represented by corals and their rate of evolution and of adaptation has likely been much faster. Therefore, relative to plants, corals may indeed retain considerable ecological similarities

among their diverse species. The relative influence of divergence time and rate of evolution are unknown for many traits, so studying these relationships is an important step in predicting how the group will fare in the future, and in devising solutions to avoid their loss.

The influence of the microbiome on host health

There is an increasing recognition of the important roles that microbes may play in the health of multicellular organisms. In addition to the classical examples of obligate symbionts and disease-causing pathogens, less intimate associations such as the dynamic communities of microbes in the human gut have been shown to be correlated with shifts in health state such as Crohn's disease (Medzhitov 2007). The realization that these more nuanced relationships are both ubiquitous and medically important has led to a movement to conceptualize all multicellular organisms not as individuals, but as *holobionts*, which straddle a blurry line between discrete biological entities and ecological systems of complex interactions (Gilbert *et al.* 2012).

Given the potential for aspects of the microbiome to influence host health, my PhD has focused on the study of microbes associated with Scleractinian corals. My early contributions to the field include involvement with a study that found that anthropogenic impacts such as nutrient pollution and overfishing cause the microbiome of the coral mucus layer to change (Zaneveld *et al.* 2016). During that study, we noticed that the changes in the microbiome were not always deterministic, but instead reflected decreased microbial consistency among stressed corals. This led to a review of the existing literature that argued that such destabilizing effects may in fact be generally characteristic of the microbiomes of stressed hosts (Zaneveld *et al.* 2017). This work may imply reduced capabilities for active regulation of microbial communities, which supports the idea that the host 'trait' of microbiome composition is not neutral. If a host must expend energy to regulate its microbiome, it has likely been evolutionarily selected to do so.

In my first co-first-author study (Chapter 1; Pollock, McMinds *et al.* 2018), I lay the groundwork for more functional analyses of coral microbiomes by broadly describing how they vary among diverse Australian corals. In Chapter 2, I use these data to address the relationships between the microbiome and coral disease susceptibility. In the progress

of this analysis, I discovered that although some aspects of the coral microbiome do have raw correlations with host species' disease susceptibility, these correlations were not statistically significant after incorporating the host phylogeny into the analysis. Because disease susceptibility itself is strongly patterned phylogenetically, it is difficult to disentangle the effects of shared evolutionary history from the effects of the evolution of a particular host trait. Realizing this, I explore in Chapter 3 a more rigorous and powerful method of investigation, and apply it to the analysis of the distributional patterns of corals' algal symbionts.

The importance of being earnest (about how phylogenetic history contributes to microbiome structure)

Organisms evolve due to drift and due to selection by the environment. Early in the development of Darwin's theories, he realized that 'the environment' also included other, evolving organisms, and that interactions with those organisms played an outsized role in selective processes (Darwin 1859). This has led to the recognition that population and community ecology are inseparable from the study of evolution, and that it is essential to understand the communities among which an organism's ancestors lived in order to understand the roots of its extant traits. As this realization developed, the communities that were studied were naturally composed of multicellular organisms, with the occasional recognition of the importance of disease and of essential microbial mutualisms. However, as technological developments over the last few decades have allowed us to explore the microbial world around us, we have begun to better appreciate the ubiquity and importance to multicellular organisms of interactions with our microbial counterparts.

Because interactions between hosts and microbes are in part governed by heritable traits in each, and because microbes can often be inherited vertically between generations of evolving hosts, it is natural that closely related hosts will tend to associate with similar communities of microbes. Such a pattern has been dubbed *phylosymbiosis* (Brucker & Bordenstein 2013). When the microbes whose identities differ between closely related hosts are at least themselves closely related, the pattern is also characterized by *cophylogeny* (Hafner & Nadler 1988). The implications of these present-day patterns

have been the subject of much debate, in part because there has been disagreement about the degree to which they help inform us about the historical associations between hosts and microbes (Moran & Sloan 2015, Douglas & Werren 2016). The term phylosymbiosis emphasizes interpretations of the pattern as the result of the intimate relationships between the host and its microbiome that are more characteristic of a discrete holobiont than a complex ecological system. Similarly, patterns of cophylogeny have been interpreted as evidence for *coevolution* or *codiversification*, with the various terms sometimes used interchangeably despite their distinct biological meanings. Given these disagreements in interpretation and emphasis, there has also been disagreement about the way we analyze microbiome data in the context of evolution and adaptation.

In A Tale of Two Phylogenies (Hadfield *et al.* 2014), an updated model-based approach for the analysis of host-symbiont relationships was introduced that incorporated the evolutionary relationships of both groups and assessed the ways each was influenced by various interacting phylogenetic scales. Hadfield *et al.* intended for the model to test for patterns in the data that are characteristic coevolution and demonstrated that such a process could indeed be detected with the method. Interestingly, the patterns that they describe as 'coevolution' and 'host phylogenetic interactions' are in fact more nuanced subsets of the patterns of phylosymbiosis and cophylogeny because, if applied to an entire microbiome, both explain more variance when related hosts have more similar microbiome compositions. The additional terms 'non-phylogenetic interaction' and 'parasite phylogenetic interaction' incorporate other distinct patterns of host-microbe specificity that are meant to represent a complete lack of host phylogenetic signal in microbiome composition, and their inclusion in the model allows them to serve as a null hypothesis to test for the presence of host phylogenetic signal. This framework provides a theoretical advantage over many existing tests for phylosymbiosis, such as topological congruence tests or simple Mantel tests, because it comprehensively addresses nuances involving the microbial phylogeny, controls for nonspecific evolution of alpha diversity and microbe prevalence, and can control for and simultaneously analyze the influence of arbitrary other factors.

Because of these theoretical advantages, I explored the possibility of applying the

Hadfield model to a study of the bacterial microbiome of Scleractinian corals, and came across a number of problems with the idea. One was immediately apparent: the MCMCglmm implementation of the model does not scale beyond a few dozen species of microbes or hosts, at which point runtime and memory usage started to become prohibitive. I thus decided to select a small number of potentially interesting subsets of the microbiome for analysis, and another problem was revealed: what arbitrary point in a phylogeny of bacteria should I choose for analysis of its descendants? At these smaller scales, a signal of cophylogeny is at least *consistent with* coevolution or codivergence of the entire group with its hosts, but that signal could be lost if bacterial relatives are included in the analysis that diverged prior to initiation of the group's more intimate relationship. Ultimately, as recounted in Chapter 1, I showed that one subset (based on the Greengenes taxonomic name *Endozoicimonaceae*) had a strong signal of cophylogeny, but that if the mechanism of codivergence was at all responsible for the pattern, it was likely limited to a small subclade of *Endozoicimonaceae* codiverging with just one subclade of the hosts (Chapter 1; Pollock, McMinds *et al*. 2018). This determination, however, was necessarily somewhat subjective, because I did not have a formal framework to detect whether specific subclades demonstrated patterns that were distinct from the microbiome as a whole or from other potential slices of the phylogenetic tree. I also found in that chapter that 20 out of 24 other bacterial families did not show signs of cophylogeny. However, I recognized that there was room for methodological improvement in part because the Hadfield model and other existing methods have no way to deal with uncertainty in phylogenetic reconstructions. Lack of a strong signal at higher levels in the other bacterial families analyzed could simply have been the result of an incorrect topology in one or both of the host and microbial trees.

With these issues in mind, I began developing a new model that had the potential for greater scaling, the flexibility to detect sub-cladal changes in patterns of cophylogeny or phylosymbiosis, and the ability to incorporate phylogenetic uncertainty. In doing so, I also discovered potential for improvement in the conceptual framework of the Hadfield model and of the idea of phylosymbiosis more generally. I reasoned that if one of the goals of analyses of cophylogeny is to estimate the degree of intimacy between two groups of organisms, then any such analysis requires a biologically meaningful reference

for comparison. Since most analyses of phylosymbiosis are not comparative in nature, they instead test their patterns against a null model that is interpreted as a complete lack of phylogenetic signal. However, the practical definition of 'zero' phylogenetic signal is entirely dependent on the scale at which the tips of the host and microbial phylogenies are defined or sampled (e.g. see Groussin *et al.* 2017). This scale often must be either chosen subjectively or results from an essentially arbitrary limit of resolution in the data. For instance, as I report in Chapter 1, of 25 bacterial families tested for cophylogeny with the Hadfield model, only four had significant signal in the cophylogeny term. However, 23 out of 25 families had some form of host specificity, such that particular bacterial types were consistently associated with some coral taxa, but not others. Most of these families did not have significant cophylogeny terms because the observed host specificity did not involve groups of hosts above the genus level resolution of the host phylogeny used in the analysis. In fact, however, even these effects can be interpreted as interactions between phylogenies, because all samples within a host genus are of course phylogenetically related to one another, and every observation of a bacterial sequence variant derives from individual cells that are phylogenetically related to one another below the resolution of the sequenced marker gene. Thus, all four 'host specificity' effects in the Hadfield model can be thought of as terms representing phylosymbiosis or cophylogeny, each at a different combination of discretized scales. Similarly, the original tests for phylosymbiosis, based on topological measures such as the Robinson-Foulds metric (Brooks *et al.* 2016), can often be arbitrarily coerced into a significant positive result. One method would be by increasing the resolution of the host phylogeny; e.g. by transforming the tips of the host phylogeny into multichotomies where each new tip corresponds to a sample rather than a species or individual. If I may partake in some *reductio ad absurdum*, I will note that even two 'independent' samples from the same individual host could be considered 'phylogenetically related': if the host cells in the immediate surroundings of each sample were different, then the microbes in each sample were interacting with clonally derived, but still distinct, biological host entities. Thus, any effect that is consistent within an individual could be considered evidence for phylosymbiosis. Although there is clearly a more precise biological interpretation for tests conducted with pools of observations at the species level (does phylogenetic history

at a scale greater than species influence the microbiome?), there would be an even more precise interpretation if host species are pooled into family-level groups, and then each family is analyzed comparatively ('does phylogenetic history *at a scale greater than family* influence the microbiome?', and 'does phylogenetic history *at a scale greater than species but less than family* influence the microbiome in family X *more than* in family Y?'). I use extreme examples only to illustrate that explicit consideration of scale is essential to understanding the results of a test for phylosymbiosis, or phylogenetic signal more broadly. Using more precise terminology and framework for interpretation allows us to better understand the scales, and potentially mechanisms, under consideration.

The field's recent discussions of phylosymbiosis have productively heightened awareness of how important it is to consider shared evolutionary history when conducting microbiome studies. However, of the many ways that evolutionary history may be manifest in a dataset, the patterns referred to as phylosymbiosis are neither all-inclusive (e.g. they do not inherently consider the history of the microbial partners) nor diagnostic (patterns that result from many distinct processes fall under its broad definition). Because of this, and because nuances of interpretability and scale are already a frequent subject in the discussion of phylogenetic signal more generally, I ultimately came to agree with critics of the holobiont framework, like Douglas & Werren (2016) and Moran & Sloan (2015), that rather than emphasize the detection of phylosymbiotic patterns broadly, a better focus would be the development of models with parameters that can be more directly associated with biological processes or mechanisms.

Thus, in my third chapter I explore a method of microbiome analysis that focuses less on hypothesis testing for the presence of phylosymbiosis, and more on controlling for phylosymbiosis and other forms of phylogenetic signal while testing for more nuanced patterns and processes. For example, I decided not to incorporate the Hadfield model's use of 'non-phylogenetic' effects, because they assumed a mechanism of a single, discrete increase in variance, at a potentially arbitrary point in time, uniformly across all tips on a phylogeny. In most cases I have considered, this does not seem like a biologically reasonable process. It does seem reasonable in the case where the tips of the tree correspond to discrete entities—such as individuals or *some* species—that may be

influenced by a large number of random latent effects. However, corals are known to have very blurry species boundaries, and they are often subject to large amounts of gene flow and morphological gradients between nominally different species. Bacterial species are even less well-defined as discrete entities. Even when relatively discrete species exist, the data we generate to describe them (such as 16S sequence variants or pools of similar variants) certainly do not consistently represent such discrete entities—some lineages of bacteria have extremely different metabolic capabilities among cells that share identical 16S sequences, whereas other lineages have extremely consistent genomic content even as their 16S sequences are highly diverged. Therefore, I decided to more flexibly model the deviations from strict Brownian evolution that are captured by the Hadfield model's 'non-phylogenetic effects'. I instead include other biological processes that reduce phylogenetic signal, such as a total-variance–stabilizing Ornstein-Uhlenbeck effect (Ives & Godfray 2006), variation in the rate of evolution along branches, and discrete evolutionary leaps associated with divergence events.

Although correlations between host and microbiome characteristics that are detected with the model still cannot definitively address questions of causation, accounting for shared evolutionary history can at least place a vote against the phylogenetically structured lurking variables that frequently confound such analyses. The development of this framework is an important step towards a more rigorous analysis of the associations among traits of the microbiome and other host traits such as disease susceptibility.

Figures and Figure Legends

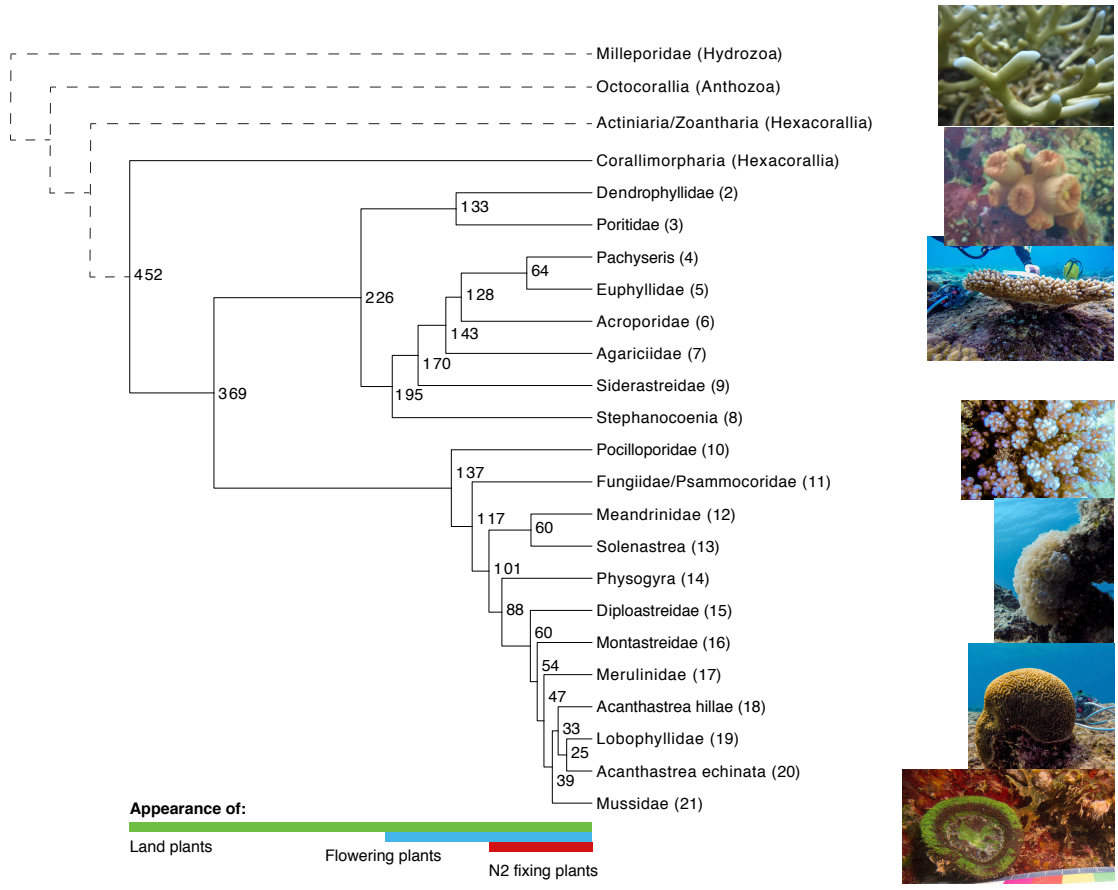


Figure 0.1. Divergence times in the Scleractinia relative to significant events in the evolution of terrestrial plants.

A time-calibrated coral phylogeny, simplified from Huang & Roy 2015 and Fukami 2008, shows that the origin of modern stony corals (Scleractinians) occurred around the same time as the origin of all terrestrial plants (from Zanne *et al*. 2013). Two other ecologically important events in the evolution of plants, the origin of flowering plants and the origin of nitrogen-fixing plants, occurred after the divergence of the two modern coral clades and after the origin of many individual coral families, respectively (see Werner *et al.* 2014). Numbers at nodes within the phylogeny represent estimated millions of years before present of each divergence event (from Huang & Roy). Numbers following taxonomic names at the tips correspond to the numbered clades from Fukami 2008. Dashed lines are taxonomic relationships to outgroups not included in the molecular

phylogenies of Huang & Roy or Fukami. Photos demonstrating the morphological diversity of corals are on the right. All photos were taken by author except the third, which was taken by Dr. Jesse Zaneveld.

Coral-associated bacteria demonstrate phylosymbiosis and cophylogeny

Authors: F. Joseph Pollock[1#], Ryan McMinds[2#], Styles Smith[1], David G. Bourne[3,4], Bette L. Willis[3,5], Mónica Medina[1,6], Rebecca Vega Thurber[2], Jesse R. Zaneveld[7*]

[1]Pennsylvania State University, Department of Biology, 208 Mueller Lab, University Park, PA 16802, USA

[2]Oregon State University, Department of Microbiology, 226 Nash Hall, Corvallis, OR, 97331, USA

[3]James Cook University, College of Science and Engineering, Townsville, Queensland 4811, Australia

[4]Australian Institute of Marine Science, Townsville, Queensland 4810, Australia

[5]ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Queensland 4811, Australia

[6]Smithsonian Tropical Research Institute, Smithsonian Institution, 9100 Panama City PL, Washington, DC 20521, USA

[7]University of Washington, Bothell, School of Science, Technology, Engineering, and Mathematics, Division of Biological Sciences, UWBB-277, Bothell, WA 98011, USA

[#]Contributed equally

[*]Corresponding author

Nature Communications 9:4921. 2018. doi:10.1038/s41467-018-07275-x

Abstract

Scleractinian corals' microbial symbionts influence host health, yet how coral microbiomes assembled over evolution is not well understood. We survey bacterial and archaeal communities in phylogenetically diverse Australian corals representing more than 425 million years of diversification. We show that coral microbiomes are anatomically compartmentalized in both modern microbial ecology and evolutionary assembly. Coral mucus, tissue, and skeleton microbiomes differ in microbial community composition, richness, and response to host vs. environmental drivers. We also find evidence of coral-microbe phylosymbiosis, in which coral microbiome composition and richness reflect coral phylogeny. Surprisingly, the coral skeleton represents the most biodiverse coral microbiome, and also shows the strongest evidence of phylosymbiosis. Interactions between bacterial and coral phylogeny significantly influence the abundance of 4 groups of bacteria– including *Endozoicomonas*-like bacteria, which divide into host-generalist and host-specific subclades. Together these results trace microbial symbiosis across anatomy during the evolution of a basal animal lineage.

Introduction

Since their first appearance around 425 million years ago, scleractinian ('stony') corals (Cnidaria: Hexacorallia: Scleractinia) have radiated into more than 1,500 species, many of which serve as the major architects of coral reef ecosystems worldwide[1]. Modern corals harbor complex communities of microorganisms, including dinoflagellates, fungi, bacteria, and archaea which are collectively termed the coral microbiome[2]. Shifts in the composition of the coral microbiome and virome are linked to changes in coral health, disease, and resistance to stressors[3–6]. It is likely that ancestral corals also harbored complex and functionally important microbial communities. Yet much remains to be understood about how these coral-microbe symbioses evolved, and which key factors influence microbial communities in modern corals. Coral diversity is too great to individually assess the biotic and abiotic factors that maintain the microbiome of every coral species. The present challenge is thus to uncover general rules for the assembly of coral microbiomes that inform estimates of the effects of microorganisms in understudied portions of the coral tree. However, disentangling the many host and environmental

features that influence the microbiome requires large and methodologically consistent surveys of phylogenetically diverse corals across geography.

Many host-microbial symbiosis studies find correlations between host phylogenetic relationships and microbial community composition, a pattern known as phylosymbiosis[7]. Phylosymbiosis has been reported for the root microbiome of flowering plants[8]; the mesohyl of marine sponges[9]; insect microbiomes[7,10]; and the gut microbiome of terrestrial mammals[11] (including *Peromyscus* deer mice[7] and wild hominids[7]). Phylosymbiotic patterns can be explained by several mechanisms, including codiversification of abundant microbial lineages with their hosts, filtering of microbial communities by host traits, or coupling between host phylogeography and environmental effects on the microbiome[7,10,12]. We are only beginning to differentiate these alternatives[10], and studies accounting for the joint effects of phylogeny, geography, and host traits are sorely needed. Moreover, different animal secretions, tissues and organs typically harbor distinct microbiomes (e.g. [13]) that may also show different patterns of phylosymbiosis, although this possibility has not yet been fully explored.

The phylum Cnidaria diverged prior to the bilaterian radiation. Thus, scleractinian coral microbiomes represent a key piece in the broader puzzle of how animal microbiomes arose. Coral mucus, tissue, and skeleton show distinct microbial community composition (e.g. [14,15]), affording the opportunity to test whether they also show different patterns of phylosymbiosis. Additionally, the high diversity and wide geographic distribution of reef-building corals presents a natural experiment for testing how host traits and environmental context influence the microbiome, and are invaluable resources for understanding how modern host-microbial symbioses evolved.

Scleractinian corals have been diversifying for longer than some more commonly studied symbiotic systems such as flowering plants and placental mammals[16]. Their microbiomes are known to be partially species-specific (e.g. [14]), and reports from other Cnidaria, such as gorgonians, suggest potential codiversification with *Endozoicomonas* symbionts[17,18]. Yet comparisons of six species of coral and an octocoral outgroup found microbiome similarities that seemed to better align with morphology than phylogeny[19], suggesting a strong influence of host traits on the microbiome. Whether scleractinian corals show

phylosymbiosis in overall community composition or cophylogeny with specific bacteria or archaea has not yet been definitively established.

The abundance of overlapping factors that affect the coral microbiome is difficult to disentangle. Many host traits are highly correlated with one another due to phylogenetic constraints, and many environmental variables are correlated due to large-scale patterns of climate and geography. Thus, analyses of these variables cannot be conducted in isolation.

We designed a comprehensive sampling and analysis strategy that asked how the microbial communities residing in the mucus, tissue, and skeleton of diverse Australian corals were shaped by host phylogeny, host functional traits, geography and environmental variables. We collected DNA samples from the mucus, tissue, and skeleton of phylogenetically diverse scleractinian species, as well as selected outgroups and environmental references. We sequenced 691 16S rRNA gene libraries from these samples, primarily targeting bacterial and archaeal members of the microbiome. We paired these microbiome data with a multigene molecular phylogeny of scleractinian corals[20], coral functional traits from the Coral Trait Database[21], and extensive *in-situ* metadata (Methods)[22]. For questions that were sensitive to host phylogeny, we integrated these diverse datasets using phylogenetic Generalized Linear Mixed Models (pGLMMs). This approach provided a unified Bayesian framework in which to test hypotheses in coral-microbe coevolution and the influence of various environmental factors on coral-microbe symbiosis.

We show that coral microbiomes differ in richness, composition, and consistency across anatomy. In all anatomical compartments, both host and environment influence the microbiome. However, the relative influence of host vs. environmental parameters varies strongly across anatomy. We confirm phylosymbiosis in coral tissue and skeleton microbiomes, yet also present evidence that host-microbial cophylogeny influences microbial abundance for only a select subset of bacterial taxa associated with corals. Notably, that subset includes certain host-specific subclades of the prominent coral symbiont *Endozoicomonas.* Together, these results help to clarify how the evolution and ecology of the coral microbiome varies across anatomy.

Results

Data collection and workflow

Coral, water, and sediment samples were collected from 21 sites around Australia spanning 17° of latitude. A total of 236 coral colonies were sampled from 32 scleractinian and 4 cnidarian outgroup taxa representing both the Hexacorallia and Octocorallia (Supplementary Data 1). Hexacorallia (anemones, corallimorpharians, zoanthids and scleractinian corals) and Octocorallia (gorgonians) are both monophyletic groups within class Anthozoa. A subset of corals was resampled at Lizard Island in summer and winter to assess seasonal effects. Up to 162 host and environmental metadata parameters were recorded or calculated for each sample (Supplementary Data 2). Combined, these data represent more than 425 million years of coral evolution[20].

A workflow summarizing the major analytical steps is presented in Supplementary Fig. 1.1. Coral samples were partitioned into mucus, tissue, and skeleton compartments (Methods), and sequenced alongside water and sediment samples from the same reefs, yielding a total of 691 samples for small subunit ribosomal RNA (16S rRNA) gene sequencing. These included 227 mucus samples, 223 tissue samples, 230 skeleton samples, and 11 additional reference samples (e.g. sediment and water; Supplementary Data S1). All samples were subjected to identical DNA extraction, PCR amplification using 515f/806r primers specific to the V4 region of the 16S rRNA gene of bacteria and archaea[23], and Illumina MiSeq sequencing. We note that despite the utility of 16S rRNA gene surveys, they are estimated to miss ~10% of environmental microbes[24], including certain archaea and the newly uncovered bacterial candidate phylum radiation[25].

Corals are regarded as challenging targets for DNA extraction. However, we found that the Earth Microbiome Project DNA extraction protocol provided sufficient DNA for analysis in most samples. After quality control, sequencing resulted in a total of 9,441,738 microbial reads (per sample median: 14,010; per sample mean: 13,664) partitioned across 129,305 unique OTUs (Methods, 97% similarity cutoff).

To avoid biases due to sequencing depth, we rarefied to even read depth (1000 sequences per sample) for most analyses (Supplementary Note 1). This strategy is conservative, in that it trades minimization of false positives for some loss of power. We also tested

alternative rarefaction depths for characterization of core microbiomes (Supplementary Data 3), comparison of multivariate dissimilarities (Supplementary Data 4) and α-diversity across compartments (Supplementary Data 5). In the specific case of differential abundance testing, we either rarefied at 1000 reads/sample or used a parametric model without rarefaction (i.e. all pGLMMs, Methods) to maximize power from read depth in each sample. In total, we detected 69 microbial phyla associated with scleractinian corals (i.e. excluding outgroups), with 56.5% of the sequenced microbes in an average sample represented by Proteobacteria, while all Archaea represented just 2% of observed sequences (Supplementary Note 2).

To compare microbial community structure to host trees, we inferred a coral phylogeny using coral mitochondrial 12S rRNA gene sequences identified in our amplicon libraries (Methods), but constrained to match the topology of the multigene molecular phylogeny of corals published by Huang and Roy[20]. Typically, these unique mitochondrial sequences (mitotypes) had a resolution of around the coral genus level (but in some cases resolved species and intra-specific lineages), and were consistent with visual taxonomic identifications of the host. This had the effect of mapping this study's samples to the multigene Huang and Roy phylogeny wherever possible, while also estimating branch lengths among additional outgroup taxa and allowing for inclusion of samples that were not visually identifiable to the species level. A conceptually similar procedure (Methods) mapped microbial reads to the Greengenes 13_8 reference phylogeny[26].

Anatomical variation in drivers of coral microbiome structure

Coral tissue, mucus, and skeleton microbiomes differed in richness (Fig. 1.1a) and microbiome composition (Fig. 1.1b). Surprisingly, the coral endolithic skeleton was richer in microbial diversity than the tissue microbiome (Fig. 1.1a; Supplementary Note 3). Differences in microbiome composition between compartments were robust to choice of multivariate dissimilarity measures (Adonis permutational p for Weighted UniFrac, Unweighted UniFrac and Bray-Curtis distance matrices < 0.001; Supplementary Note 4).

Compartments differed in core microbiome membership (Supplementary Note 5), the fraction of the microbiome that was core (Supplementary Figure 1.2a) and inter-colony variability (Supplementary Figure 1.2b). Observed core microbiomes were consistent

with past reports of the coral mucus core microbiome (Supplementary Note 6). Core microbiome analysis also confirmed the presence of Candidatus *Amoebophilus* (an intracellular symbiont of eukaryotes) as present in >50% of tissue microbiomes (consistent with [14]; see Supplementary Note 7). Overall, mucus microbiomes were notable for their relative stability between colonies and their high abundance of core vs. variable microbes (Supplementary Note 8).

Across all compartments, host species was the single most important variable structuring the coral microbiome in our data (Fig. 1.3, Supplementary Note 9, Supplementary Data 4). Broader taxonomic levels were also associated with microbiome composition, with more specific taxonomic levels always explaining more microbiome variance than more general taxonomic levels. This finding held across several dissimilarity metrics and rarefaction depths, and in all 3 compartments (Supplementary Note 10, Supplementary Data 4). The influence of the coral host was thus a commonality of coral mucus, tissue, and skeleton microbiomes.

However, microbiomes associated with the three portions of coral anatomy differed in the *extent* of their relative responsiveness to host vs. environmental factors (Fig. 1.1c; Supplementary Fig. 1.3; Supplementary Data 6; additional discussion in Supplementary Note 11). For each environmental and host parameter, we tested its relative influence on coral mucus, tissue, and skeletal microbiomes (Fig. 1.1c). We then clustered host and environmental parameters in terms of their effects on the microbiome across compartments. Intrinsic host-based traits clustered separately from environmental traits. This was driven by the fact that environmental factors (e.g., season, temperature and turf algal competition) had a stronger influence on mucus microbiomes than tissue or skeleton microbiomes; whereas the coral species and its functional traits (e.g., growth form and disease susceptibility; Fig. 1.1c) had a stronger influence on tissue and skeletal microbiomes than mucus. Intriguingly, the diverse endolithic skeletal microbiomes were nearly as responsive to many host traits as the tissue microbiome (Fig. 1.1c), and showed the strongest response to the divide between Robust and Complex clade corals.

The finding that coral anatomical compartments differ in overall community composition raises the possibility that one could predict a given sample's compartment from

knowledge of its microbial community. To quantify the accuracy with which a sample's compartment can be predicted from bacterial community membership at the genus level, a supervised classification model was developed using random forest analysis (a machine learning method). This model was 74.3% accurate – a 2.58-fold improvement on error rates compared to random guessing -- demonstrating that a predictable set of bacteria are shared within compartments but differ among them.

We also used machine learning methods to quantify how much information the microbial community of each compartment conveyed about a suite of categorical host and environmental traits and host physiological and phylogenetic parameters (Supplementary Data 7). Consistent with our dissimilarity analysis (i.e., Fig. 1.1c), tissue microbiomes were better predictors of host factors like host genus (34% accuracy, 1.35x lower error rates than random guessing) and vertical transmission of *Symbiodinium* (86% prediction accuracy; 3.41-fold more accurate than random guessing) than were mucus microbiomes (Supplementary Data 7). Exploration of anatomical differences in coral microbiomes using machine-learning methods also revealed that the coral skeleton microbiome could better predict the deep phylogeny of the coral host (i.e. membership in the 'Complex' or 'Robust' clade) than could the microbiome of the coral tissue (Supplementary Results, Supplementary Data 7). Conversely, mucus communities were much better predictors of environmental features like contact with turf algae (82% accuracy; 2.14x lower error rates than random guessing) and sampling location (53% accuracy; 1.46x lower error rates than random guessing).

Together these multivariate and machine learning results clarify that while host species influences the microbiome across anatomy, the extent of host vs. environmental influence on coral microbiomes is not consistent in coral mucus, tissue, and skeleton. They further suggest that mucus microbiomes are useful for detecting environmental perturbations and that skeleton communities warrant greater attention as a diverse community strongly structured by host traits. Because coral compartments differed in both their composition and responsiveness to host and environmental variables, we report results of all subsequent analyses separately for each.

Latitude and colony size influence the coral microbiome

In addition to these general observations, two specific findings emerged that bear mentioning (Fig. 1.2). First, the latitude of the sampling location significantly influenced the richness and composition of coral microbiomes. Moving away from the equator, coral microbiomes became less rich (Fig. 1.2a). This effect was significant for coral mucus and tissue, but not skeleton microbiomes, even after accounting for the uneven distribution of species across locations (Fig 1.2a and legend). In addition to its effects on richness, latitude had a small but significant influence on microbiome composition, accounting for between 1-4% of variance in microbiome composition (Weighted UniFrac adjusted Adonis R^2), depending on the compartment (Fig. 1.1c). Phylogenetic GLMMs were fit separately to each compartment and showed that more bacterial genera were positively than negatively correlated with latitude (11%, 13%, and 9% of genera positively correlated with latitude in mucus, tissue, and skeleton, compared to 4%, 1%, and 4% negatively correlated with latitude in these compartments) (Fig. 1.2; Supplementary Data 6). Together these results suggest that patterns of diversity in coral microbiomes may mirror latitudinal diversity gradients seen in free-living communities[27].

We also found that proportionally larger corals (those closer to their species' maximum recorded size) showed differential microbiome composition and richness relative to smaller specimens (Fig. 1.2b). Effects of coral relative size on microbiome richness were significant in coral tissue and skeleton, but not mucus, after accounting for phylogeny (Fig. 1.2b). Coral size also had minor but statistically significant effects on microbiome composition in coral skeleton (Supplementary Note 12; Supplementary Data 4). Two bacterial genera, *Aurantimonas* and *Balneola,* were significantly reduced in larger corals (Supplementary Data 6). *Aurantimonas* has been proposed as the causative agent of White Plague Type II[28] and *Balneola* was previously identified as an indicator of sewage pollution in the Red Sea[29]. These trends may reflect increased vulnerability of smaller corals to opportunistic pathogens or may simply reflect normal shifts in microbiome composition over the course of coral development[30].

Coral phylogeny structures microbiome richness and composition

Phylosymbiosis refers to the evolutionary pattern in which the phylogeny of a related group of host organisms correlates with changes in multivariate community dissimilarities among their microbiomes[31]. Mantel tests assessed phylosymbiosis in the coral microbiome (Supplementary Data 8). These tests quantify the correlation between matrices of coral host phylogenetic distances and multivariate dissimilarity as measured by the Bray-Curtis or Weighted UniFrac measures. Using Bray-Curtis dissimilarities, more closely related corals had more similar microbiomes in both tissue (Mantel $r = 0.16$, $p = 0.0001$; Supplementary Fig. 1.4a, dashed red regression line) and skeleton compartments (Mantel $r = 0.18$, $p = 0.0001$), but not mucus (Mantel $r = 0.02$, $p = 0.18$). Using the phylogenetically-aware Weighted UniFrac method deemphasized fine variation at the tips of the microbial tree, resulting in a significant signal of phylosymbiosis in skeleton, but not tissue or mucus microbiomes (Supplementary Data 8).

Our Mantel test results demonstrate patterns consistent with phylosymbiosis in skeleton and perhaps tissue microbiomes, but do not clarify the evolutionary scales over which these patterns emerged. Therefore, we used Mantel correlograms to assess how these correlations varied across multiple scales of phylogenetic divergence. Across all compartments and dissimilarity measures, microbiomes of the most closely related coral hosts were significantly more similar than expected by chance (Mantel $r > 0$; $p < 0.05$). In general, tissue and skeletal microbiomes became gradually more dissimilar throughout the entire range of host phylogenetic distances (Supplementary Fig. 1.4b, Supplementary Data 8). Mucus microbiomes, on the other hand, did not become more dissimilar as host phylogenetic distance increased past the second distance class (Supplementary Data 8).

We employed a similar procedure to test the evolution of microbiome richness in corals. As richness is a univariate rather than multivariate quantity, we conducted these tests using Moran's *I* as a measure of univariate autocorrelation. These phylogenetic correlograms demonstrated that, like community composition, richness was significantly more similar among closely related corals than expected by chance (Moran's *I* 95% lower CI > 0; Supplementary Fig. 1.4c, red confidence intervals). Additionally, richness was more *dissimilar* than expected among samples that were separated by phylogenetic

distances of approximately 0.2 to 0.25 (Moran's *I* 95% CI < 0; Supplementary Fig. 1.4c; blue confidence intervals), which corresponded roughly to between-family distances in our tree. At greater phylogenetic distances, they were no more or less similar than expected. These trends were consistent across coral mucus, tissue, and skeleton microbiomes. Coral microbiome richness is therefore influenced by the evolutionary histories of host corals. Importantly, the scales of phylogenetic divergence at which these effects appear, suggest that the radiation of modern reef-building coral families (between roughly 25 and 65 mya) was accompanied by large changes in microbiome richness, with changes continuing to accumulate during more recent speciation events. What's more, these results demonstrate that the phylogenetic histories of corals partially constrain the composition of their tissue and skeletal microbiomes and the richness of all coral compartments. In other words, corals and their microbiomes exhibit phylosymbiosis[32].

Limited phylogenetic signal in the distribution of bacterial genera

Phylosymbiosis results from a number of different mechanisms: the steady evolution of host traits that directly influence the microbiome (e.g., by excluding certain microbes); spatial patterning of hosts that indirectly influence the microbiome via environmental or ecological interactions (e.g., dispersal to areas with intensive turf algae competition), or long-term codiversification between hosts and specific microbial symbionts[12,41].

To address these alternatives, we tested all microbial genera for associations with tips of the coral tree (host identity) or wider regions of the coral tree (host phylogeny). Genera were defined based on Greengenes taxonomic annotations, and therefore included some imprecise pools of unannotated taxa, but were deemed sufficient for our intended analyses (a complementary fine-scale approach is pursued below). Both host identity and host phylogeny were assessed using the genus-resolution 12S mitochondrial RNA gene markers extracted from amplicon libraries for each sample, and pGLMMs (see Methods) were used to separate the effects of environmental and physiological variables from host effects.

Even after accounting for some of the most important environmental and physiological factors from the multivariate analysis (e.g., geographic region, turf algal contact, disease susceptibility, and maximum corallite width), most microbial genera showed host-

specific abundance patterns (Supplementary Table 1.5). Yet most coral-associated microbes were correlated with host identity, not host phylogeny. Phylogenetic GLMMs estimated that the abundances of between 62% (tissue) and 75% (mucus) of microbial genera were significantly correlated with at least one host mitotype (host identity), depending on the compartment analyzed (Supplementary Data 6). For example, 100/446 microbial genera detected in tissue microbiomes were significantly more abundant in the *Acropora* mitotype than in others. Overall, 276/446 (62%) of microbial genera were associated with a host mitotype, but only 13/446 (3%) of genera in coral tissue were associated with host phylogeny (Supplementary Data 6). Genera associated with host phylogeny include *Candidatus* Amoebophilus (Cytophagales*:* SGUS912), a taxon previously identified as a core coral microbiome member across three species[15]. *Candidatus* Amoebophilus was associated in the skeleton with the coral clade formed by both *Seriatopora* and *Stylophora*, rather than with individual 'host identity' mitotypes (for additional discussion see Supplementary Results). Mucus microbiomes showed fewer genera (1.6%) associated with host phylogeny than tissue, while skeletal microbiomes showed more (4.9%). Taken together, this analysis confirmed that while the coral microbiome is highly host-specific, only a restricted subset of the microbiome members show preferences for entire groups of related corals.

Cophylogenetic analysis identifies coral-bacterial interactions

The above GLMM analyses were conducted at the level of microbial genera, but finer-scale taxonomic variation is likely to exist. Also, the previous analysis identified only the response of microbial genera to host phylogeny, rather than any potential interactions between microbial and coral phylogenies. Therefore, we ran pGLMMs incorporating both coral and microbial phylogeny on fine-scale microbial sequence variants using Minimum Entropy Decomposition (MED)[33]. These methods tested whether corals showed patterns of cophylogeny with any of their microbial associates, which in this context refers to the tendency for groups of related microbes to be associated with groups of related hosts. Such patterns can arise from coevolution or codiversification, and may thus be a sign of intimate symbiosis, mutualistic or otherwise.

Because these analyses are computationally intensive, only the most prevalent microbial taxa were tested. A total of 25 bacterial family-level groups present in >50% of samples from at least one coral compartment were selected for detailed analysis. All but two of the families tested had some form of host specificity, with members either associated with particular coral mitotypes (representing species or genera) or particular regions of the coral tree. More formally, each of these bacterial families showed one of the following interaction effects (Fig. 1.3): host identity by bacterial identity; host identity by bacterial phylogeny; host phylogeny by bacterial identity; or host phylogeny by bacterial phylogeny (i.e., cophylogeny). One-to-one associations between individual bacterial sequence variants and individual coral hosts (i.e., host identity by bacterial identity interaction effects) were only significant in three bacterial groups: Clostridiaceae (mucus), unclassified Myxococcales (mucus), and unclassified Kiloniellales (mucus and tissue). In 19 of 25 families, host identity interacted significantly with bacterial phylogeny, meaning individual coral mitotypes were significantly associated with clades of related bacteria. However, the converse pattern did not occur: no individual bacterial sequence variants were significantly associated with clades of related coral hosts.

Four bacterial groups exhibited significant cophylogenetic effects (i.e. host phylogeny by bacterial phylogeny interaction): Clostridiaceae, *Endozoicomonas*-like bacteria (Endozoicimonaceae in Greengenes), unclassified Kiloniellales, and unclassified Myxococcales (Fig. 1.3, red box). Cophylogeny in *Endozoicomonas*-like bacteria was detected in both tissue and mucus (ICCs, 95% lower bounds: 0.20 and 0.17, respectively) (Fig. 1.3, 1.4). Cophylogeny in Clostridiaceae and unclassified Myxococcales was detected within the coral skeleton only (ICCs, 95% lower bounds: 0.06, and 0.29, respectively), and it was detected in unclassified Kiloniellales in only the tissue compartment (95% ICC lower bound: 0.03).

Together, these results show that while overall coral microbiome composition and richness do track phylogeny, and the majority of microbial genera show significant host-specificity, only a small subset of coral-associated microbial diversity shows larger-scale interactions between coral and bacterial phylogeny.

Endozoicomonas partition into host-generalist and host-specific clades

Endozoicomonas-like bacteria are important coral symbionts[34], and in our data showed the strongest signal of cophylogeny among bacteria found in coral tissues. We therefore analyzed this group in greater depth. Inspection of the phylogeny of *Endozoicomonas*-like bacteria (Methods) revealed two major coral-associated divisions within the group (Fig. 1.4, Supplementary Fig. 1.4): one in which most strains were host-specific (hereafter 'Clade HS' for 'Host-Specific'), and another where most strains had a cosmopolitan distribution across multiple hosts (hereafter, 'Clade HG' for 'Host Generalist).

Within the host-specific clade HS, two bacterial sub-clades were strongly associated with the two major lineages of corals ('Complex' or 'Robust' corals). We have termed these clades of *Endozoicomonas*-like bacteria 'HS-R' for 'Host-Specific: Robust' and 'HS-C' for 'Host-Specific: Complex'. All of these clades and subclades were well-supported by posterior probabilities (posterior probabilities: Clade HG, 1.00; HS-R 0.92; HS-C 0.72) with the exception of Clade HS, which was only weakly supported (posterior probability 0.33).

To further assess the relationship between corals and Clade HS *Endozoicomonas*-like bacteria, we fitted a GLMM that included all corals but only clade HS bacteria, and another that included only the Robust clade corals and clade HS-R bacteria. The cophylogeny terms from both these tests were highly significant (ICCs, 95% lower bounds: 0.34 and 0.21, respectively).

The coral-associated members of Clade HS from this study were more closely related to *Endozoicomonas* strains previously reported to live in symbiosis with diverse non-scleractinian hosts (gorgonians, mollusks, sponges, and other marine invertebrates) than they were to members of Clade HG. Thus it appears that *Endozoicomonas*-like bacteria have formed novel associations with scleractinian corals multiple times throughout their evolution, but that such host-swapping has been a relatively rare occurrence.

Discussion

Scleractinia have been diversifying for almost half a billion years[20]. In about half that timespan, flowering plants appeared, diversified and evolved important specialized microbial symbioses in specific lineages[35]. Here we demonstrate that a phylogenetic framework for analysis of coral microbes can reveal how scleractinian corals' evolutionary history, host traits and the local environment interact to shape coral microbiomes. Our results test longstanding hypotheses that bear on potential coral-microbe coevolution, and add quantitative details and taxonomic breadth to several previously explored patterns in coral microbiology.

We originally hypothesized that corals would show signs of phylosymbiosis throughout their entire phylogenetic history. While our results are in accord with this hypothesis in coral skeleton and tissue, the same is not true for the coral mucus microbiome. Despite documented variability in the chemical composition of coral mucus between species[32], and significant host-specificity in the mucus microbiome, host specificity in the mucus microbiome was limited to relatively recent divergences and was not significantly structured by larger scales of host phylogeny. Importantly, because this analysis focused on the entire Scleractinian order, it did not test whether patterns of phylosymbiosis occur in the mucus within specific coral lineages or at intrageneric timescales generally. In contrast to the patterns in the mucus, the coral skeleton, which has been less intensively studied than mucus and tissue, showed both the greatest microbiome richness and the strongest signal of long-term phylosymbiosis. These findings emphasize that different anatomical regions of animal hosts may show distinct evolutionary patterns. This observation will be relevant for studies in other systems (e.g. mammals) where most studies of host-microbe coevolution have emphasized a single body site (e.g. the distal gut).

Phylosymbiosis can emerge as a consequence of multiple mechanisms, including codiversification of many lineages, microbial habitat filtering by host traits, or the interaction of host and microbial biogeography[12]. We tested whether the most prevalent coral-associated bacteria demonstrated cophylogenetic patterns with their hosts. We used pGLMMs to compare the prevalence of individual sequence variants within particular

bacterial families among diverse coral hosts. This approach allowed us to disentangle the effects of geographic area, cophylogeny, and distinct associations between individual hosts and microbes. Of the 25 bacterial families tested, cophylogentic interactions significantly influenced the abundance of only 4. Thus, although many coral-associated bacteria are host-specific, and the overall composition of coral microbiomes tracks phylogeny (i.e., phylosymbiosis), only a select minority of coral-associated bacterial families show cophylogenetic signals consistent with long-term host-microbe codiversification. This result emphasizes that while host-microbe cophylogeny likely contributes to phylosymbiosis, other factors, such as biogeographic effects and phylogenetically patterned host traits are likely very important in producing this pattern.

That coral microbes differ in their extent of cophylogeny with their host also emphasizes that the microbiome is not a single unit of selection, but instead contains diverse players that vary greatly in the extent of their history of association with the host and one another[12]. Host specificity of certain microbes with extant coral species, community-level phylosymbiosis across the coral microbiome, and cophylogeny of certain microbial lineages with their host are all distinct concepts that should be distinguished. We recommend that observations of phylosymbiosis be accompanied by finer-scale tests of host-microbe cophylogeny in order to identify specific microbial lineages that may have coevolved or codiversified with their hosts. These may warrant additional investigation as potential 'key players' in the microbiome, but a time-calibrated microbial phylogeny would be necessary to test for it.

The microbial families that show signs of cophylogeny may be associated with important host functions that have led to stable symbiotic relationships across extended evolutionary time. The pGLMM methods used here allow identification of such taxa from a broader symbiotic community, even in the absence of strict one-to-one associations between hosts and symbionts, and will be relevant to other study systems. We identified four bacterial lineages displaying signs of cophylogeny with their coral hosts, all of which represent important targets for future study. One lineage, the *Endozoicomonas*-like bacteria, has previously been hypothesized to have codiversified with their coral hosts throughout the evolution of Scleractinia[36]. These results are

relatively consistent with this hypothesis for one subclade of *Endozoicomonas*, but suggest that the abundant variants found in well-studied *Porites* corals are more cosmopolitan in their distribution. A greater geographic breadth of samples and representatives of the many azooxanthellate scleractinians will help inform this notion further, and a test of codiversification specifically will require a better-resolved and time-calibrated *Endozoicomonas* phylogeny.

In addition to the *Endozoicomonas*, three other groups of functionally distinct bacteria showed significant patterns of cophylogeny: the proposed mutualist Kilioniellales[37], the predatory 'wolf pack' bacteria Myxococcales, and a group of organisms generally hypothesized to be pathogens of corals, the Clostridiaceae[4,37-39]. Future work on these taxa may provide insight into their broader roles in coral evolution and health.

Our survey of Australian coral microbial diversity provides the most conclusive evidence to date that phylosymbiosis has occurred between corals and their microbiomes. Despite this, cophylogeny between scleractinian corals and their microbial symbionts is likely restricted to a small subset of bacterial families. The results of this survey further quantify the relative influence of host and environmental drivers on the microbial diversity of coral mucus, tissue and skeleton. A still more comprehensive picture of coral microbiology will be gained with future efforts that expand analyses to global sample datasets (including potentially informative samples from deep-water or Caribbean corals), development of improved statistical models (e.g. by relying less on arbitrary taxonomic thresholds; see for instance the emerging 'ClaaTU' method[40]), and connection of these patterns of microbial diversity to other members of the coral microbiome such as *Symbiodinium*. In particular, the addition of deep-water, azooxanthellate corals could fill in important gaps in the phylogeny and help test the generality of phylosymbiosis in coral microbiomes.

Data Availability

Raw sequence data, metadata, OTU and MED representative sequences are publicly available at https://doi.org/10.6084/m9.figshare.c.3855466.v2. Raw sequence data are also deposited at the European Nucleotide Archive under accession number

PRJEB28183. Analysis code is available on GitHub: https://github.com/zaneveld/GCMP_Australia_Coevolution

Acknowledgements

The authors would like to acknowledge Tasman Douglass, Margaux Hine, Frazer McGregor, Kathy Morrow, Katia Nicolet, Cathie Page, and Gergely Torda for their field assistance and Lyndsy Gazda, Jamie Lee Proffitt, Gabriele Swain, and Alaina Weinheimer for their assistance in the laboratory. The authors also acknowledge the staff of the Coral Bay Research Station, Lizard Island Research Station, Lord Howe Island Marine Park, Lord Howe Island Research Station, and RV Cape Ferguson for their logistical support. This work was supported by a National Science Foundation Dimensions of Biodiversity grant (#1442306) to RVT and MM.

Author Contributions

Contributions: RVT, JZ, and MM designed the experiment. FJP, RM, RVT and MM conducted the fieldwork. RM, FJP, and SS generated the libraries. FJP, RM, and JZ performed the data analysis. BW and DB assisted in the sample collection and provided critical data for meta-analysis. FJP, RM, JZ, MM, and RVT wrote the manuscript, and all authors contributed to its editing.

Conflict of Interest Statement

The authors declare no conflict of interest.

Methods

Selection of target sites

We aimed to collect coral specimens spanning coral phylogenetic diversity from a variety of Australian reefs. We targeted collection based on the 21 major coral clades defined in one of the most recent molecular phylogenies available at the start of the project[41]. Many of these monophyletic groups have since been defined as family-level taxa. Corals were collected at several sites on the east and west coast of Australia. These included Ningaloo Reef (Western Australia), Lizard Island, multiple reefs along the northern Great Barrier Reef, and Lorde Howe Island. Samples at Lizard Island were collected in both Summer and Winter, allowing for comparison of seasonal effects at one site across diverse corals.

Collection of metadata

During sampling, each coral, outgroup species, water, and sediment sample was associated with MIxS metadata[42]. This was accomplished by recording standardized metadata about each site prior to dives, and using an underwater metadata sheet (available at https://doi.org/10.6084/m9.figshare.5326870.v1). These metadata included basic features of coral species (as identified in the field), location, depth, water temperature, but also diver annotation of contact with macroalgae, turf algae or cyanobacteria (and the percent of the coral in contact); the presence of any visible tissue loss or disease signs; and coral color (using the Coral Reef Watch color charts[43]. Additionally, photographs of each coral were taken and released via openly accessible third party websites. They are easy to browse and thoroughly keyworded with taxonomy, location, and sample ID metadata on Flickr: https://flic.kr/s/aHsk9mjb54, and permanently archived in raw camera format with a spreadsheet linking filenames to colony names on FigShare at https://doi.org/10.6084/m9.figshare.5318236.v2.

Coral sampling

All coral samples were collected by AAUS-certified scientific divers, in accordance with local regulations. Relevant permit numbers are: CITES (PWS2014-AU-002155, 12US784243/9), Great Barrier Reef Marine Park Authority (G12/35236.1, G14/36788.1), Lord Howe Island Marine Park (LHIMP/R/2015/005), New South Wales Department of Primary Industries (P15/0072-1.0, OUT 15/11450), US Fish and Wildlife Service (2015LA1632527, 2015LA1703560), and Western Australia Department of Parks and Wildlife (SF010348, CE004874, ES002315). Only healthy corals were collected.

One goal of the project was to compare microbial diversity associated with the coral mucus, tissues and skeletons across many coral colonies. Each of these compartments represents a simplification of more complex structure, and much work remains to be done on the finer-scale distribution and dynamics of microorganisms across coral anatomy. For this project, we felt that a consistent reporting of these compartments across diverse corals represented a tractable step forward, given the scale of the project. Mucus was collected by gently agitating the surface of corals for ~30 seconds with a blunt 10 mL syringe. Exuded mucus or surface water (if no visible mucus was exuded) was then

collected by suction. On the surface, settled mucus typically formed a distinct visible layer within the syringe. This was expelled into a cryogenic vial and stored in a dry shipper charged with liquid nitrogen for subsequent processing.

Tissue and skeletal samples were collected from each colony by hammer and chisel, or (for branching corals) by bone shears. These fragments were placed in sterile WhirlPaks and returned to the surface where they were snap frozen in a liquid nitrogen dry shipper until processing. In the laboratory, tissue was washed with sterile seawater (which removed visible mucus and detritus), then separated from skeleton using pressurized air of between approximately 800 and 2,000 PSI (an 'air gun'). Skeleton was sampled using a sterile chisel to isolate a ~1 cm^3 region of skeleton that was not in direct contact with coral tissue. Skeleton samples were collected without regard to endolithic algae presence or absence (i.e., endolithic algae were neither specifically targeted nor excluded). Tissue slurries and skeleton samples were added directly to a MoBio PowerSoil Kit (MoBio Laboratories, Carlsbad, California) bead tube (which contains, among other things, a solution of guanidinium preservative) and stored at -80 °C until DNA extraction.

Sampling of reference samples

Because reef water and adjacent sediment might have an effect on the microbiota of corals from the same reef (especially in coral mucus), reef water and sediment were sampled at multiple sites. Surface seawater samples (1 L) were filtered through 0.22 μm Millipore Sterivex filters (Sigma-Aldrich, St. Louis, MO, USA) and reef sediment samples (2 mL) were collected in sterile cryogenic vials. Samples were snap frozen in a liquid nitrogen dry shipper, and subsequently stored at -80 °C until DNA extraction.

For comparison with corals from the same reef, we also opportunistically sampled non-scleractinian cnidarians from the genera *Millepora* (fire corals), *Palythoa* (zoanthids), *Heliopora* (blue corals), and *Lobophytum* (soft corals).

16S library preparation, sequencing, and initial quality control

DNA was extracted from skeleton, tissue, mucus, and environmental samples using the MoBio Powersoil DNA Isolation Kit. Two-stage amplicon PCR was performed on the V4 region of the 16S rRNA gene using the 515F/806R primer pair that targets bacterial

and archaeal communities[23]. Extraction blank controls were also included in amplification and sequencing for quality assurance. The average concentration of extracted DNA used for PCR was 10.8 with a standard error of 1.2. First, 30 PCR cycles were performed using 515F and 806R primers (underlined) with linker sequences at the 5' ends: 515F_link (5'-ACA CTG ACG ACA TGG TTC TAC AGT GCC AGC MGC CGC GGT AA-3') and 806R_link (5'-TAC GGT AGC AGA GAC TTG GTC TGG ACT ACH VGG GTW TCT AAT-3'). Each 20 µL PCR reaction was prepared with 9 µL 5Prime HotMaster Mix (VWR International), 1 µL forward primer (10 µM), 1 µL reverse primer (10 µM), 1 µL template DNA, and 8 µL PCR-grade water. PCR amplifications consisted of a 3min denaturation at 94 °C; 30 cycles of 45 s at 94 °C, 60 s at 50 °C and 90 s at 72 °C; and 10 min at 72 °C. Next, amplicons were barcoded with Fluidigm barcoded Illumina primers (8 cycles) and pooled in equal concentrations for sequencing. The amplicon pool was purified with AMPure XP beads and sequenced on the Illumina MiSeq sequencing platform (using V3 chemistry) at the DNA Services Facility at the University of Illinois at Chicago.

QIIME (v1.9)[44] was used to process all 16S sequence libraries. Primer sequences were trimmed, paired-end reads merged, and QIIME's default quality-control parameters used when splitting libraries. Chimeras were removed and 97%-similarity OTUs picked using USEARCH 7.0[45], QIIME's subsampled open-reference OTU-picking protocol[46], and the 97% Greengenes 13_8 reference database[26]. Taxonomy was assigned using UCLUST, and reads were aligned against the Greengenes database using PyNAST[47]. FastTreeMP[48] was used to create a bacterial phylogeny with constraints defined by the Greengenes reference phylogeny. Following quality control, 9,441,738 usable reads remained. The number of per sample reads ranged from 2 to 38,523 with a median of 14,010, mean of 13,644, and standard deviation of 7,565. Reads were partitioned across 129,305 unique OTUs (97% similarity cutoff). Sequencing success did not show any obvious trends with regards to host taxonomy or geographic location.

A 'canonical' rarefied OTU table was created and used for all downstream analyses except the linear model analyses. To create this table, OTUs were filtered out of the starting table if their representative sequences failed to align with PyNAST to the

Greengenes database or if they were annotated as mitochondrial or chloroplast sequences. The beta_diversity_through_plots.py script was then used to rarefy the resulting table to exactly 1000 sequences per sample, and to calculate from this rarefied table multivariate dissimilarity measures including Bray-Curtis, Binary Jaccard, Weighted UniFrac, and Unweighted UniFrac. Also from this table, α-diversity statistics were calculated using alpha_rarefaction.py, including the number of OTUs observed, evenness, and Faith's Phylogenetic Diversity.

Mitochondrial annotation and quality control

The primers used in this study were designed to selectively amplify the V4 region of bacterial and archaeal 16S rRNA gene, but we have noticed in many of our studies that they (and other standard primer sets) tend to strongly amplify corals' mitochondrial 12S rRNA gene, which is the homolog of the bacterial 16S rRNA gene. Because our samples included species that were not used in the Huang and Roy 2015[20] phylogeny (including, critically, all outgroup taxa), these 'off-target' host mitochondrial reads were used to inform phylogenetic analyses and for an additional layer of quality-control. First, split_libraries_fastq.py was run on the raw forward reads without any quality trimming. Then, primers and adaptor sequences were removed, and USEARCH used to de-replicate 100% identical sequences. A frequency table was created and the data were filtered to contain only sequence variants with a total count of at least 100. Greengenes taxonomy was assigned to the remaining sequence variants with UCLUST as before, and sequence variants that had no match in the Greengenes database (e.g. putative non-bacterial or archaeal sequences) were isolated. For each host species, sequence variants were manually submitted to NCBI's BLASTn web interface in order of their total abundance, comparing against the entire nr database. If a variant's top 20 hits were annotated as coral mitochondria of any species, the sequence was copied to a FASTA file of host sequences. If all three compartments of a single coral individual of the same putative species still had unclassified variants that were more abundant, then manual annotation of those variants continued until either another coral mitochondrial sequence variant was found or there were no more variants in those samples that were more abundant than the previously annotated mitochondria. Using this method, no host sequences were found for some

species of coral. The process was repeated for these species individually, without first discarding sequences that had counts of less than 100. In this way, mitochondrial sequences were eventually identified for every sample in the study.

Once all host species mitochondrial sequences were identified, the original frequency table of all unique sequence variants was filtered to contain only the identified host sequences. For each individual sample, the most abundant mitochondrial type was determined, and this information was then added to the sample's metadata as its '12S genotype'. Then, all selected host sequences were aligned using MAFFT[49] and *de novo* phylogenies were constructed in BEAST 2.4.2[50], with a chain length of 10 million, thinning interval of 1000, a log-normal relaxed clock model, and the site model selected using bModelTest[51]. The maximum clade credibility tree was selected using TreeAnnotator with a burn-in of 25% and common ancestor heights. This tree was compared to the expected topology (monophyletic Anthozoa, Hexacorallia, and Scleractinia, and otherwise matching the Huang and Roy 2015[20] molecular tree) to identify potential mismatches among the observed sequences and the field species identification. Regions of the tree with topology that differed from expectation were manually inspected.

Using this strategy, two coral individuals were noted whose field identifications placed them in the family Merulinidae, but whose sequence variants were strongly indicative of a relationship with the family Lobophylliidae. In these instances, further analyses verified that the same mitotype was detected in all three compartments of the same individual. Photos from collections in the field were consulted, and both were ultimately determined to have been misidentified in the field and in fact belonged to the genus *Echinophyllia*. Their metadata and annotations were updated to reflect this.

Aside from these two taxa, it was determined that unexpected topology in the *de novo* phylogeny was a result of imprecise resolution of the 12S V4 marker. For example, sequences from *Millepora*, *Palythoa*, and both octocoral species were placed in a monophyletic clade including the Complex corals, though they properly belong as outgroups to all Scleractinia. These errors emphasized the limitations of our opportunistic host sequence data to build a *de novo* phylogeny. Thus, having confirmed identifications

of host species to within the resolution of the 12S marker, a new phylogeny was constructed with the topology constrained to exactly match the Huang and Roy 2015[20] molecular phylogeny and the known relationships of outgroup taxa. In cases where a single 12S genotype belonged to members of a polyphyletic group of taxa, we created separate tips for each monophyletic group. The mitochondrial sequence alignment and BEAST 2.4.2 were used to estimate relative branch lengths on this tree by supplying the starting tree and turning off all topology operators. The resulting tree was used for all phylogenetic analyses. As the branch lengths in this tree are derived from a relaxed clock model and limited sequence data, they are likely to represent some average between divergence times and degree of molecular evolution. Thus, analysis using these branch lengths represents a compromise between assuming correlation of traits is proportional to time since divergence and assuming that correlation of traits is proportional to overall evolutionary change since divergence.

Annotation of coral life history strategy

To assess connections between coral traits and microbiome structure, coral species sampled in this study were mapped to functional traits. These host features were added to the microbial mapping file, and used for tests of microbiome structure vs. host traits.

Coral life history strategies from Darling *et al*.[52] ('weedy', 'competitive', 'stress-tolerant', and 'generalist') were digitized and associated with coral species. Some species have recently been moved between genera based on updated phylogenetic evidence[53]. In these cases, both the original species name and the revised name are noted in the metadata. In some cases, species sampled were not annotated in Darling *et al*.[52]. These were not assigned an annotation if annotated members of the same genus had mixed life-histories, or if only a single species of the same genus had been annotated. In cases where at least two members of the genus had been annotated and all annotated members shared the same life-history strategy, the same annotation was assigned to other members sampled from the genus.

Annotation of coral functional traits

Metadata associated with each species sampled was annotated with 28 reproductive, biogeographic, and morphological traits from the Coral Trait Database (CTDB) v.

1.1.123. These traits included basic details on coral distribution (abundance worldwide and on the Great Barrier Reef, range size, northerly and southerly limits, upper and lower depth limits), reproduction (sexual system, mode of larval development, propagule size, presence of Symbiodinium in propagules), phylogeny (genus and species ages), morphology (growth form, skeletal density, corallite maximum width, maximum growth rate) and conservation (IUCN Red List Category).

Adonis analysis of factors affecting microbial composition

We tested the influence of multiple host and environmental factors on the microbial community of each compartment. These results are presented in Fig. 1.1c and Supplementary Fig. 1.2, while the raw underlying data is presented in Supplementary Data S3. Throughout the analysis care was taken to account for the effects of rarefaction depth (we tested the robustness of the results at rarefaction depths of 1000 or 10000 sequences/sample), β-diversity distance measure (we tested three distance measures), the degrees of freedom in each parameter (we used adjusted R^2 values to account for differences in degrees of freedom), and to stringently control for the number of comparisons performed (using Bonferroni correction).

β-diversity distance matrices were calculated from separate OTU tables for coral mucus, tissue, and skeleton (outgroups and environmental samples were not included in this analysis). We calculated distance matrices using Weighted UniFrac distances, Unweighted UniFrac distances or Bray-Curtis dissimilarities. Then, for each host or environmental factor, the distance matrix was filtered to just those samples for which metadata were available (i.e. excluding 'Unknown' values). This prevented 'Unknown' values from being treated as a bona fide category in downstream statistical tests. The filtered distance matrix was then tested for clustering by factor using permutational tests (as implemented in Adonis in QIIME 1.9.1; 999 permutations per test). Because categories that can take on more values (e.g. species) are biased upwards in raw R^2 values, we calculated adjusted R^2 values for each category. These adjusted R^2 values are primarily useful in that they allow for fair comparison between factors with differing degrees of freedom. Therefore, we present adjusted R^2 values when comparing factors, but raw R^2 values when discussing the percentage of variance explained (adjusted

R^2 values can no longer be interpreted as percent of variance explained). Importantly, we took care to separately filter the QIIME mapping file to exclude 'Unknown' values for each parameter under consideration. Failure to do so can lead to continuous variables (columns containing only numbers) being treated as categorical in QIIME, due to the presence of text values. This in turn can strongly influence inferred R^2 values.

Summary of Adonis analysis of microbial Beta-diversity

To present a summarized view of the Adonis analysis of microbial community beta-diversity, we compiled the R^2 and p value obtained from each individual Adonis analysis. In Supplementary Fig. 1.1 the compiled adjusted R^2 are presented in a heatmap. In Fig. 1.1c, we compared the relative influence of each host or environmental parameter on different host compartments by Z-score normalizing Adonis R^2 values within columns. This has the effect of showing which compartments are most strongly influenced by a particular factor, independent of how influential that factor was overall. We present both views into the data because the unnormalized Adonis R^2 values better emphasize the absolute magnitude of the microbiome response to each factor, whereas the Z-score normalized values better illustrate common patterns across compartments in host vs. environmental parameters. We emphasize that in both Fig 1.1c. and Supplementary Fig. 1.2, clustering of rows and columns was performed without any prior specification of which factors were host vs. environmental. Thus, observed clustering of host vs. environmental factors emerges from features of the microbial communities themselves.

Machine learning analyses

All machine learning analyses were conducted through the supervised_classification.py script in QIIME (v1.9)[44]. This script implements random forest classification, which is a machine learning method for supervised classification. We used default parameters, which classify samples using inferred forests of 500 decision trees. We applied random forest classifiers to two tasks: 1) testing whether we can predict if a DNA sample came from coral mucus, tissue, or skeleton using microbial 16S rRNA data alone and 2) predicting whether within each coral compartment we can predict certain categorical features of a sampled coral using its microbiome (contact with turf algae, reef name, complex vs. robust clade membership, etc.). The results from random forest analysis of

coral compartments are presented in the main text. The results for random forest analysis of host and environmental parameters are presented in Supplementary Data S6. Because error rates typically scale with both the number of categories - it is easier to predict the correct category for a binary category than one with 100 possibilities for example – we took care to consider the proportional increase in random classification relative to a baseline formed by random guessing. For both the compartment classification task and the trait classification task, we tested random forest classification on microbial phyla, orders or genera. For the compartment classification task we also tested random forest classification with 97% OTUs directly or predicted functional repertoires of coral microbiomes as inferred using the PICRUSt software. However, this was computationally expensive and yielded < 0.1% improvement in classification error rates over classification based on microbial genera, and was therefore not pursued further.

Statistical analyses on the effect of phylogeny on the microbiome

Phylogenetic analyses were conducted in *R* v3.3.1[54]. The packages *ape* (v3.5)[55] and *paleotree* (v2.7)[56] were used to manipulate trees and to calculate cophenetic distances. Univariate phylogenetic correlograms of α-diversity and distance-to-centroid measures were implemented using the package *phylosignal* (v1.1)[57]. Mantel tests and mantel correlograms of multivariate dissimilarities vs. phylogenetic distances were implemented using *vegan* (v2.5-2*)*[58]. Size classes for the mantel correlograms were defined manually. Following Sturge's rule, we created 11 distance classes. Due to the structure of the host phylogeny and the sampled species, four discrete phylogenetic distances greater than ~0.3 existed, corresponding to (1) all comparisons between the two major coral clades, (2) all comparisons between scleractinians and *Palythoa* (Zoantharia), (3) all comparisons between hexacorals and octocorals, and (4) all comparisons between anthozoans and *Millepora* (Hydrozoa). We first created distance classes that corresponded to each of these four discrete comparisons, then created the remaining seven distance classes by spacing them evenly across the smaller phylogenetic distances.

Phylogenetic Generalized Linear Mixed Models (pGLMMs) for analysis of the entire community were implemented using the package *MCMC.OTU* (1.0.10)[59]. *MCMC.OTU* wraps the package *MCMCglmm* (v2.24)[60] to fit a model whereby sequencing depth is

accounted for by using a sample's total read count as the base level of a fixed per-OTU effect. Compositionality is accounted for by the inclusion of a per-sample random effect, and a number of other parameters and priors are set with defaults that are sensible for microbiome studies, such as 'global effects' of each specified factor (which control for and test effects on α-diversity) and independent error variance for each OTU. Analysis of the entire set of 97% OTUs was computationally impractical, so OTUs were first collapsed from the pre-rarefaction OTU table into their annotated genera using QIIME's summarize_taxonomy.py. The package *phyloseq* (1.18.1)[61] was then used to import and manipulate this table and its associated metadata. Samples with total counts less than 1000 or that were lacking relevant metadata were removed. The purgeOutliers command was applied to the data with an otu.cut value of 0.0001. For the first, more comprehensive GLMM, the command mcmc.otu was run with maximum corallite width, disease prevalence, and binary turf contact as fixed effects; geographic area, host phylogeny, and host identity as random effects; a chain length of 125000, thinning interval of 5, and burn-in of 25000; and with the inverse of the host phylogenetic covariance matrix supplied with the ginverse option. Subsequently, the command was run again with latitude and then coral colony size as the sole fixed effect, and with only host phylogeny as a random effect. Significance for each term was determined by calculating 95% credible intervals with HPDinterval and isolating those that did not include zero.

Cophylogenetic analyses

We reasoned that microbial groups that are most intimately associated with corals (whether commensal, mutualistic, or parasitic) are likely to have evolved in ways that led to patterns of cophylogeny with their hosts. A preliminary pipeline was developed to screen the microbiome for such groups. First, joined sequences were re-processed with the Minimum Entropy Decomposition (MED) pipeline[33], discarding MED nodes with substantive abundances less than 100. Taxonomy was assigned to the resulting MED representative sequences as before with OTUs. Family-level groups of microbes were analyzed independently because higher taxonomic levels would be unlikely to have evolved within the same timescale as scleractinians, and lower taxonomic levels were more likely to contain misannotations. It was computationally impractical to analyze all

microbial families, so only the most prevalent in each compartment were tested. And arbitrary threshold of 50% prevalence was chosen. For each family in each compartment, all MED nodes were isolated that were the most abundant representative in at least one sample. This conservative approach was done partly out of concern that spurious sequences generated by sequencing error could influence the downstream phylogenetic analyses, and partly to reduce each dataset to a size that was practical for phylogenetic inference and GLMMs. The representative sequences were then combined from these nodes with reference sequences for each family. Reference sequences were randomly subsampled from the Greengenes 13_8 99% OTU database such that each dataset contained the MED nodes of interest, 75 random full-length 16S sequences belonging to the family of interest, plus 10 random 'outgroup' sequences belonging to any other family from the same order.

Each collection of sequences was then aligned using MAFFT in QIIME. Phylogenetic trees were built using BEAST 2.4.2 with a chain length of 100 million, thinning interval of 1000, a log-normal relaxed clock model, and the site model selected using bModelTest. The maximum clade credibility tree for each group was selected using TreeAnnotator with a burn-in of 25% and common ancestor heights.

A separate pGLMM was then fit for each microbial family in each tissue compartment. The raw MED table was imported into *R* using *phyloseq* and filtered to contain only samples with counts greater than 1000. The resulting table was merged with each microbial family's phylogenetic tree using *phyloseq*, a process that automatically filters all sequences from the table that are not represented on the tree. Samples were further filtered from this table if they did not retain a count of least 10. Phylogenetic covariance matrices based on the bacterial and host phylogenies were then generated[62]. Phylogenetic covariance matrices based on the bacterial and host phylogenies were generated using the function inverseA on each host tree. The Kronecker product of the resulting matrices was then computed for use as the 'coevolutionary' covariance matrix. The Kronecker product of each phylogenetic covariance matrix and an identity matrix was computed for use as microbial identity x host phylogeny and microbial phylogeny x host identity interaction effects[62]

Binary models were fit with *MCMCglmm* using a single fixed effect of the log of the sequencing depth, 'global' random effects of host phylogeny, host identity, microbial phylogeny, and microbial identity, all combinations of host-by-microbe phylogenetic and identity random interaction effects, and a geographic area-by-microbial identity random interaction effect. Altogether this approach is similar to the models described in reference[62]. Our models were fit with a chain length of 1,250,000, thinning interval of 50, and burn-in of 250,000. After the model was fit, convergence was assessed by verifying that the Effective Sample Sizes (ESS) of all covariance terms were greater than 200. Intraclass correlation coefficients (ICCs) were calculated for each iteration, with 95% credible intervals calculated with HPDinterval. Factors with ICC lower credible bounds greater than 0.01 were considered significant.

To independently analyze subclades of *Endozoicomonas*-like bacteria, a custom QIIME-formatted taxonomy database was created with sequence annotations corresponding to clades C, HS, HS-R, and HS-C from the initial analysis. Taxonomy was then assigned to all MED nodes and *Endozoicimonaceae* Greengenes reference sequences using UCLUST with max-accepts set to 1. The above procedure of filtering, selecting reference sequences, building a phylogeny, and fitting pGLMMs, was then repeated based on each annotated subclade instead of each family. HS-R within only Robust clade corals was also analyzed by first filtering other samples from the dataset.

Figures and Figure Legends

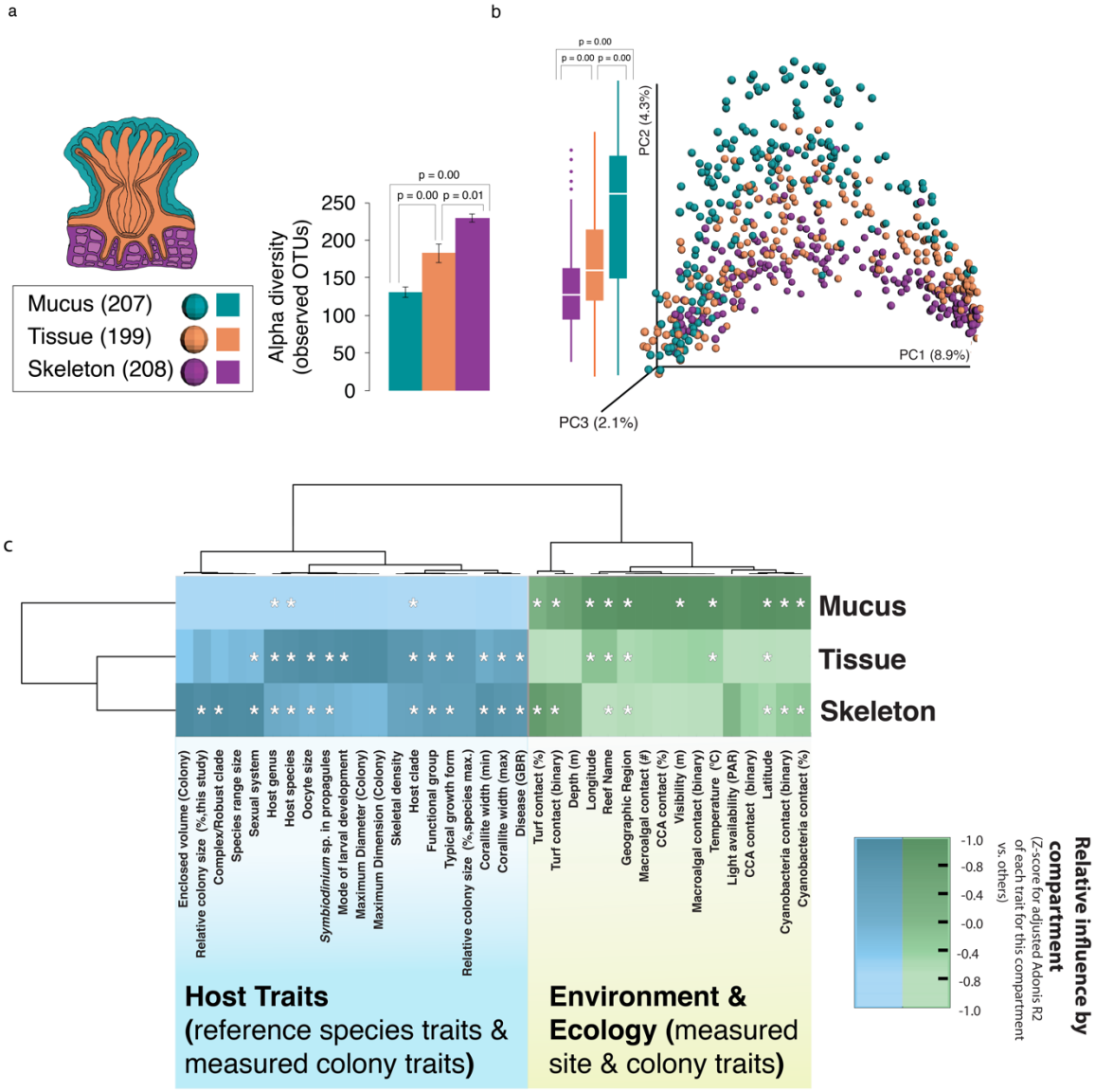


Figure 1.1. Anatomical differences in coral microbiomes.

Coral mucus, tissue, and skeleton microbiomes differ in richness, composition, and response to host vs. environmental factors based on 16S rRNA gene sequence data. a) Microbial community richness (observed OTUs) in coral mucus (teal), tissue (orange) and skeleton (purple), assessed at an even depth of 1000 reads per sample. P-values reflect Tukey's HSD. b) Principal coordinates plot of coral-associated microbial communities (Unweighted UniFrac; n = 614). Reads were rarefied to 1000 reads per

sample. Coral compartments show significant differences in community composition (Adonis R2 = 0.028; permutational $p < 0.001$). The percent variation explained by the principal coordinates is indicated at the axes. Boxplots of the second PC elucidate differences among compartments. P-values reflect Tukey's HSD. c) Relative influence of host and environmental factors on microbiome composition (Weighted UniFrac, Adonis adjusted R2) in each compartment. Darker cells for a compartment indicate that it is more strongly influenced by that trait than the other compartments (Adonis adjusted R2 values z-score normalized within columns). Cell values reflect adjusted R2, which penalizes R2 for each factor downward to allow for fair comparison among factors with varying degrees of freedom. Asterisks indicate a significant effect of that factor (Adonis permutational $p < 0.05$) on the microbiome in that compartment, following stringent Bonferroni correction across all traits and compartments. While both host and environmental factors influenced all compartments, host factors tended to influence coral tissue and skeleton more strongly than mucus, whereas host environment more influenced mucus microbiomes. All values in the table, plus other combinations of rarefaction depth and multivariate dissimilarity measure are presented in Supplementary Data 4.

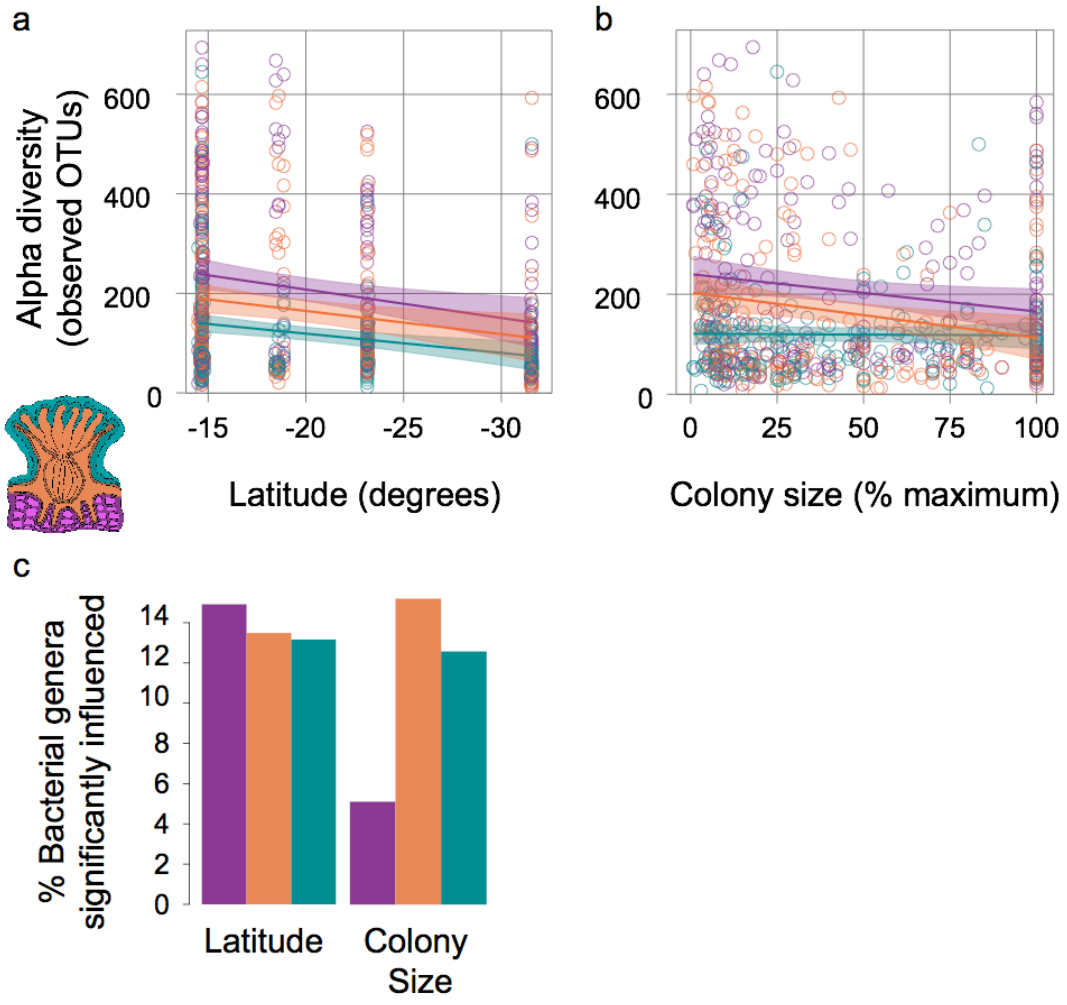


Figure 1.2. Effects of latitude and coral relative colony size on coral microbiomes.

In all panels, we rely on phylogenetic Generalized Linear Mixed Models (pGLMMs; Methods), which account for potential confounding effects of coral phylogeny, for effect size and significance. a) Microbial community richness (observed OTUs) as a function of latitude and coral anatomy (teal, coral mucus; orange, tissue; purple, skeleton). For visualization of latitudinal effects on richness, linear correlations are shown with colored lines, and their 95% confidence intervals are shown by shaded areas. Associations between latitude and microbiome richness were significant in coral mucus and tissue, but not skeleton (pMCMC: mucus, 0.0018; tissue, 0.0004; skeleton 0.468; pGLMM effect sizes: mucus, 0.026; tissue, 0.035; skeleton, 0.007). b) Microbiome richness as a function

of coral colony size relative to the maximum recorded size for each species and coral anatomy. Relative colony size vs. microbiome richness was visualized with linear regression. A negative association between coral relative size and microbiome richness was significant in tissue and skeleton, but a positive association in mucus was not significant (pMCMC: mucus, 0.86; tissue, 0.0008; skeleton, 0.02; pGLMM effect sizes: mucus, 0.028; tissue, -0.591; skeleton, -0.392). c) Percent of tested microbial genera significantly associated with latitude and colony size in phylogenetically-controlled pGLMMs (Supplementary Data 6).

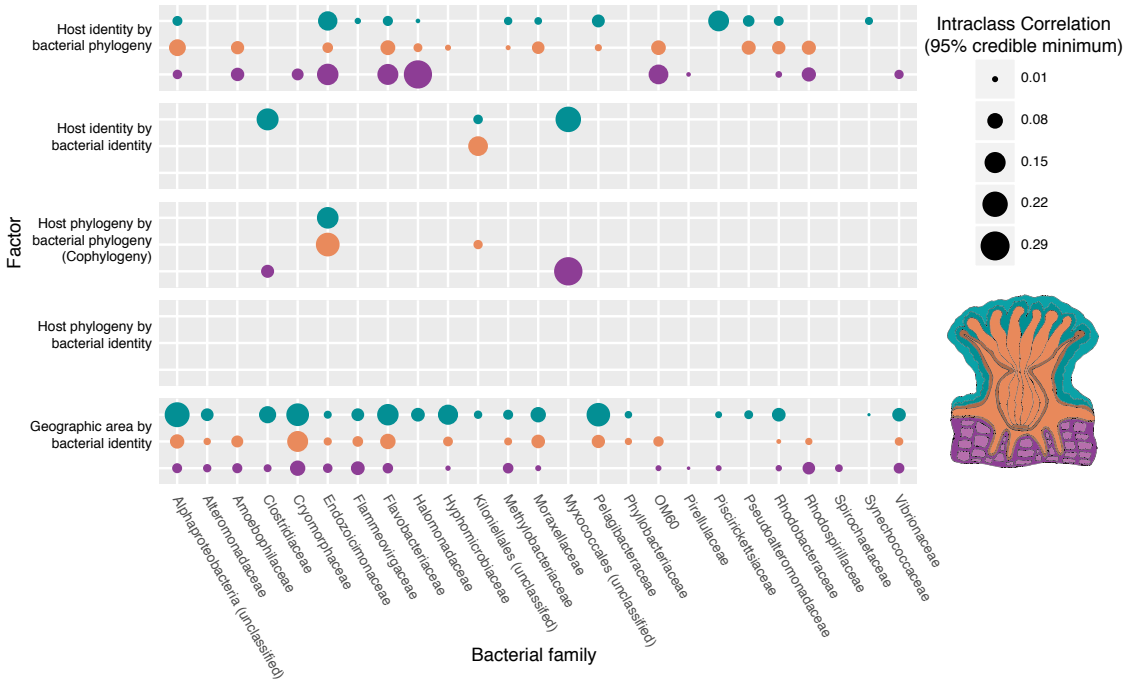


Figure 1.3. Effects of host identity, phylogeny, and cophylogeny on bacterial families. Results are derived from co-phylogenetic GLMM analysis within prevalent bacterial families, and incorporate geographic area, bacterial and coral host identity, and bacterial and coral host phylogeny (see Methods and workflow in Fig. 1). Each block of rows corresponds to a factor in the model (main effects are not shown). Each row within a block corresponds to a tissue compartment (teal = mucus, orange = tissue, and purple = skeleton; see coral polyp illustration), while each column corresponds to an independent model fit for the specified microbial group. Dots were plotted only for 'significant' factors (ICC lower bound > 0.01). The size of each dot represents the intra-class correlation coefficient (ICC) 95% credible lower bounds from co-phylogenetic linear model analysis. While coral host identity is associated with bacterial phylogeny for most prevalent bacterial families (top block of rows), only 4 bacterial families show co-phylogeny with corals (middle block of rows).

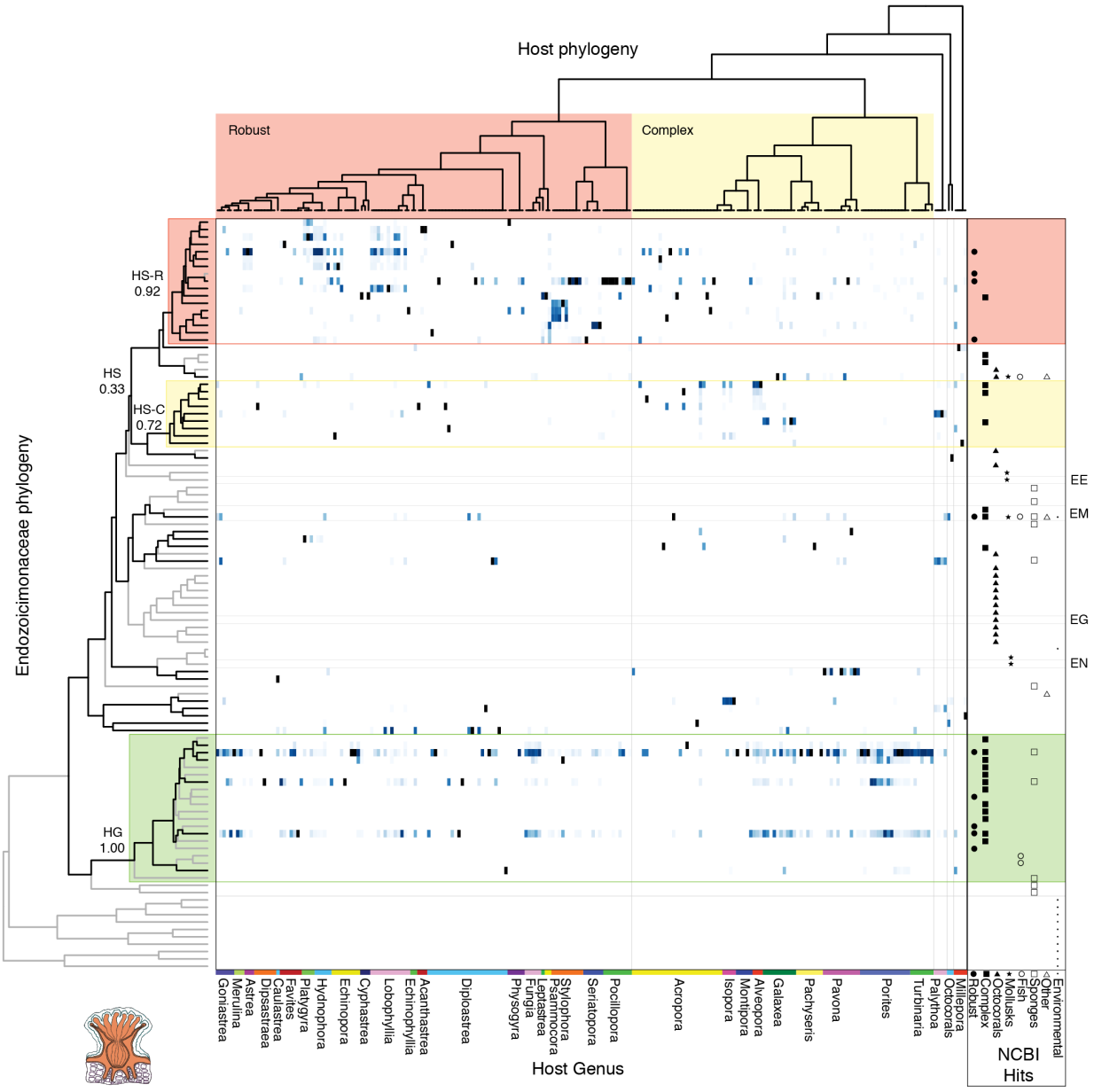


Figure 1.4. Distribution of Endozoicomonas-like bacteria across coral hosts.

The heatmap illustrates patterns of association between Endozoicomonas-like bacteria and coral hosts. Colored cells represent the relative abundance of Endozoicomonas-like bacterial sequences (out of the total abundance of Endozoicomonas-like bacteria) in each coral host, plotted on a scale from 0% (white) to 100% (dark blue). The x-axis is arranged by coral host phylogeny, which is shown at the top. The y-axis is arranged by the

phylogeny of the most abundant Endozoicomonas-like bacterial sequences observed in host tissues, which is shown to the left (Bayesian posterior support values are shown for clades of interest). Clade HG (Host Generalist; green box) is prevalent in diverse species spanning both the Complex and Robust clades. Clades HS-R (Host-specific: Robust; pink box) and HS-C (Host-specific: Complex; yellow box) are composed of host-specific Endozoicomonas-like lineages. On the right, the host organisms of each sequence variant's perfect matches in NCBI's nr database are shown. Cultured and named strains are identified with abbreviations (EE: Endozoicomonas elysicola, EM: Endozoicomonas montiporae, EG: Endozoicomonas gorgoniicola, EN: Endonucleobacter bathymodioli). Sequences in clades HS-R and HS-C are consistently associated with Robust and Complex clade corals, respectively (see Supplementary Figure 1.5 for more detail).

References

1. Horton, T. *et al.* World Register of Marine Species. (2018) doi:10.14284/170
2. Rohwer, F., Seguritan, V., Azam, F. & Knowlton, N. Diversity and distribution of coral-associated bacteria. *Mar. Ecol. Prog. Ser.* **243,** 1–10 (2002).
3. Zaneveld, J. R. *et al.* Overfishing and nutrient pollution interact with temperature to disrupt coral reefs down to microbial scales. *Nat. Commun.* **7,** 11833 (2016).
4. Sunagawa, S. *et al.* Bacterial diversity and White Plague Disease-associated community changes in the Caribbean coral *Montastraea faveolata*. *ISME J.* **3,** 512–521 (2009).
5. Sato, Y., Willis, B. L. & Bourne, D. G. Successional changes in bacterial communities during the development of black band disease on the reef coral, *Montipora hispida*. *ISME J.* **4,** 203–214 (2010).
6. Vega Thurber, R. *et al.* Metagenomic analysis of stressed coral holobionts. *Environ. Microbiol.* **11,** 2148–2163 (2009).
7. Brooks, A. W., Kohl, K. D., Brucker, R. M., Opstal, E. J. van & Bordenstein, S. R. Phylosymbiosis: relationships and functional effects of microbial communities across host evolutionary history. *PLOS Biol.* **14,** e2000225 (2016).
8. Yeoh, Y. K. *et al.* Evolutionary conservation of a core root microbiome across plant phyla along a tropical soil chronosequence. *Nat. Commun.* **8,** 215 (2017).
9. Thomas, T. *et al.* Diversity, structure and convergent evolution of the global sponge microbiome. *Nat. Commun.* **7,** 11870 (2016).
10. Sanders, J. G. *et al.* Stability and phylogenetic correlation in gut microbiota: lessons from ants and apes. *Mol. Ecol.* **23,** 1268–1283 (2014).
11. Ley, R. E. *et al.* Evolution of mammals and their gut microbes. *Science* **320,** 1647–1651 (2008).
12. Douglas, A. E. & Werren, J. H. Holes in the Hologenome: why host-microbe symbioses are not holobionts. *mBio* **7,** e02099 (2016).
13. Costello, E. K. *et al.* Bacterial community variation in human body habitats across

space and time. *Science* **326,** 1694–1697 (2009).
14. Apprill, A., Weber, L. G. & Santoro, A. E. Distinguishing between microbial habitats unravels ecological complexity in coral microbiomes. *mSystems* **1,** e00143-16 (2016).
15. Ainsworth, T. D. *et al.* The coral core microbiome identifies rare bacterial taxa as ubiquitous endosymbionts. *ISME J.* **9,** 2261–2274 (2015).
16. Thompson, J. N. *The Coevolutionary Process*. (University of Chicago Press, 1994).
17. Rivière, M. L., Garrabou, J. & Bally, M. Evidence for host specificity among dominant bacterial symbionts in temperate gorgonian corals. *Coral Reefs* **34,** 1087–1098 (2015).
18. van de Water, J. A. J. M. *et al.* Comparative assessment of Mediterranean gorgonian-associated microbial communities reveals conserved core and locally variant bacteria. *Microb. Ecol.* **73,** 466–478 (2017).
19. Sunagawa, S., Woodley, C. M. & Medina, M. Threatened corals provide underexplored microbial habitats. *PLOS ONE* **5,** e9554 (2010).
20. Huang, D. & Roy, K. The future of evolutionary diversity in reef corals. *Philos. Trans. R. Soc. B Biol. Sci.* **370,** (2015).
21. Madin, J. S. *et al.* The Coral Trait Database, a curated database of trait information for coral species from the global oceans. *Sci. Data* **3,** 160017 (2016).
22. Willis, B. L., Page, C. A. & Dinsdale, E. A. Coral disease on the Great Barrier Reef. in *Coral Health and Disease* 69–104 (Springer, Berlin, Heidelberg, 2004). doi:10.1007/978-3-662-06414-6_3
23. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6,** 1621–1624 (2012).
24. Eloe-Fadrosh, E. A., Ivanova, N. N., Woyke, T. & Kyrpides, N. C. Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat. Microbiol.* **1,** 15032 (2016).
25. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523,** 208 (2015).
26. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6,** 610–618 (2012).
27. Hillebrand, H. On the generality of the latitudinal diversity gradient. *Am. Nat.* **163,** 192–211 (2004).
28. Denner, E. B. M. *et al. Aurantimonas coralicida* gen. nov., sp. nov., the causative agent of white plague type II on Caribbean scleractinian corals. *Int. J. Syst. Evol. Microbiol.* **53,** 1115–1122 (2003).
29. Ziegler, M. *et al.* Coral microbial community dynamics in response to anthropogenic impacts near a major city in the central Red Sea. *Mar. Pollut. Bull.* **105,** 629–640 (2016).
30. Williams, A. D. *et al.* Age-related shifts in bacterial diversity in a reef coral. *PLoS ONE.* **10,** e0144902 (2015).
31. Brucker, R. M. & Bordenstein, S. R. The capacious hologenome. *Zoology* **116,** 260–261 (2013).
32. Meikle, P., Richards, G. N. & Yellowlees, D. Structural investigations on the mucus from six species of coral. *Mar. Biol.* **99,** 187–193 (1988).
33. Eren, A. M. *et al.* Minimum entropy decomposition: Unsupervised oligotyping for

sensitive partitioning of high-throughput marker gene sequences. *ISME J.* **9,** 968 (2015).
34. Neave, M. J. *et al.* Differential specificity between closely related corals and abundant *Endozoicomonas* endosymbionts across global scales. *ISME J.* **11,** 186–200 (2017).
35. Werner, G. D. A., Cornwell, W. K., Sprent, J. I., Kattge, J. & Kiers, E. T. A single evolutionary innovation drives the deep evolution of symbiotic N2-fixation in angiosperms. *Nat. Commun.* **5,** 4087 (2014).
36. Bayer, T. *et al.* The microbiome of the Red Sea coral *Stylophora pistillata* is dominated by tissue-associated *Endozoicomonas* Bacteria. *Appl. Environ. Microbiol.* **79,** 4759–4762 (2013).
37. Soffer, N. *et al*. Potential role of viruses in white plague coral disease. *ISME J.* **8,** 271-283 (2014).
38. Kellogg, C. A. *et al*. Comparing bacterial community composition between healthy and white plague-like disease states in *Orbicella annularis* using PhyloChip™ G3 microarrays. *PLoS ONE* **8,** e79801 (2013).
39. Closek, C. J. *et al.* Coral transcriptome and bacterial community profiles reveal distinct Yellow Band Disease states in *Orbicella faveolata*. *ISME J* **8,** 2411-2422 (2014).
40. Gaulke, C. A. *et al.* Ecophylogenetics reveals the evolutionary associations between mammals and their gut microbiota. *bioRxiv* 182212 (2017).
41. Fukami, H. *et al.* Mitochondrial and nuclear genes suggest that stony corals are monophyletic but most families of stony corals are not (Order Scleractinia, Class Anthozoa, Phylum Cnidaria). *PLOS ONE* **3,** e3222 (2008).
42. Yilmaz, P. *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* **29,** 415–420 (2011).
43. Siebeck, U. E., Marshall N. J., Klüter, A., Hoegh-Guldberg, O. Monitoring coral bleaching using a color reference card. *Coral Reefs* **25,** 453–460 (2006).
44. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7,** 335 (2010).
45. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinforma. Oxf. Engl.* **26,** 2460–2461 (2010).
46. Rideout, J. R. *et al.* Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ* **2,** e545 (2014).
47. Caporaso, J. G. *et al.* PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinforma. Oxf. Engl.* **26,** 266–267 (2010).
48. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5,** e9490 (2010).
49. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
50. Bouckaert, R. *et al.* BEAST 2: A software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **10,** e1003537 (2014).
51. Bouckaert, R. R. & Drummond, A. J. bModelTest: Bayesian phylogenetic site model

averaging and model comparison. *BMC Evol. Biol.* **17,** 42 (2017).
52. Darling, E. S. *et al.* Evaluating life-history strategies of reef corals from species traits. *Ecol. Lett.* **15,** 1378-1386 (2012).
53. Budd, A. F., Fukami, H., Smith, N. D. & Knowlton, N. Taxonomic classification of the reef coral family Mussidae (Cnidaria: Anthozoa: Scleractinia). *Zool. J. Linn. Soc.* **166,** 465–529 (2012).
54. R Development Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2008).
55. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinforma. Oxf. Engl.* **20,** 289–290 (2004).
56. Bapst, D. W. paleotree: an R package for paleontological and phylogenetic analyses of evolution. *Methods Ecol. Evol.* **3,** 803–807 (2012).
57. Keck, F., Rimet, F., Bouchez, A. & Franc, A. phylosignal: an R package to measure, test, and explore the phylogenetic signal. *Ecol. Evol.* **6,** 2774–2780 (2016).
58. Oksanen, J. et al. vegan: Community Ecology Package. (2008).
59. Green, E. A. *et al.* Quantifying cryptic *Symbiodinium* diversity within *Orbicella faveolata* and *Orbicella franksi* at the Flower Garden Banks, Gulf of Mexico. *PeerJ* **2,** e386 (2014).
60. Hadfield, J. D. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J Stat Soft* **33,** 1-22 (2010). doi:10.18637/jss.v033.i02
61. McMurdie, P. J. & Holmes, S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLOS ONE* **8,** e61217 (2013).
62. Hadfield, J. D., Krasnov, B. R., Poulin, R. & Nakagawa, S. A tale of two phylogenies: comparative analyses of ecological interactions. *Am. Nat.* **183,** 174–187 (2014).

Connections among the microbiome, coral disease susceptibility, and coral life history strategies

Authors: Ryan McMinds[1#*], F. Joseph Pollock[2#], Styles Smith[2], David G. Bourne[3,4], Bette L. Willis[3,5], Mónica Medina[2,6], Rebecca Vega Thurber[1*], Jesse R. Zaneveld[7]

[1]Oregon State University, Department of Microbiology, 226 Nash Hall, Corvallis, OR, 97331, USA

[2]Pennsylvania State University, Department of Biology, 208 Mueller Lab, University Park, PA 16802, USA

[3]James Cook University, College of Science and Engineering, Townsville, Queensland 4811, Australia

[4]Australian Institute of Marine Science, Townsville, Queensland 4810, Australia

[5]ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Queensland 4811, Australia

[6]Smithsonian Tropical Research Institute, Smithsonian Institution, 9100 Panama City PL, Washington, DC 20521, USA

[7]University of Washington, Bothell, School of Science, Technology, Engineering, and Mathematics, Division of Biological Sciences, UWBB-277, Bothell, WA 98011, USA

[#]Contributed equally

[*]Corresponding author

Abstract

Global change is predicted to increasingly drive future outbreaks of diseases of animals. Scleractinian corals, the foundation species of tropical reefs, have already significantly declined due to massive disease epizootics that are linked to the individual and combined effects of overfishing, climate change induced sea-surface temperature anomalies, and pollution. Identifying the factors that lead to increased disease susceptibility has become a major target for conservation efforts. Yet identifying the proper alignment of host-pathogen-environment factors leading to disease outbreaks has remained challenging due to a paucity of studies that can disentangle variation among the host phylogeny, the animal microbiome, and environmental parameters. We surveyed bacterial and archaeal communities in healthy Australian corals and compared them to their host species' phylogeny, overall life-history strategy, and average disease susceptibility. We found that although the richness of the microbiome was highly correlated with the average disease susceptibility of the host genera, controlling for phylogenetic relatedness made this effect non-significant. However, the abundances of many individual microbial genera were significantly correlated disease susceptibility even when controlling for the hosts' relatedness. Life history traits were also linked to disease susceptibility and microbiome compositionality and richness. Together, these findings connect microbial symbiosis to the evolution of disease susceptibility and life-history strategies in a basal animal lineage.

Introduction

Coral microbiome composition is thought to influence resistance to stress and disease (Lesser *et al.* 2007, Daszak *et al.* 2000), yet analyses of the associations among microbiome characteristics and disease resistance require data that spans diverse coral taxa and geographic range. We previously found that the average disease susceptibility of coral genera was one of several variables that explained a significant amount of variation in the composition of the coral microbiome (Pollock, McMinds *et al.* 2018). To test if there are systematic differences in the microbiomes of disease-resistant vs. disease-susceptible coral taxa, microbial sequence data were paired with 10-year, genus-level surveys of coral disease on mid-shelf reefs on the northern Great Barrier Reef (Willis *et al*. 2004, Pollock, McMinds *et al.* 2018). Given that corals and their pathogens exhibit

complex coevolutionary dynamics, we used total disease prevalence as a proxy for corals' investment in disease resistance. Because the vast majority of coral colonies (99.6%) sampled in this past survey were visibly healthy, our analysis assessed if and how healthy coral microbiomes vary between disease-susceptible and disease-resistant coral genera.

Host traits such as investment in disease resistance do not evolve independently, and previous studies have described two to four relatively discrete coral life-history strategies that are composed of numerous highly correlated species traits. Although the original analyses that described these groups did not consider the species' disease susceptibility, it is reasonable to assume that it is a trait that varies in tandem with those strategies. Therefore we also assessed the relationships between these life history strategies and microbiome composition and richness.

We collected data from three anatomical coral compartments, mucus, tissue, and skeleton, and analyzed each separately. We compared naïve correlations of these data against microbiome traits to phylogenetically controlled linear model analyses. We show that strong phylogenetic patterning of traits precludes robust conclusions about the relationship between disease susceptibility and microbiome richness, but that more robust conclusions could be reached regarding the abundances of individual microbial genera. We also show that there are significant relationships between microbiome richness and certain life-history strategies.

Results

Convergent microbiome richness and disease susceptibility in two groups of corals

In all compartments, microbiome richness was significantly negatively correlated with the disease susceptibility of coral species when tested with phylogenetically naïve Spearman correlations (Fig. 2.1b; tissue, $R^2 = 0.123$, $p < 0.001$; skeleton: $R^2 = 0.137$, $p < 0.001$; mucus: $R^2 = 0.029$, $p = 0.016$). However, these correlations were not significant after accounting for coral phylogeny with phylogenetic GLMMs (Fig. 2.1b; tissue, pMCMC = 0.8; skeleton: pMCMC = 0.4; mucus: pMCMC = 0.12). This observation suggests that the apparent connection between disease susceptibility and reduced

microbiome richness might be driven by strong trends in a few, closely related coral groups. We visualized the correspondences among disease susceptibility, microbiome richness, and phylogenetic relatedness by mapping each trait to the coral phylogeny as a continuous trait, then performing a maximum likelihood estimate of ancestral states (Fig. 2.1a). This reveals the convergent increases in disease susceptibility and reductions in microbiome richness in two important coral groups: *Acropora* in the complex clade and the Pocilloporidae within the robust clade.

Disease susceptibility correlates with microbiome composition

Using phylogenetically naïve Adonis tests for compositional correlations, the long-term susceptibility of coral genera to disease on the Great Barrier Reef was strongly associated with microbiome structure in all compartments (all $p < 0.05$), accounting for 18-29% of the variation in overall microbial community composition. Despite the proposed role of coral mucus as a barrier against disease, we found that the correlation between disease susceptibility and microbiome structure was more pronounced in tissue ($R^2 = 0.289$) and skeleton microbiomes ($R^2 = 0.250$) than in mucus ($R^2 = 0.183$). Although these high-level multivariate summaries did not account for phylogenetic autocorrelation, univariate phylogenetic GLMMs generally supported these findings, showing that ~11, 18, and 16% of microbial genera in the tissue, skeleton, and mucus, respectively, are significantly correlated with disease susceptibility (Figure 2.1c).

Coral life-history strategy correlates with microbiome structure

The groups that converged on high disease susceptibility and low microbiome richness (the acroporids and Pocilloporidae) belong to distinct clusters of corals with similar life-history strategies ('competitive' and 'weedy' in the classification of Darling *et al*. 2012). We therefore explored the idea that coral life-history strategy might drive convergence in microbiome richness and disease susceptibility.

We characterized the influence of coral life-history strategy on microbiome composition and richness. Coral life-history strategy was significantly associated with 10-11% of variance in microbiome composition in the tissue and skeleton, but only 3% of variance in mucus (Fig. 2.2a-e; all Adonis $p << 0.05$). Using phylogenetic GLMMs, we contrasted each of the four Darling life history strategies serially (Figure 2.3), (1) fast vs. slow, or

‘competitive’ corals vs. all others; (2) ‘weedy’ vs. ‘generalist’ and ‘stress-tolerant’, and (3) ‘generalist’ vs. ‘stress-tolerant’. There was no difference in richness between fast- and slow-growing corals in any compartment (pMCMC, tissue: 0.28; skeleton: 0.13; mucus: 0.83). Weedy corals were less rich than generalists or stress-tolerant corals in the tissue compartment, but not skeleton or mucus (pMCMC, tissue: 0.046; skeleton: 0.10; mucus: 0.52). Generalists were less rich than stress-tolerant corals in the tissue and skeleton, but not mucus (pMCMC, tissue: 0.006; skeleton: 0.02; mucus: 0.5).

Discussion

In this analysis, we show that the combination of microbiome richness and disease susceptibility alone almost perfectly separated coral genera into life-history strategies defined by Darling *et al.* (Fig. 2.2f), even though neither factor was originally used in defining the groups (Darling *et al.* 2012). We also show that two independent lineages (*Acropora* and the Pocilloporidae) that have distinct life-history strategies are united by convergent reduced microbial symbiont richness. We thus explored the relationships among those groups, the phylogenetic history of corals, microbiome richness, and disease susceptibility. Although a trend is suggestive of a link between richness and disease susceptibility, low phylogenetic replication precludes confident conclusions about this relationship. Further global sampling of related species with different disease susceptibilities and life-history strategies will be needed to help disentangle these traits further, but these results raise the question of whether other animal taxa with well-defined life-history tradeoffs also show broad patterns of increased disease susceptibility and reduced microbiome diversity.

Although convergence in microbiome richness remains to be clarified, corals with different life-history strategies and disease susceptibilities were shown to host distinct microbial communities, even when controlling for phylogenetic relatedness. Ultimately, determining whether there are causal links driving these patterns will require laboratory experiments involving the transfer of specific microbes or microbial communities among coral species. The observation that life-history strategy is correlated with microbiome structure and disease susceptibility highlights the intertwined nature of physiology and microbiology (Sunagawa *et al.* 2010).

This framework demonstrates the importance of controlling for shared evolutionary for analysis of the coral microbiomes of scleractinian corals'. Our results test longstanding hypotheses that bear on potential coral-microbe coevolution, and add quantitative detail and taxonomic breadth to several previously understood patterns in coral microbiology. However, lack of clear relationships among the measured traits also suggests the need for further development of analysis methods that deliver results with more precise interpretations.

Data Deposition

Raw sequence data, metadata, OTU and MED representative sequences, and analysis code are available at https://doi.org/10.6084/m9.figshare.c.3855466.v2.

Acknowledgements

The authors would like to acknowledge Tasman Douglass, Margaux Hine, Frazer McGregor, Kathy Morrow, Katia Nicolet, Cathie Page, and Gergely Torda for their field assistance and Lyndsy Gazda, Jamie Lee Proffitt, Gabriele Swain, and Alaina Weinheimer for their assistance in the laboratory. The authors also acknowledge the staff of the Coral Bay Research Station, Lizard Island Research Station, Lord Howe Island Marine Park, Lord Howe Island Research Station, and RV Cape Ferguson for their logistical support. This work was supported by a National Science Foundation Dimensions of Biodiversity grant (#1442306) to RVT and MM.

Author Contributions

Contributions: RVT, JZ, and MM designed the experiment. FJP, RM, RVT and MM conducted the fieldwork. RM, FJP, and SS generated the libraries. RM, FJP, and JZ performed the data analysis. BW and DB assisted in the sample collection and provided critical data for meta-analysis. RM, FJP, JZ, MM, and RVT wrote the manuscript.

Methods

Selection of target sites

The reefs alongside the continent of Australia represent ~435 million years of coral evolution. Thus to comprehensively target corals across their phylogenetic diversity we sampled representatives of a majority of the 21 major coral clades[48]. We collected

specimen from Ningaloo Reef, Lizard Island, multiple reefs along the northern Great Barrier Reef, and Lorde Howe Island (Extended Data Figure S1). Samples at Lizard Island were collected in the two major seasons to account for annual variation.

Collection of metadata

During sampling, each coral, outgroup species, water, and sediment sample was associated with MIxS metadata (Yilmaz *et al.* 2011; for details see Pollock, McMinds *et al.* 2018). Briefly metadata included basic features of coral species (as identified in the field), location, depth, water temperature, and any visual health cues. Additionally, photographs of each coral were taken and released via openly accessible third-party websites.

Coral sampling

All samples were collected by AAUS-certified scientific divers, in accordance with local regulations. Relevant permit numbers are: CITES (PWS2014-AU-002155, 12US784243/9), Great Barrier Reef Marine Park Authority (G12/35236.1, G14/36788.1), Lord Howe Island Marine Park (LHIMP/R/2015/005), New South Wales Department of Primary Industries (P15/0072-1.0, OUT 15/11450), US Fish and Wildlife Service (2015LA1632527, 2015LA1703560), and Western Australia Department of Parks and Wildlife (SF010348, CE004874, ES002315).

Corals contain three main gross anatomical features which have distinct microbiomes: the surface mucus layer, the tissues (composed of two true tissue layers the gastroderm and epithelia along with an intermediate non-cellular layer, the mesophyll), and the skeleton which is laid down by the coral throughout its lifetime. We standardized the collection of these anatomical compartment. Briefly, mucus was collected by the agitation and negative pressure method while tissue and skeletal samples were collected from each colony by hammer and chisel, or (for branching corals) by bone shears, snap frozen and washed with sterile seawater prior to any downstream collections. Tissue was removed from skeleton using an air gun and ~1 cm^3 region of skeleton not in direct contact with coral tissue was collected with a sterile chisel. Microbiome DNA was collected using the MoBio PowerSoil Kit (MoBio Laboratories, Carlsbad, California) bead tube method.

16S library preparation and sequencing, sequence quality control and initial data processing

Two-stage amplicon PCR was performed on the V4 region of the 16S rRNA gene using the 515F/806R primer pair that targets bacterial and archaeal communities as previously described (Pollock, McMinds, *et al,* 2018). QIIME v1.9 (Caporaso *et al.* 2010a) was used to process all 16S sequence libraries. Primer sequences were trimmed, paired-end reads merged, and QIIME's default quality-control parameters used when splitting libraries. Chimeras were removed and 97%-similarity OTUs picked using USEARCH 7.0 (Edgar 2010), QIIME's subsampled open-reference OTU-picking protocol (Rideout *et al.* 2014), and the 97% GreenGenes 13_8 reference database (McDonald *et al.* 2012). Taxonomy was assigned using UCLUST, and reads were aligned against the GreenGenes database using PyNAST (Caporaso *et al.* 2010b). FastTreeMP (Price *et al.* 2010) was used to create a bacterial phylogeny with constraints defined by the GreenGenes reference phylogeny. Following quality control, 9,441,738 usable reads remained.

A 'canonical' rarefied OTU table was created and used for all downstream analyses except the linear model analyses. To create this table, OTUs were filtered out of the starting table if their representative sequences failed to align with PyNAST to the GreenGenes database or if they were annotated as mitochondrial or chloroplast sequences. The beta_diversity_through_plots.py script was then used to rarefy the resulting table to exactly 1000 sequences per sample, and to calculate from this rarefied table multivariate dissimilarity measures including Bray-Curtis, Binary Jaccard, Weighted UniFrac, and Unweighted UniFrac. Total numbers of observed OTUs per sample were calculated using this table and alpha_rarefaction.py.

The V4 primers we used in this study were designed to amplify the bacterial and archaeal 16S rRNA gene, but they also amplify the 12S rRNA gene of coral mitochondria. We used these additional data for quality control and for phylogenetic analysis of the host corals, as previously described (Pollock, McMinds *et al.* 2018).

Annotation of coral life history strategy and functional traits

Coral life history strategies from Darling *et al*. 2012 ('weedy', 'competitive', 'stress-tolerant', and 'generalist') were added to the mapping file and used for tests of

microbiome structure vs. host traits. Species that did not exist in Darling *et al.* were assigned the same annotation as other member of the same genus if there were multiple species represented and all had the same life history strategy.

Statistical analyses investigating the effect of phylogeny on microbiome traits

Phylogenetic analyses were conducted in *R* v3.3.1 (R Development Core Team 2008). Beta diversity distance-to-centroid values were first calculated using the pairwise distance matrices generated in QIIME and the betadisper function in the package *vegan* v2.4-1 (Oksanen *et al.* 2008), with bias-adjustment.

The packages *ape* v3.5 (Paradis *et al.* 2004) and *paleotree* v2.7 (Bapst 2012) were used to manipulate trees and to calculate cophenetic distances. Univariate phylogenetic correlograms of α-diversity and distance-to-centroid measures were implemented using the package *phylosignal* v1.1 (Keck *et al.* 2016). Phylogenetic Generalized Linear Mixed Models (pGLMMs) assessing the correlation between microbiome richness and disease susceptibility were conducted separately for each compartment using MCMCglmm with disease susceptibility as a fixed effect, host phylogeny as a random effect (with the inverse of the host phylogenetic covariance matrix supplied with the "ginverse" option), a chain length of 1,300,000 iterations, burn-in of 300,000, and thinning interval of 100. Significance was determined with the pMCMC output generated with the function HPDinterval. Assessment of the correlation of richness with life-history strategy was conducted equivalently, except that life-history strategy was treated as a nested categorical effect with contrasts defined as 'fast vs. slow' (the mean of weedy and competitive vs. the mean of generalist and stress-tolerant), 'weedy vs. competitive', and 'generalist vs. stress-tolerant'.

Phylogenetic Generalized Linear Mixed Models (pGLMMs) assessing the abundance of individual microbial genera were implemented using the package *MCMC.OTU* 1.0.10 (Green *et al.* 2014). *MCMC.OTU* wraps the package *MCMCglmm* v2.24 (Hadfield 2010) to fit a multivariate model whereby sequencing depth is accounted for by using a sample's total read count as the base level of a fixed per-OTU effect, compositionality is accounted for by the inclusion of a per-sample random effect, and a number of other

parameters and priors are set with defaults that are sensible for microbiome studies, such as 'global effects' of each specified factor (which control for and test effects on α-diversity) and independent error variance for each OTU. Analysis of the entire set of 97% OTUs was computationally impractical, so OTUs were first collapsed from the pre-rarefaction OTU table into their annotated genera using QIIME's summarize_taxonomy.py. The package *phyloseq* 1.18.1 (McMurdie & Holmes 2013) was then used to import and manipulate this table and its associated metadata. Samples with total counts less than 1000 were removed. The purgeOutliers command was applied to the data with an otu.cut value of 0.0001. The command mcmc.otu was run with disease prevalence as a fixed effect; host phylogeny as a random effect; a chain length of 125000, thinning interval of 5, and burn-in of 25000; and with the inverse of the host phylogenetic covariance matrix supplied with the "ginverse" option. Significance for each term was determined by calculating 95% credible intervals with HPDinterval and isolating those that did not include zero.

Ancestral state reconstruction of coral disease prevalence and α-diversity.

The R package *phylocom* was used to reconstruct ancestral coral disease susceptibility and α-diversity. First, trees and trait tables were filtered to include the same lineages, which excluded species on the coral tree not sampled in this analysis. Next, the fastAnc function was used to generate a maximum likelihood estimate of ancestral trait values for each trait under a Brownian motion model. These reconstructions were then mapped to the tree using the contmap function. Because some non-scleractinian outgroup taxa had microbiome richness data but not disease data, these are plotted only on the microbiome richness tree.

Figures and Figure Legends

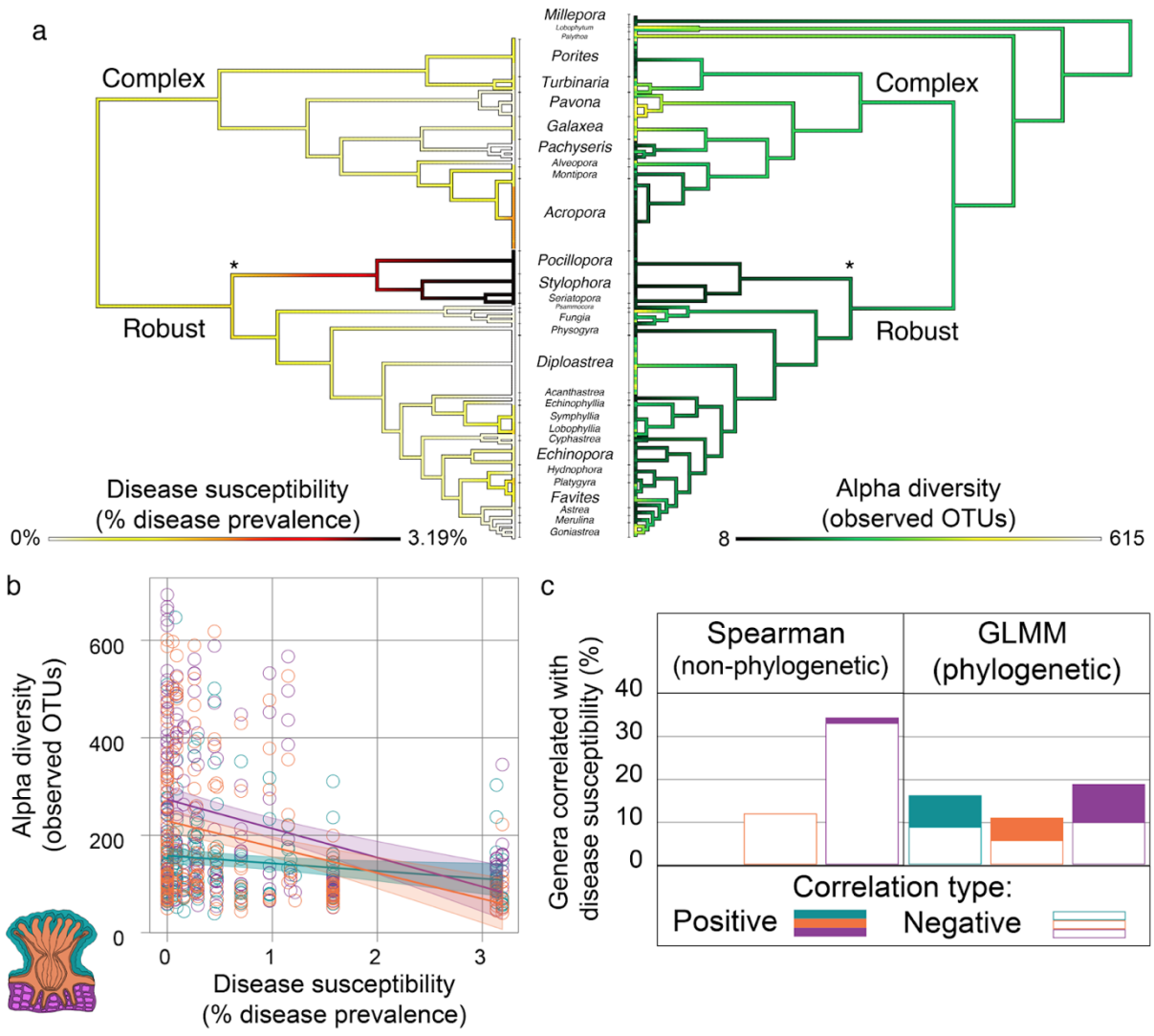


Figure 2.1. Coral microbiome richness correlates with disease susceptibility.

a) Ancestral state reconstruction of genus-wide disease susceptibilities of northern Great Barrier Reef corals (left) and coral microbiome richness (observed OTUs; right). Asterisks indicate the node uniting the family Pocilloporidae (i.e., *Pocillopora*, *Stylophora*, and *Seriatopora*). b) Coral microbiome richness decreases with disease susceptibility. c) Percent of microbial genera associated with disease susceptibility using either phylogenetically-naive Spearman correlations or a phylogenetically-aware

generalized linear mixed model. Filled bars indicate significant positive correlations and empty bars indicate significant negative correlations in mucus (teal), tissue (orange) and skeleton (purple) compartments.

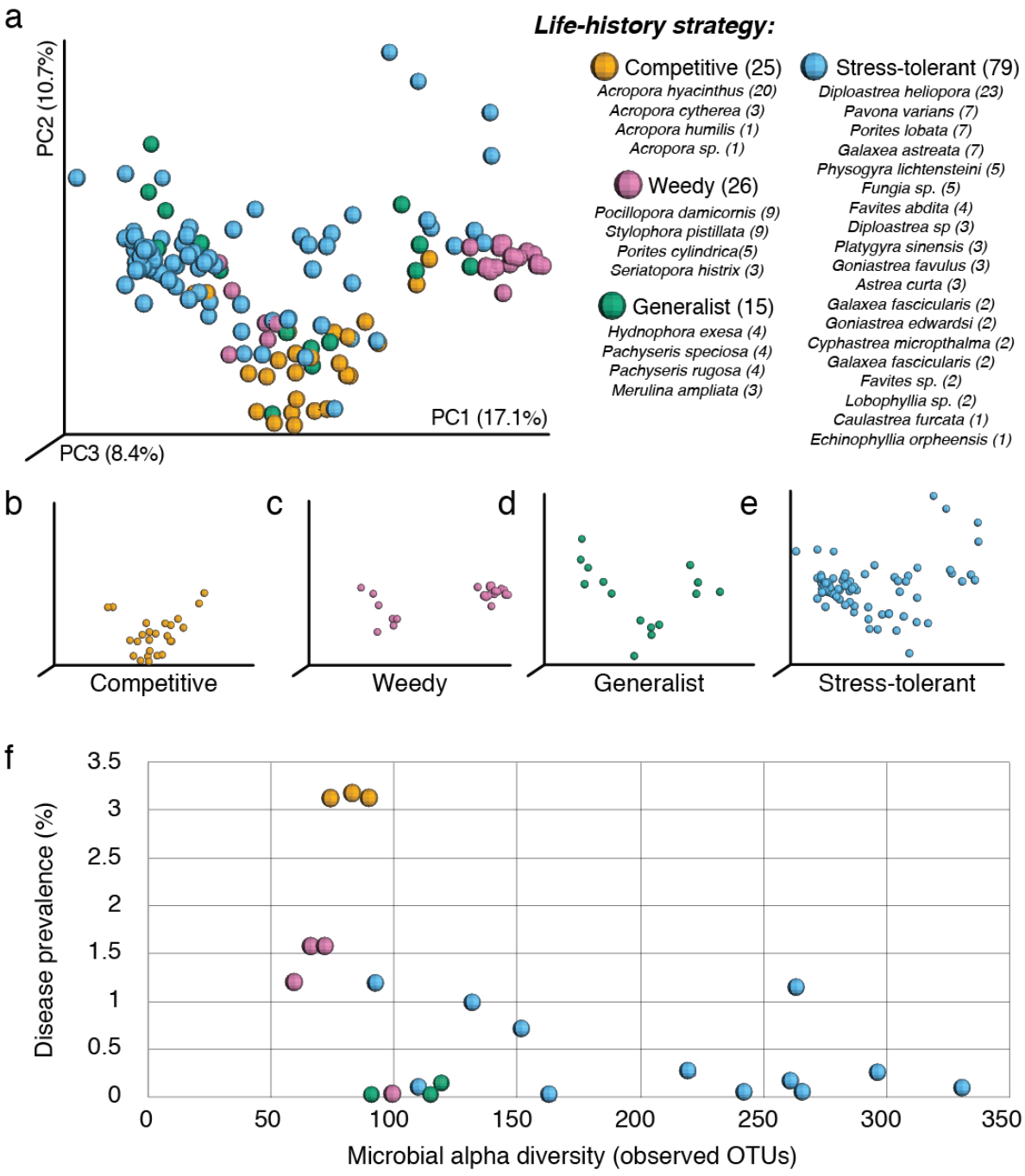


Figure 2.2. Coral life-history strategies influence microbiome composition and richness.

a) PCoA plot of Weighted UniFrac distances between coral tissue microbiomes. Colors reflect coral life-history strategy (Darling *et al.* 2012). The legend indicates the number of tissue samples from corals in each functional group, and the specific coral species to which they belong. b-e) Panels show each functional group separately to allow easier visualization of overlapping samples. Functional group subsets are taken directly from, and share the same axes as, panel a. f) Scatter plot of tissue microbiome richness (observed OTUs) vs. disease prevalence for each species, colored by functional group (data from tissue compartment only). Corals with different life-history strategies show differences in microbial richness.

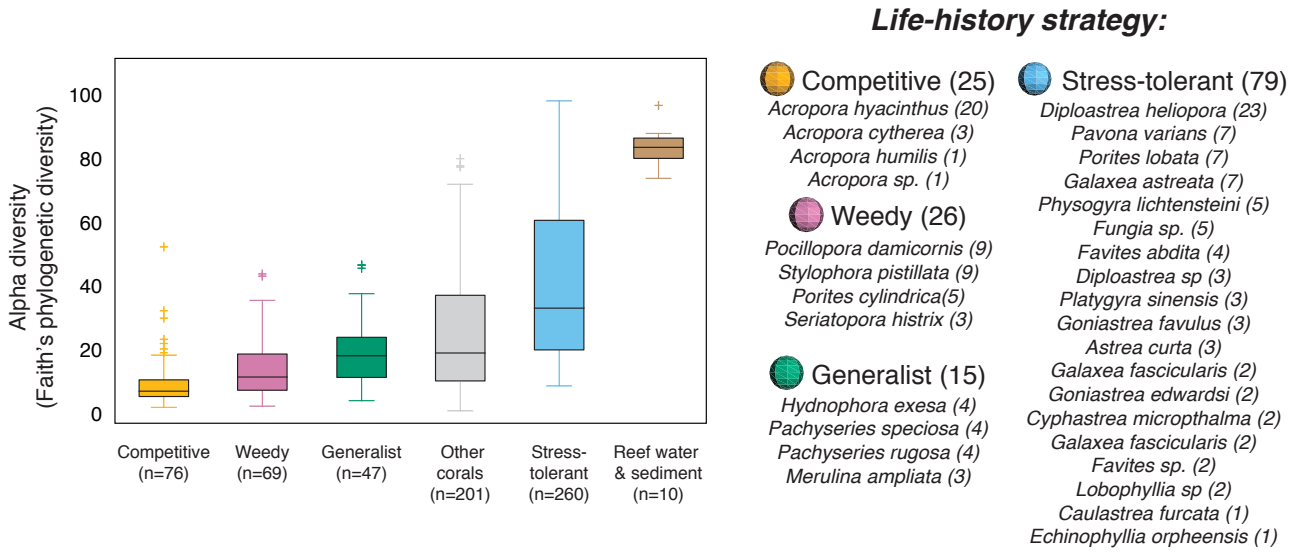


Figure 2.3. Coral life-history strategies influence tissue microbiome richness.

Corals with different life-history strategies show differences in microbial richness (data from tissue compartment only).

References

Bapst, D. W. paleotree: an R package for paleontological and phylogenetic analyses of evolution. *Methods Ecol. Evol.* **3,** 803–807 (2012).

Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7,** 335 (2010).

Caporaso, J. G. *et al.* PyNAST: a flexible tool for aligning sequences to a template

alignment. *Bioinforma. Oxf. Engl.* **26,** 266–267 (2010).
Darling, E. S. et al. Evaluating life-history strategies of reef corals from species traits. *Ecol. Lett*. 15, 1378–1386 (2012).
Daszak, P., Cunningham, A. A. & Hyatt, A. D. Emerging Infectious Diseases of Wildlife-- Threats to Biodiversity and Human Health. *Science* 287, 443–449 (2000).
Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinforma. Oxf. Engl.* **26,** 2460–2461 (2010).
Green, E. A. *et al.* Quantifying cryptic *Symbiodinium* diversity within *Orbicella faveolata* and *Orbicella franksi* at the Flower Garden Banks, Gulf of Mexico. *PeerJ* **2,** e386 (2014).
Hadfield, J. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software.* doi:10.18637/jss.v033.i02
Keck, F., Rimet, F., Bouchez, A. & Franc, A. phylosignal: an R package to measure, test, and explore the phylogenetic signal. *Ecol. Evol.* **6,** 2774–2780 (2016).
Lesser, M. P., Bythell, J. C., Gates, R. D., Johnstone, R. W. & Hoegh-Guldberg, O. Are infectious diseases really killing corals? Alternative interpretations of the experimental and ecological data. *J. Exp. Mar. Biol. Ecol*. 346, 36–44 (2007).
McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6,** 610–618 (2012).
McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE* **8,** e61217 (2013).
Oksanen, J. *et al. vegan: Community Ecology Package*. (2008).
Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinforma. Oxf. Engl.* **20,** 289–290 (2004).
Pollock, F. J. et al. Coral-associated bacteria demonstrate phylosymbiosis and cophylogeny. *Nat. Commun*. 9, 4921 (2018).
Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5,** e9490 (2010).
R Development Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2008).
Rideout, J. R. *et al.* Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ* **2,** e545 (2014).
Sunagawa, S., Woodley, C. M. & Medina, M. Threatened Corals Provide Underexplored Microbial Habitats. *PLOS ONE* 5, e9554 (2010).
Willis, B. L., Page, C. A. & Dinsdale, E. A. *Coral Disease on the Great Barrier Reef*. in Coral Health and Disease 69–104 (Springer, Berlin, Heidelberg, 2004). doi:10.1007/978-3-662-06414-6
Yilmaz, P. et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol*. 29, 415–420 (2011).

Novel generalized linear model framework identifies a clade of Symbiodiniaceae that is associated with coral skeletal samples, but not tissue or mucus

Authors: Ryan McMinds[*1], Jesse Zaneveld[2], F. Joseph Pollock[3], Mónica Medina[3], Christian R. Voolstra[4], Ben Hume[4], Rebecca Vega Thurber[1]

[1]Oregon State University, Department of Microbiology, Corvallis, OR, 97331, U.S.A.
[2]University of Washington, Bothell, School of Science, Technology, Engineering, and Mathematics, Division of Biological Sciences, UWBB-277, Bothell, WA 98011, USA
[3]Pennsylvania State University, Department of Biology, 208 Mueller Lab, University Park, PA 16802, USA
[4]Red Sea Research Center, Division of Biological and Environmental Science and Engineering (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia

[*]Corresponding author

Abstract

Neutral evolutionary processes can introduce strong correlations between unrelated traits. In any analysis of the traits of evolving organisms, these neutral correlations should be explicitly considered and controlled for. Methods to do so are common for univariate analyses, but large multivariate datasets are often analyzed in a more abstract manner that makes phylogenetic analyses more complicated. Using the Stan modeling framework, we introduce here a model for the analysis of host-microbe prevalence data, which scales to larger datasets than a previously published generative model and incorporates more flexible patterns of evolution. We apply the model to analysis of coral endosymbionts of the family Symbiodiniaceae, and find that the rate of evolution has increased in two host-specific lineages. One of these lineages is closely related to a clade of symbionts that is found only in coral skeletal samples, suggesting an interesting evolutionary transition either to or from indirect symbiosis with another eukaryotic coral symbiont.

Introduction

Under neutral processes of evolution, where traits evolve according to drift alone, shared evolutionary history can induce strong correlations among functionally independent traits (Grafen 1992, Kruuk & Hadfield 2007). The degree to which traits are affected by this phylogenetic autocorrelation is related to the rate at which those traits evolve – traits that evolve extremely rapidly (or even change within a lifetime, e.g. if they are completely uninheritable) contain very little phylogenetic autocorrelation, whereas traits that evolve slowly contain more phylogenetic autocorrelation. Given this, there has for decades been much research into the best ways to analyze biological data, whether and how to 'control' for phylogenetic autocorrelation, how to interpret the results of statistical tests, and how to define the most useful null hypothesis for given biological questions. Although there may be no perfect consensus, a number of authors have concluded that statistical models that incorporate phylogenetic information should be the default or null assumption for analyses of biological data (Grafen 1992, Moran & Sloan 2015, Douglas & Werner 2016).

As a result of multiple processes, such as vertical inheritance and environmental filtering, the microbiomes of host organisms should be expected to contain some amount of

phylogenetic signal just like any other trait (Mazel *et al.* 2018). Recognizing this, the term 'phylosymbiosis' has recently been coined to refer to phylogenetic signal in microbiome composition (Brucker & Bordenstein 2013), and the interpretation of such signal has been a frequent topic for discussion in recent literature (e.g. Moran & Sloan 2015, Douglas & Werner 2016). Likewise, the phylogenetic relatedness of microbes often contains valuable information for understanding the factors that influence their distribution. For instance, early-diverging clades of Bacteria are associated with differences in mammalian diet, although individual bacterial species are not (Groussin *et al.* 2017).

Despite the importance of phylogenetic relationships in microbiome research, many of the most commonly applied statistical methods do not explicitly consider them, and those that do tend to drastically simplify the dynamics of complex communities. For instance, use of the UniFrac community distance metric (Lozupone & Knight 2005) assumes strict Brownian evolution of microbes and reduces all the complicated abundance patterns among samples to a single abstract distance. Incorporating phylogenetic relationships of host organisms into distance metric analyses is difficult and often makes interpretation even more abstract (e.g. Revell 2009). For univariate analyses, a common way to incorporate different levels of microbial hierarchy is to analyze a dataset in the same way repeatedly, with microbial taxa aggregated at different phylogenetic or taxonomic scales each time (Mazel *et al.* 2018, Zaneveld *et al.* 2016, Pollock, McMinds *et al.* 2018). This procedure of course inflates type-I error until multiple testing corrections are applied, which then reduce the power of the analysis (Gaulke *et al.* 2018).

In light of this, there is a need for the development of holistic methods of analysis that can assess factors that influence the composition and variability of host-associated microbiomes while incorporating both host and microbial phylogenies. One such method is through the use of phylogenetic generalized linear models, which have been frequently applied to both univariate and multivariate datasets in the context of phylogenetically structured data (Revell 2010), and that have recently been developed to extend to *co*-phylogenetic datasets with interacting phylogenies (Ives & Godfray 2006, Hadfield *et al.* 2014, Björk *et al.* 2017). In such models, the highest resolution of biological units can be

used as base data, but phylogenetic relatedness between the units allows information to be pooled hierarchically across scales. However, existing linear models for cophylogenetic analysis suffer from a number of drawbacks. The first is scalability: application of one model in a recent analysis of coral-associated bacteria required almost a week of runtime for a dataset with 66 microbial units, 72 host units, and ~200 samples (Pollock, McMinds *et al.* 2018). Another problem is the flexibility of evolutionary models that have been implemented. It has been shown, for example, that phylogenetic signal in microbiome composition often diminishes gradually over time, suggesting that Brownian evolution may not be an appropriate model for datasets that cover long evolutionary timescales (Groussin *et al.* 2017, Pollock, McMinds *et al.* 2018). The coestimation of 'host phylogeny' and 'host identity' effects a number of recent models accounts for host specificity that does not contain phylogenetic signal, but in a manner that assumes the addition of discrete effects layered on top of strict Brownian evolution (Hadfield *et al.* 2014, Björk *et al.* 2017). This framework does not match the data observed in mammals and corals (Groussin *et al.* 2017, Pollock, McMinds *et al.* 2018), which suggest *gradual* decrease in phylogenetic signal, and it does not lend itself to biologically interpretable analyses of scale. A previous model introduced by Ives and Godfray (2006), on the other hand, incorporates gradual decrease in signal through the estimation of an Ornstein Uhlenbeck (OU) effect (Butler & King 2004), but the model was not extended to cases where additional effects such as location or sampling effort should be estimated. Additionally, one of the more interesting questions in analysis of host associated microbiomes is not just whether particular microbes are conserved in particular hosts, but whether particular host or microbial groups change the rate at which they evolve to form new associations. Modeling rate shifts should allow us to detect subclades of hosts and microbes whose members are more sensitive to one another than to others.

Toward these goals, we have developed a model using the Bayesian modeling framework Stan (Monnahan *et al.* 2017). Stan's Hamiltonian Monte Carlo sampler and efficient compilation of C code allows it to handle models of much greater scale and complexity than other samplers such as MCMCglmm (Hadfield 2010) or JAGS (Plummer 2003). At its core, this model is a logistic linear model with phylogenetic interaction effects that

correspond to deviations in the probability that a given microbe will be observed in a given sample. However, we have incorporated parameters that (a) scale the expected phylogenetic variance according to estimates of the strength of an Ornstein-Uhlenbeck stabilizing effect; (b) model shifts in the overall rate of evolution; and (c) incorporate uncertainty in divergence times and topology. Summary of these parameters enables direct comparison of degrees of host specificity among different clades with the trees, robustly compares effects at multiple phylogenetic scales, and places the strength of host specificity into a biologically interpretable context through comparison with the strength of other effects such as geographic location.

We use the model to analyze a novel dataset of associations between Scleractinian corals and their endosymbiotic dinoflagellates of the family Symbiodiniaceae (until recently genus *Symbiodinium,* LaJeunesse *et al.* 2018). The associations among corals and the Symbiodiniaceae have been studied in detail for decades. When associated with corals, Symbiodiniaceae are housed inside gastrodermal cells, where they photosynthesize and provide their hosts with energy and nutrients (Gates & Ainsworth 2011, Stat *et al.* 2008). Various environmental stressors, such as sustained periods of increased light levels and water temperatures, can lead to the breakdown of this symbiosis, wherein the endosymbiont is either expelled from the host tissue or broken down within it (Gates & Ainsworth 2011). The family Symbiodiniaceae is known to be composed of a diverse array of free-living forms, obligate symbionts, and facultative symbionts, and it is well established that its members are characterized by specificity to all of host, environmental, and geographic variables (Stat *et al.* 2008, Stat *et al.* 2009, Fay *et al.* 2012). Given the important role that this symbiosis plays for the health of entire coral reef ecosystems, there has been great interest in the dynamics that influence the success of a given pairing of host and symbiont in the environment (Lesser *et al.* 2013).

However, the study of this symbiosis has been hampered by the lack of a reliable genetic marker for identification and phylogenetics of the Symbiodiniaceae (Fay *et al.* 2012). Much of the existing taxonomic work on the group has relied on sequences amplified from the nuclear ITS2 region of the genome, but the high intragenomic copy number of this gene makes it difficult to interpret amplicon data from DNA collected in the

environment (Santos *et al.* 2004, Fay *et al.* 2012). In the face of these complicated datasets, researchers have found that certain combinations of sequence variants, or profiles, can be used as diagnostic markers for Symbiodiniaceae 'types'. Until recently, these profiles have been largely characterized using denaturing gradient gel electrophoresis (DGGE), which has low throughput, can miss rare but diagnostic intragenomic variants, and requires a separate step of gel excision and sequencing in order to get any sense of phylogenetic relationships among variants (Sampayo *et al.* 2009, Silverstein *et al.* 2012). Newer high throughput sequencing (HTS) technologies have enabled the discovery of previously undescribed symbiont diversity, and deeper sequencing of amplicon pools has the potential to increase the resolution and phylogenetic relationships of Symbiodiniaceae types through profiling. Recently, a tool called SymPortal has been developed to infer these profiles from HTS amplicon datasets in a manner analogous to traditional DGGE profiling.

The analysis presented in this paper represents a preliminary assessment of the Symbiodiniaceae in corals collected as part of the Global Coral Microbiome Project (GCMP), one of the first attempts to conduct Symbiodiniaceae sequence profiling using HTS amplicon data, and a proof-of-principle for the framework in which to assess the influence of other interesting factors, such as host disease susceptibility, when the full dataset (containing more thorough phylogenetic and geographic replication) becomes available. The analysis reveals the presence of weak phylogenetic signal in both microbes and hosts, and a surprising pattern of evolution within one subclade of Symbiodiniaceae: one clade of symbionts has some members that are unusually specific to particular hosts, and other members that are not host-specific but are found only in the skeletal compartment of corals.

Results

Data and model summaries

After processing and filtering the data, 252 individual colonies from 72 species of cnidarians were analyzed. Although each colony was sampled three times (once for each of the skeletal, tissue, and mucus compartments), the samples from some compartments were excluded from the final dataset after quality control. Thus the 252 colonies were

represented by a total of 598 samples. Of these samples, the mean sequencing depth was 6893 reads. Ninety-six ITS2 profiles were inferred by SymPortal and analyzed with the model.

Sample characteristics, such as host species, are specified in the model using model matrices. ‘Main effects’ of each of sample characteristic and microbial type are included in addition to interaction effects between microbes and sample characteristics. Main effects of sample characteristics correspond to differences in non–microbe-specific alpha diversity (a given factor level may be associated with a greater number of microbial types, regardless of the identity of each microbe). Main effects of microbes correspond to differences in non–sample-specific prevalence or host range (a given microbial type may be generally more common than another, regardless of any sample characteristics such as host species). ‘Interaction effects’ correspond to specific associations between microbes and sample characteristics—e.g. if a given microbe is only found in samples that come from a given host species. Overall influence of sample characteristics is compared through partitioning of the estimated variance for each factor (Fig. 3.1). This partitioning is done in a hierarchical manner for nested factors such as individual reefs within larger geographic regions, so the combined influence of these factors is pooled and they can be compared to one another more directly (compare Figs. 3.1 & 3.2).

Host species and geographic location explain the most variance in Symbiodiniaceae occurrence patterns

In applying this model to the Symbiodiniaceae dataset, five samplewise factor groups were modeled in addition to the identity of the host: tissue compartment, sampling date, sequencing depth, colony identity, and location of sampling. Thus total variance was first partitioned among 13 factor groups corresponding to the main and interaction effects of each of these and of host, plus the main effect of microbes. Three of the 13 factor groups included in the GLM explained a median of 84.7% of the variance (69.0% - 91.9%, 95% credible interval, Fig. 3.1). Of them, host-microbe specificity explained the most, a median of 37.9% of the model variance (28.4% - 46.1%). Specificity of microbes to sampling location explained 28.1% (20.1% - 35.7%). Specificity to individual colonies (in some senses the ‘residual’ variance) explained 17.6% of the estimated variance

(12.8% - 23.1%). Sampling location was treated as a hierarchical variable subdivided into levels of 'ocean' (Pacific, Indian, or Atlantic), 'ocean area' (Coral Sea, Tasman Sea, South Pacific, Mediterranean, Red Sea, Caribbean, Eastern Indian, South China Sea, and Western Indian), and finally, individual reefs. Of the variance explained by specificity to location, 'ocean area' explained the most variance, at 68.5% (45.1% - 88.4%, Fig. 3.2). No alpha diversity factors had a median estimate of more than 2.2% of the model variance, although the 95% upper credible bounds for 'tissue compartment' and 'sequencing depth' were 10.1% and 18.8%, suggesting that these factors could be considerably more influential than their medians (0.4% and 2.2%, respectively) would imply.

Phylogenetic signal in Symbiodiniaceae occurrence patterns is weak

The mean estimate for the strength of the mean-reverting Ornstein-Uhlenbeck (OU) effect on host evolution was 7.8 (3.5 - 13.1, 95% credible interval). This translates to a mean of just 20% (10% - 50%, 95% credible interval) of variance that is explained by the first 90% of the evolution of hosts (since the divergence of Hydrozoa and Anthozoa). The mean OU effect strength for microbes was 4.9 (2.6 – 8.6, 95% credible interval), translating to 40% (20% - 60%, 95% credible interval) of modern variance occurring in the first 90% of the evolution of microbes (since the divergence of *Symbiodinium sensu stricto*, former '*Symbiodinium* Clade A', from the rest of the Symbiodiniaceae; Fig. 3.3).

Two instances of increased evolutionary rates in coral-Symbiodiniaceae specificity

When considering the summed effects of all ancestral nodes (the contrast between each node and the grand mean), there were two microbial clades with significant changes in their rate of evolution within distinct host clades (Figure 3.4). A subset of the C15 radiation (containing DIVs C15, C22b, C15l, C15n, C15.8, and C15s) had significantly higher variance among the massive *Porites* corals *P. lobata* and *P. lutea* (logit 0.08, 95% credible lower bound). Another *Cladocopium* subclade (containing DIVs C1, C1l, C42.2, C42a, C1d, C8, C8d, C42e, C1z, C1b, and C1ab) had significantly higher variance among the Indo-Pacific Pocilloporidae (logit 0.004, 95% credible lower bound). The timing of

these rate increases could not be ascertained with 95% confidence, in part because the effective sample sizes were low for contrasts between internal nodes.

Numerous specific associations between microbial types and sample characteristics

There were 314 significant host-microbe associations. These consisted of 50 out of 62 reliably identified host coral species associated with 77 out of 96 Symbiodiniaceae ITS2 profiles. Nine entire clades of Symbiodiniaceae were positively associated with ten groups of hosts: i112 (four profiles containing C116 variants) with the clade containing all *Pavona*; i139 (four profiles containing C3 and C50 variants) with the clade containing all *Acropora*; i142 (four more profiles containing C50 variants) with the clade containing all *Acropora* and independently with *Favites abdita*; i149 (two profiles containing C40 variants) with all *Echinopora* and independently with *Hydnophora exesa*; i155 (three profiles containing C1 variants) with all *Leptastrea*; i156 (two of the variants within i155) with *Pocillopora grandis, Pocillopora damicornis, Pavona varians*, and all *Cyphastrea*; i167 (two profiles containing C123 & C1 variants) with *Seriatopora hystrix*; i175 (two profiles containing D2.2 variants) with *Diploastrea heliopora*; i177 (two profiles containing other D2 variants) with *Echinopora mammiformis*; i181 (six profiles containing *Durusdinium* variants) with *Stylophora pistillata*; and i189 (four profiles containing A3 variants) with *Millepora*. In addition, numerous other entire host clades were positively associated with individual Symbiodiniaceae profiles. For instance, profile 13312 (C40-C3-C115-C40c-C40e) was separately associated with five distinct host clades: the clade uniting Diploastreidae, Lobophyllidae, and Merulinidae, and the clades corresponding to each of the genera *Physogyra*, *Turbinaria*, *Pachyseris*, and *Galaxea*.

There were 25 Symbiodiniaceae profiles significantly associated with specific ocean areas, four associated with particular sampling days, one associated with a specific reef, and one profile (C15h) plus a closely related clade of four profiles (all C116 variants) associated with the coral skeleton (Table 3.1). Two profiles (including 13312) were additionally associated with higher sequencing depths, and there were 189 associations between Symbiodiniaceae profiles and individual coral colonies. Negative effects were limited to nine profiles that were more likely to be found with lower sequencing depths.

There were no microbial clades or profiles associated with changes in nonspecific prevalence across all samples, and no sample characteristics, host clades, or host species that were associated with changes in nonspecific alpha diversity.

Rare occurrence of mixed infections

There were 115 samples (19.2% of samples) that were predicted to have multiple Symbiodiniaceae type profiles. However, we suspected that profiles significantly associated with lower sequences depths were artifacts of the SymPortal type profiling algorithm because eight of them were single-DIV profiles and all occurrences of them were in samples that also contained a superset profile. Therefore we removed them from data matrix before further assessing the prevalence of mixed infections. After doing so, mixed infections of corals occurred in 94 samples (15.7% of samples). Only 22 (3.6% of samples) contained three or more profiles, and three (0.5% of samples) contained four or more.

Computational performance of the model

Useful effective sample sizes for most parameters were gained from the model by running it for only 2,048 iterations and sampling only 1,024 draws per chain, although some evolutionary rate shift parameters did not have enough effective samples for posterior inference. Although each chain was run using a different draw from the posterior host phylogeny, almost all comparable parameters converged on similar values across chains (Rhats < 1.1). There were no divergent transitions or iterations that hit the maximum tree depth after warmup. The model took 22 hours to run with very little variation among chains and consumed ~3.2 gigabytes of memory per chain. All effects are listed in Supplementary Data 3.1.

Discussion

In this study, we have introduced a method for estimating and controlling for phylogenetic signal in microbiome composition (phylosymbiosis) using a model-based approach with biologically interpretable parameters and summaries. A single analysis with this model enables inferences often attained through many separate analyses, such as multivariate tests like PERMANOVA and univariate regressions of microbial abundances or alpha diversity statistics. Co-estimating all these effects allows each to be controlled

for in the interpretation of the others, and can provide clearer insight into the biological processes that produce the patterns.

For instance, a common paradigm in the study of coral-algal symbiosis is that both hosts and symbionts can be generally divided into categories of 'generalists' vs. 'specialists' (Fabina *et al.* 2012). However, in analysis of this dataset, no significant differences were found in the main effects of either host or microbe, which would correspond to differences in overall degree of specialization. Instead, many significant interactions were found between tips and clades of each group, indicating varying degrees of specificity at multiple phylogenetic scales. This reflects the fact that the terms generalist and specialist are ill-defined without reference to the phylogenetic scales being considered (Stat *et al.* 2006). A given coral species may have generalist qualities in that it can associate with Symbiodiniaceae of all of clades A, C, and D, but simultaneously be a specialist in that it only associates with particular species within each of those clades. Another coral species could display the opposite pattern, associating with any number of clade C symbionts (a generalist in terms of the absolute number of potential partners), but never with symbionts of clades A or D (a specialist in terms of phylogenetic range). In this analysis, an interaction effect between a tip on the host tree and an internal node on the microbial tree indicates a 'specialist' type association with that microbial clade, but a 'generalist' type of association *among* the members of that microbial clade. A lack of significant effects with internal nodes accompanied by the presence of significant effects between tips indicates finer-scale specificity. This model-based framework thus helps avoid oversimplification of generalist/specialist patterns.

In addition to assessments of present-day patterns of association, there is interest in the rate at which such patterns may change through the evolution of both host and symbiont. Multiple studies have now demonstrated that assumptions of uniform rates of evolution lead to inaccurate reconstruction of phylogenies from nucleotide sequence data (Drummond & Suchard 2010, Wertheim *et al.* 2010, Revell *et al.* 2012). Although the assumption has not been thoroughly assessed with traits other than nucleotide sequences, it seems prudent to assume that rates of evolution of most traits are variable, and to allow for such variation more broadly in phylogenetic analyses. Models that have been

developed to incorporate rate variation in phylogenetic inference should find natural translations in the analysis of other traits. For this analysis, we included parameters for rate variation equivalent to the lognormal correlated relaxed molecular clocks frequently used in phylogenetics. We chose not to scale the prior for magnitude of rate shifts in according the branch lengths because doing so would assume very gradual change over time, whereas it has been demonstrated that changes in rate can occur very abruptly (Drummond *et al.* 2006). One process that could contribute to abrupt changes in rate is the initialization of codivergence, where branch lengths should become irrelevant and each speciation event should be accompanied by an increase in total variance. Modeling rate shifts as we did, with correlation to ancestral rates but without consideration of branch lengths, can be seen as a compromise model. Although this suggests the need for continued development of the model, the current implementation was able to lend valuable insight into two interactions between clades of hosts and symbionts, which may be evolving at increased rates relative to one another due to codivergent or coevolutionary mechanisms, or due unidirectional evolutionary responses by one party to the other.

A surprising finding of this study was the specific association of a group of symbionts nested within the C15 radiation with the skeletal compartment of corals (Table 3.1; Figure 3.5). We did not expect the composition of the compartments within a single coral to be different; the tissue compartment effect was included in this study partly for the simple reason that the samples existed. The separate compartment samples were collected more for the study of their bacterial communities than for analysis of Symbiodiniaceae (Pollock, McMinds *et al.* 2018). Regardless, inclusion of both the microbial phylogeny and the tissue compartment factor enabled the discovery this association, which would probably have been missed if each profile were analyzed as independent units. Each individual variant of C116 was only found in a single sample, or at most two samples, so statistical support would not have existed for the association. However, the incorporation of the phylogeny in the analysis allowed pooling of information and revealed the clade's overall greater prevalence in skeletal samples. Given the relative rarity of C15 type Symbiodiniaceae in non-*Porites* corals (Krueger & Gates 2012), it is tempting to speculate that many of the records that do exist reflect the presence of the one of the

skeleton-specific ecotypes observed here rather than one of the otherwise *Porites*-specific ecotypes. Indeed, a number of such reports are either unclear about the anatomical compartments analyzed or used methods that clearly include skeletal material (e.g. Gong *et al.* 2018). In light of this, care should be taken in interpretation of the presence of this type of Symbiodinium in coral samples. It would seem most likely that its association with corals is indirect; that it is a symbiont of another eukaryotic denizen of coral skeletons. Previous work has shown, for instance, that a Symbiodiniaceae type related to the C15 radiation (C66; see Figure 3.5b) is a symbiont of the coral-associated acoel worm *Waminoa* (Barneah *et al.* 2007), and that C15 types are found in foraminifera and on the surface of macroalgae (Venera-Ponton *et al.* 2010). Alternatively, there is the possibility that the skeleton specific C15/C116 ecotypes are located in different tissue sections of the coral animal, or that they live extracellularly in the skeleton. Distinguishing among these possibilities could be accomplished without more sampling in the field by an analysis of *Waminoa* or other eukaryotic sequences in metagenomic or other amplicon data from the existing GCMP samples. To understand the direction of evolution within this group (i.e. whether the ancestor of the group was located in diverse coral skeletons and evolved a specific association with *Porites* host tissue, or if the opposite transition occurred, or if switches between the ecotypes have occurred multiple times) a more robust phylogenetic analysis should be performed on this group of Symbiodiniaceae (Figure 3.5).

A number of improvements can be made to this modeling framework. Importantly, alternate models should be explored for the evolution of rate shifts. For the analysis of the evolution of nucleic acid sequences, uncorrelated models for rate variation such as random local clocks have been shown to be more reliable (Drummond & Suchard 2010), but many of these models incorporate discrete parameters that do not translate well into models in Stan. Stan requires continuously differentiable posterior gradients or else must marginalize over all discrete combinations of possibilities, which can be computationally expensive. It may be important to consider whether gradual evolution of rates should be modeled, similar to the way location effects are modeled, rather than the current shifts at each node that have priors that are not proportional to branch lengths (Drummond *et al.* 2006, Ronquist *et al.* 2012, Jhwueng & Maroulas 2015). The exact opposite pattern, where each divergence event is associated with a sudden and discrete increase in

variance, could be co-estimated to account for processes such as the start of strict codiversification, as discussed above. While the currently implemented compromise model allows for both processes to occur, it cannot distinguish between them. Incorporating both of these processes of rate variation separately into a model could help reveal the processes that lead to rate shifts such as those observed in the current study.

Additionally, the model could be improved by handling phylogenetic uncertainty better in a number of ways. The current model does so in part by allowing input node heights to vary prior to the layering on of OU effects and evolutionary rate shifts, but the prior confidence in all node heights is estimated with a single parameter. It should instead be feasible to incorporate summary statistics such as the standard deviation of each node's height from a set of posterior phylogenies, which could be used to individually calibrate the priors for each node. To sample more phylogenies of both host and symbiont and incorporate more topological uncertainties, significant speed or memory improvements would be needed. However, the current model ran in 1/6 of the time of one alternative, even though it incorporated ~3X the number of samples and ~1.5X the number of microbial taxa. Additionally, there is a strong potential to increase the speed of model fitting through reparameterization because Stan's performance is highly dependent on the geometry of raw parameter space. Summaries across chains could also be improved by weighting the draws from each chain according to the posterior probability of the phylogeny used for that chain, or even after updating the posterior support for each topology using the microbiological data through the use of bridge sampling (Gronau 2017).

Finally, it should be noted that, given the scaling of branch lengths using an Ornstein-Uhlenbeck process, the current contrasts involving deeper nodes in the phylogeny do not necessarily represent contrasts with ancestral states, but rather contrasts between modern mean values for each hierarchical set of species. The calculation of ancestral states would provide additional insight into the evolution of these groups, and would require relatively simple reparameterizations in the model. These improvements are all ideal subjects of research for the future development of comprehensive analyses of host-microbe interactions.

Acknowledgements

The authors would like to acknowledge all of the members of the Global Coral Microbiome Project who assisted us in the collections of these samples worldwide, including but not limited to: Tasman Douglass, Margaux Hine, Frazer McGregor, Cathie Page, Gergely Torda, Lyndsy Gazda, Jamie Lee Proffitt, Gabriele Swain, Alaina Weinheimer, Katia Nicolet, David Bourne, Kathy Morrow, Karen Weynberg, Maren Ziegler, Anna Roik, Mark Vermeij, Kristen Marhaver, Pedro Frade, Ben Mueller, the Burkepile lab, Le Club de Plongee Suwan Macha, Jean-Pascal Quod, Jerome Payet, Amelia Foster, Danwei Huang, Yossi Loya, and Raz Tamir. Thank you Steven Hubbs for preparing many samples in the lab for sequencing. Thanks are also extended to Johannes Bjork for introducing RM to the world of species distribution modeling and discussing theory early on. The authors also acknowledge the staff of the Coral Bay Research Station, Lizard Island Research Station, Lord Howe Island Marine Park, Lord Howe Island Research Station, RV Cape Ferguson, Gump Marine Station in Mo'orea, French Polynesia, and the Caribbean Marine Biological Institute in Curaçao for their logistical support. We would also like to thank Stan forum members for their friendly and skilled guidance.

This work was supported by a National Science Foundation Dimensions of Biodiversity grant (#1442306) to RVT and MM.

Author Contributions

Contributions: RM, RVT, JZ, and MM designed the experiment. RM and FJP collected samples and extracted DNA. RM and BH performed the data analysis. RM wrote the manuscript.

Conflict of Interest Statement

The authors declare no conflict of interest.

Materials and Methods

Selection of target sites

We aimed to collect coral specimens spanning coral phylogenetic diversity from a variety of reefs across the globe. We targeted collection based on the 21 major coral clades

defined in one of the most recent molecular phylogenies available at the start of the project (Fukami *et al.* 2008). Many of these monophyletic groups have since been defined as family-level taxa. A majority of this diverse set of coral samples were collected at several sites on the east and west coast of Australia. These included Ningaloo Reef (Western Australia), Lizard Island, multiple reefs along the northern Great Barrier Reef, and Lorde Howe Island. Samples at Lizard Island were collected in both Summer and Winter, allowing for comparison of seasonal effects at one site across diverse corals. Additional coral samples were collected from reefs in Saudi Arabia, Israel, Réunion, Curaçao, Singapore, and Mo'orea (Supplementary Table 3.1). As of this analysis, only the samples belonging to *Porites* and *Pocillopora* corals were sequenced from Curaçao, Singapore, Réunion, and Saudi Arabia. These two genera were selected for initial analyses due to their extensive ranges and their large sample size within our collection.

Coral sampling

All coral samples were collected by AAUS-certified scientific divers, in accordance with local regulations. Relevant permit numbers are: CITES (PWS2014-AU-002155, 12US784243/9), Great Barrier Reef Marine Park Authority (G12/35236.1, G14/36788.1), Lord Howe Island Marine Park (LHIMP/R/2015/005), New South Wales Department of Primary Industries (P15/0072-1.0, OUT 15/11450), US Fish and Wildlife Service (2015LA1632527, 2015LA1703560), and Western Australia Department of Parks and Wildlife (SF010348, CE004874, ES002315). Only healthy corals were collected.

One goal of the project was to compare microbial diversity associated with the coral mucus, tissues and skeletons across many coral colonies. Each of these compartments represents a simplification of more complex structure, but for this project, we felt that a consistent reporting of these compartments across diverse corals represented a tractable step forward. Mucus was collected by gently agitating the surface of corals for ~30 seconds with a blunt 10 mL syringe. Exuded mucus or surface water (if no visible mucus was exuded) was then collected by suction. On the surface, settled mucus typically formed a distinct visible layer within the syringe. This was expelled into a cryogenic vial and stored in a dry shipper charged with liquid nitrogen for subsequent processing.

Tissue and skeletal samples were collected from each colony by hammer and chisel or with bone shears. These fragments were placed in sterile WhirlPaks and returned to the surface where they were snap frozen in a liquid nitrogen dry shipper until processing. In the laboratory, tissue was washed with sterile seawater (which removed visible mucus and detritus), then separated from skeleton using pressurized air of between approximately 800 and 2,000 PSI (an 'air gun'). Skeleton was sampled using a sterile chisel to isolate a ~1 cm^3 region of skeleton that was not in direct contact with coral tissue. Skeleton samples were collected without regard to endolithic algae presence or absence (i.e., endolithic algae were neither specifically targeted nor excluded). Tissue slurries and skeleton samples were added directly to a MoBio PowerSoil Kit (MoBio Laboratories, Carlsbad, California) bead tube (which contains, among other things, a solution of guanidinium preservative) and stored at -80 °C until DNA extraction.

ITS2 library preparation and sequencing

DNA was extracted from skeleton, tissue, and mucus using the MoBio Powersoil DNA Isolation Kit. Two-stage amplicon PCR was performed to amplify the Symbiodiniaceae ITS2 region. First, 35 PCR cycles were performed using ITS2 primers (underlined) with adaptor sequences at the 5' ends: ITS_Dino-forward_MAf (5'- TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG GTG AAT TGC AGA ACT CCG TG-3') (Pochon *et al.* 2001) and its2rev2-reverse_MAr (5'-GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACAG CCT CCG CTT ACT TAT ATG CTT-3') (Stat *et al.* 2009). Each 12.5 µL PCR reaction was prepared with 6.25 µL AccuStart II ToughMix, 0.25 µL forward primer (10 µM), 0.25 µL reverse primer (10 µM), 0.5 µL template DNA, and 5.25 µL MilliQ water. PCR cycles consisted of a 5 min denaturation at 94 °C; 35 cycles of 40 s at 94 °C, 120 s at 59 °C and 60 s at 72 °C; and 5 min at 72 °C. Next, amplicons were barcoded with a second PCR using primers consisting of Nextera adaptors and indices from Schloss *et al*. (2011). The second PCR reaction consisted of 12.5 µL ToughMix, 1 µL of each primer, 9.5 µL MilliQ water, and 1 µL of the PCR product from the first step, and a cycle of 5 min denaturation at 94 °C and 12 cycles of 30 s at 94 °C, 30 s at 63 °C, and 30 s at 72 °C, and a final 10 min extension at 72 °C. Products of this reaction were purified with AMPure XP beads, pooled in equimolar

concentrations, and sequenced on the Illumina MiSeq sequencing platform (using 350 bp V3 chemistry) at the Center for Genome Research and Biocomputing (CGRB) at Oregon State University.

SymPortal ITS2 type profiling, initial quality control, and phylogenetics

Raw ITS2 sequence reads were demultiplexed with fastq-multx (Aronesty 2013) and submitted to the SymPortal pipeline to generate a table of profiles. Briefly, the SymPortal pipeline uses Mothur (Schloss *et al.* 2009) to trim adaptor sequences, merge paired-end reads, and filter low-quality sequences from the data. It then performs minimum entropy decomposition with the *decompose* script (Eren *et al.* 2014). Then, it separates the sequence variants into bins corresponding to their genus (formerly lettered major clade), and searches for groups of variants within each bin that co-occur frequently among all samples in the database. The theory is similar to that traditionally used in *Symbiodinium* taxonomy, which relied on ITS2 DGGE profiles to identify unique types. The output of the SymPortal pipeline is one table with counts corresponding to the number of sequence reads in a sample that are inferred to belong to each profile, and another table with counts for each *defining intragenomic variant* (DIV) of which profiles are composed.

A phylogeny of Symbiodiniaceae ITS2 profiles was generated by first combining DIV sequences with a database of ITS2 sequence database composed of all the sequences in the file ‘mec12869-sup-0001-FileS1.fasta’ from Arif *et al.* 2014 plus four outgroup ITS2 sequences retrieved from NCBI (accessions KT389903: *Gyrodiniellum shiwhaense*, DQ195357: *Gymnodinium beii*, JN558110: *Polarella glacialis*, and LC068842: *Biecheleria brevisulcata*). Sequences were aligned using mafft (Katoh & Standley) with options “--adjustdirectionaccurately”, “--ep 0”, “--maxiterate 1000”, and “--genafpair”. A phylogenetic gene tree was created using FastTree (Price *et al.* 2010) with a GTR model and four rate categories. This tree was rooted in R using the outgroups, then all non-DIV sequences were trimmed from the tree. The pairwise unweighted UniFrac distances between profiles was calculated using this tree and a matrix that related the profiles (as if they were samples in a community data matrix) to the DIVs that identify them. The resulting distance matrix was provided to *ape*’s (Paradis *et al.* 2004) function fastme.bal to produce a minimum evolution ‘species’ phylogeny, which was rooted and then made

ultrametric with the default parameters (including correlated rates of evolution) of the *ape* function chronos.

Host phylogenetics

Host phylogenies were randomly sampled from the set of posterior draws provided by Huang & Roy 2015. Outgroup taxa that were not included in that study were added by first generating posterior draws of a phylogeny that included only the outgroup taxa, *Acropora palmata* and *Pocillopora damicornis*. This phylogeny was created by combining the multigene protein alignments from Pratlong *et al.* 2016 with 12S mitochondrial data generated by our lab and downloaded from NCBI. An analysis was conducted using BEAST2 v2.4 (Bouckaert *et al.* 2014) with bModelTest v0.3.2 (Bouckaert & Drummond 2017) as a site model for mitochondrial evolution, a gamma site model with 4 rate categories and an LG substitution model for the Pratlong alignments, a relaxed log normal clock, a single linked tree, monophyly constraints on Scleractinia, and 100 million iterations. After a burn-in of 25 million samples, samples were saved every 5,000 iterations, for a total of 15,000 posterior samples saved. From these samples, ten were randomly selected, and ten trees from the Huang and Roy collection were also randomly selected. Each outgroup tree was scaled such that the age of the Scleractinian-Corallimorpharian split matched the age of that split in the paired Huang and Roy tree, then the two trees were concatenated together.

Bayesian species distribution model

The distributions of the various Symbiodiniaceae species among samples were analyzed using a novel logistic linear model implemented in Stan. The model incorporates the phylogenetic relatedness of both hosts and symbionts and can incorporate phylogenetic uncertainty both by allowing node heights to vary during model fitting and by combining draws from chains run with independent draws from a posterior sample of phylogenies. Node height adjustment is accomplished by first calculating for each node in each input phylogeny the logit of the proportion of the distance between its parent node and the tips of the tree. This value is then adjusted during model fitting by adding to it an offset drawn from a normal distribution (with standard deviation drawn from an exponential distribution of mean 1.0) and recalculating node heights starting at the root. Rather than

requiring an *a priori* assumption of a strict and stationary Brownian model, the model dynamically estimates both the strength of an Ornstein-Uhlenbeck attraction toward the mean (Butler & King 2004; drawn from an exponential distribution,with mean 1.0), and changes in the rate of evolution. Changes in rate of evolution are incorporated by drawing variance scaling parameters for each subclade from a lognormal distribution of median one and (log) standard deviation of 1.0. Interactions between subclades of the two trees are also incorporated using lognormal scaling parameters, in a way that can be viewed as allowing the rate of evolution of the host to be different from the perspective of each microbial clade, while shrinkage draws them toward the global means. The variance of the lognormal scaling parameters is estimated by partitioning a total 'meta variance' into three categories: global host modifiers, global microbial modifiers, and host-microbe interaction modifiers. A flat, symmetric Dirichlet prior is placed on the proportions of total meta variance, and the total meta variance itself is the square of a parameter drawn from an exponential distribution (with a mean of $0.1 \times \sqrt{3}$).

The probability of observing a given microbe in each sample is calculated by adding predictor terms corresponding to the interactions between a model matrix that maps samples to their hosts, ancestral nodes, and arbitrary other factors; and a matrix that maps microbial identities to their ancestral nodes. In other words, in a dataset with $T_m$microbial species, A_m microbial ancestral nodes, T_h host species, A_h host ancestral nodes, and E_s estimated samplewise effects, the model uses $N_p = (T_m + A_m) \times (T_h + A_h + E_s)$ total predictors. Each predictor has a normal prior with location zero and a variance estimated as described below. Samplewise effects are grouped into $F_s <= E_s$ factors that have shared variance parameters, and hierarchically related factors are assigned to $G_s <= F_s$ groups. In the case of factors designed with sum-to-zero contrasts, the reference level is calculated as the negative sum of all other levels, this value is also given the same normal prior as all the other levels, and to preserve their marginal variance, it is scaled by $\frac{1}{1-\frac{1}{N_l}}$, where N_l is the total number of levels.

Each factor's variance is partitioned out of an estimated total variance, which is the square of a parameter drawn from an exponential distribution (with an expectation of

$\sqrt{3 + 2 \times G_s}$ in this study). Each of $K = 3 + 2 \times G_s$ categories gets a proportion of the total variance, with proportions assigned even prior probabilities (a flat symmetric Dirichlet prior). This is a similar method to the DECOV prior used in the package *rstanarm* (Goodrich *et al.* 2018). The number of categories is derived by recognizing the inherent three factors of host alpha diversity (the 'main effects' of each host node), microbial prevalence (the 'main effects' of each microbial node), and host-microbe specificity (the 'interactions' between host and microbial nodes), and allowing each samplewise factor to have separate variances for their influence on alpha diversity (their 'main effects') and specificity (their 'interactions' with each microbial node). The variance assigned to each group is then further subdivided among the factors that belong to that group with another flat Dirichlet prior.

Within each category that interacts with a phylogeny, variance is further partitioned in proportion to branch lengths. For samplewise factors, the modified microbial tree is scaled so that the mean root-to-tip distance is one, and the variance of each node's interaction with the factor is the product of its leading branch length with the factor's total variance. Variance for interactions between nodes in the host and microbial phylogeny is calculated by taking the product of each node's modified global branch length, multiplying it by the lognormal scaling parameter specific to that pair of nodes, and then scaling the resulting matrix such that the mean root-to-tip distance is equal to that of the 'host specificity' factor.

Location effects are parameterized in a non-centered manner, such that the raw parameters are drawn from a standard Normal distribution, then scaled, and finally combined through multiplication with model matrices. Although it is possible to generate a model matrix that relates a single vector of parameters to all interactions between hosts and microbes via the Kronecker product of a samplewise model matrix and a microbial ancestry matrix, this approach was found to be computationally impractical. In raw form, such a matrix is extremely large, and consumes excessive memory. Stan's sparse matrix utilities reduced the memory load, but the runtime was still impractically slow. The practical solution was to organize the raw parameters in a matrix indexed by each samplewise effect and each microbial ancestral node. This matrix is multiplied first by a

samplewise model matrix, and subsequently by the microbial ancestry matrix on the second dimension.

This model was run using samplewise factors corresponding to the ocean, ocean subregion, and reef from which a sample was taken (all collected into a 'location' group), the date on which it was taken, the name of the colony, the tissue compartment, and the sample's total sequencing depth. The sequencing depth was log transformed, centered, and scaled to have a standard deviation of 1.0. A model matrix was constructed using sum-to-zero contrasts for all other factors. Samples of whole corals including all three compartments were included by fixing their tissue compartment effects to zero. The model was run separately for each of ten randomly sampled host phylogenies using *rstan* v2.17.3 (Monnahan *et al.* 2017), 2048 iterations, default parameters for adapt_delta, max_treedepth, and warmup, to retain 1024 post-warmup samples per tree.

Model result summaries and hypothesis testing

After model fitting, the draws from each tree were combined into a single 'stanfit' object that treated each tree as a separate chain. The significance of both location and (log-scale) variance effects was assessed by determining whether A) 95% of draws across all chains were either positive or negative, B) the effective sample size was greater than 200, and C) the corresponding Rhat was less than 1.1. Host node effects were only tested if an equivalent contrast existed across all ten sampled phylogenies. A number of different contrasts were considered. The contrast between the value of a given node and its parental node corresponds directly to the scaled parameters. Contrasts between the value of a given node and its 'grandparent' can be calculated by adding the value indexed by the node to the value indexed by its parent, and contrasts with more basal nodes can be calculated similarly. To compare a given node's value to the grand mean, all ancestral values are summed. Assessing multiple such contrasts can be useful because each parameter may in itself have a wide confidence interval, while deeper contrasts combine the information of more observations and may be estimated with more confidence.

Assessment of potential biases induced by the model or by sampling strategy

Model biases were checked by two methods: sampling without data (i.e. 'sampling from the prior'), and sampling with the samples shuffled. There were no parameters estimated to significantly differ from zero in either case.

Figures and Figure Legends

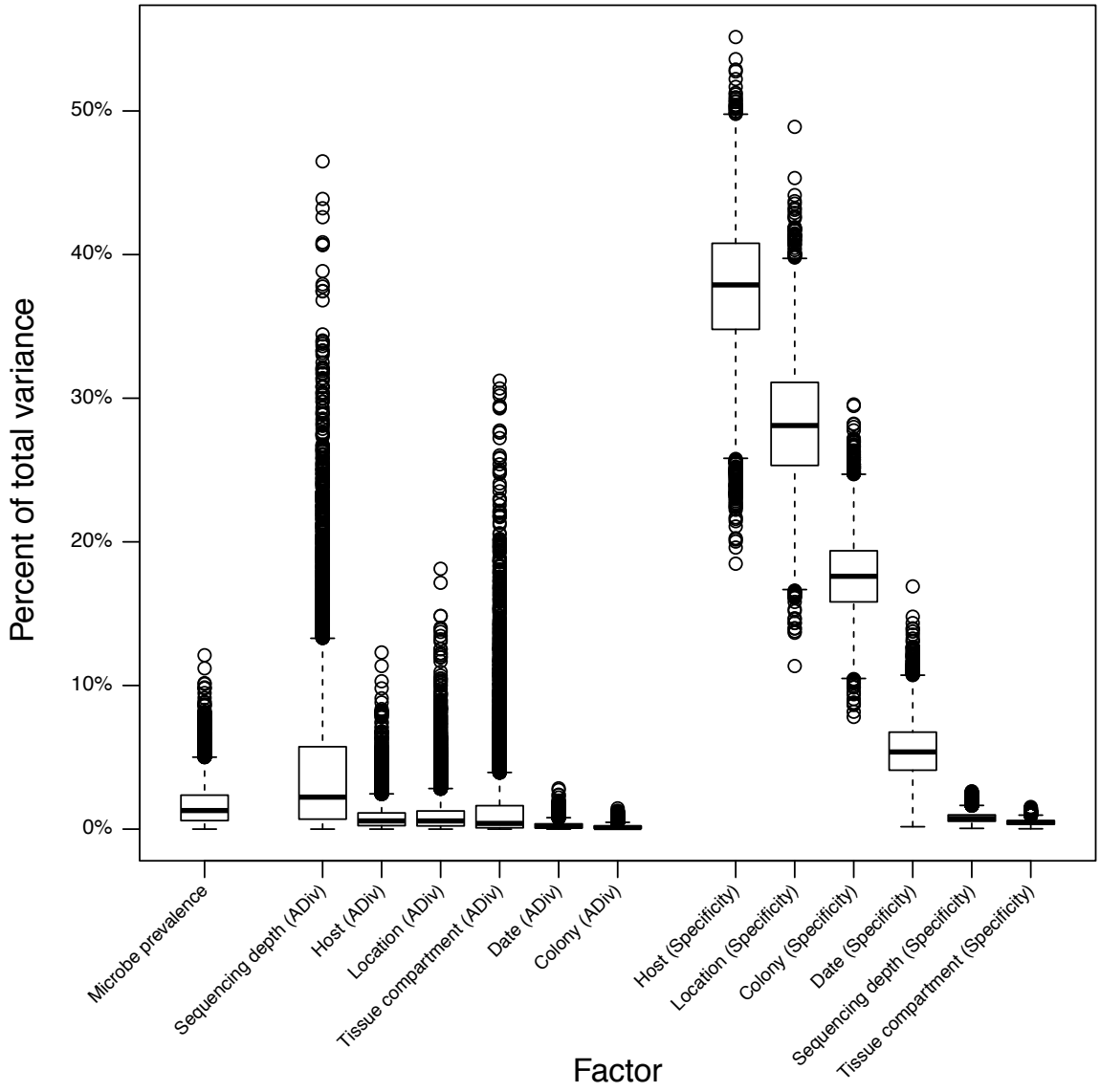


Figure 3.1. Host specificity and location specificity explain a majority of variance in the distribution of *Symbiodiniaceae*.

All draws of the relative influence of each factor in the model are plotted, with median estimates as horizontal lines and boxes ranging from the 25th to the 75th percentiles. Effects of specificity were generally much stronger than effects of alpha diversity or prevalence, and the majority of variance could be explained by specificity to hosts, locations, and individual colonies.

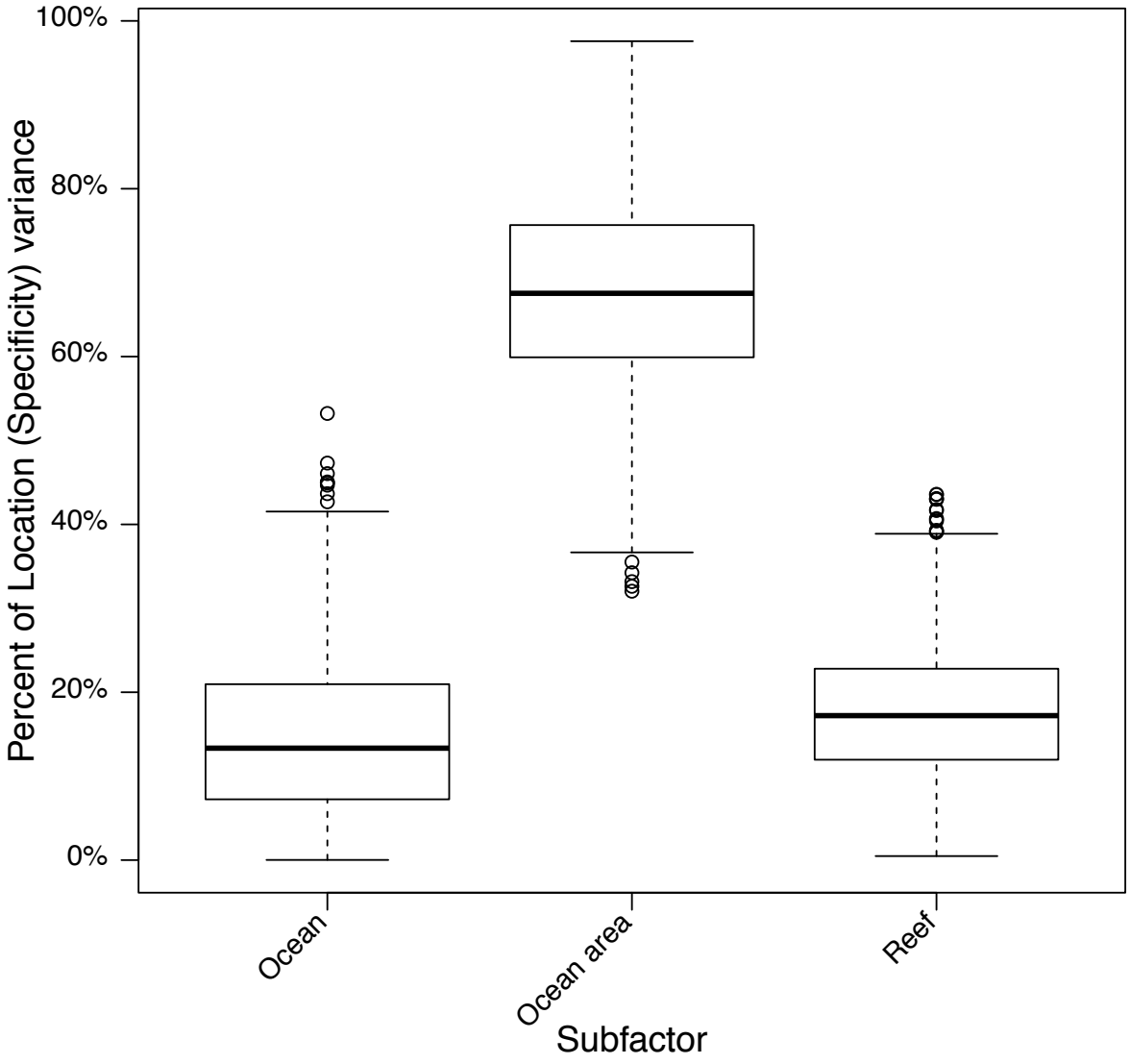


Figure 3.2. Intermediate geographic scales explain the majority of location-specific patterns.

The variance explained by specificity to location is subdivided into levels of geographic hierarchy. Intermediate geographic scales explained more variance than either differences between oceans or differences between individual reefs, suggesting that local environmental variables may have a strong influence on the composition of communities of Symbiodiniaceae.

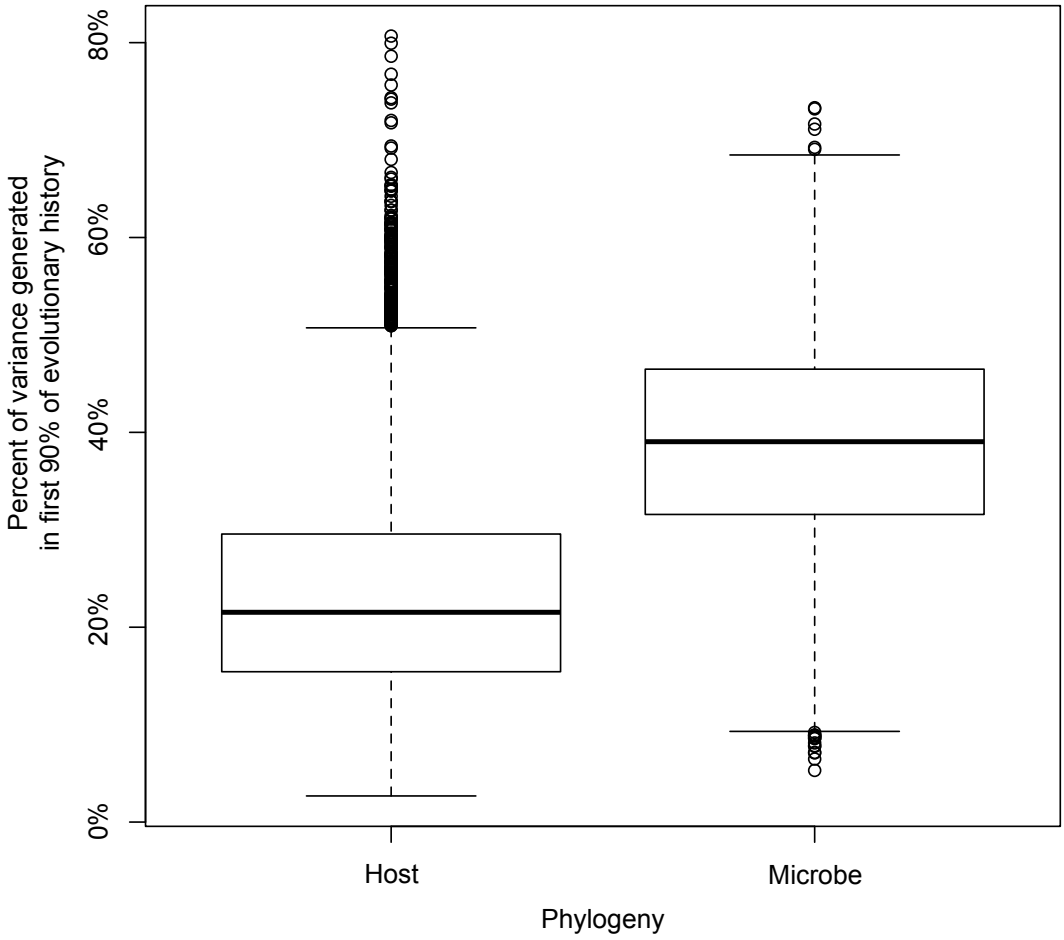


Figure 3.3. Partitioning of variance at different scales of the host and microbial phylogenetic history.

Given an Ornstein-Uhlenbeck (OU) process of evolution, phylogenetic signal decreases throughout time. The branch lengths of an ultrametric phylogeny can be scaled to represent this loss of phylogenetic signal by reducing the expected contribution to variance by edges near the root of the tree. To translate the value of the OU effect

parameter into a more interpretable scale, an arbitrary temporal threshold of 90% of evolution since the root was chosen, and the amount of variance explained prior to that time was calculated. A large amount of variance explained by historical host evolution can be interpreted as a signal of phylosymbiosis. The low values observed in this system indicate that phylogenetic signal is present but not strong, and demonstrate that using a Brownian model of evolution (with variance exactly proportional to time) would not be appropriate for analysis of these data. Note that the values for the host and microbe are not comparable to one another because the ages of their roots are different.

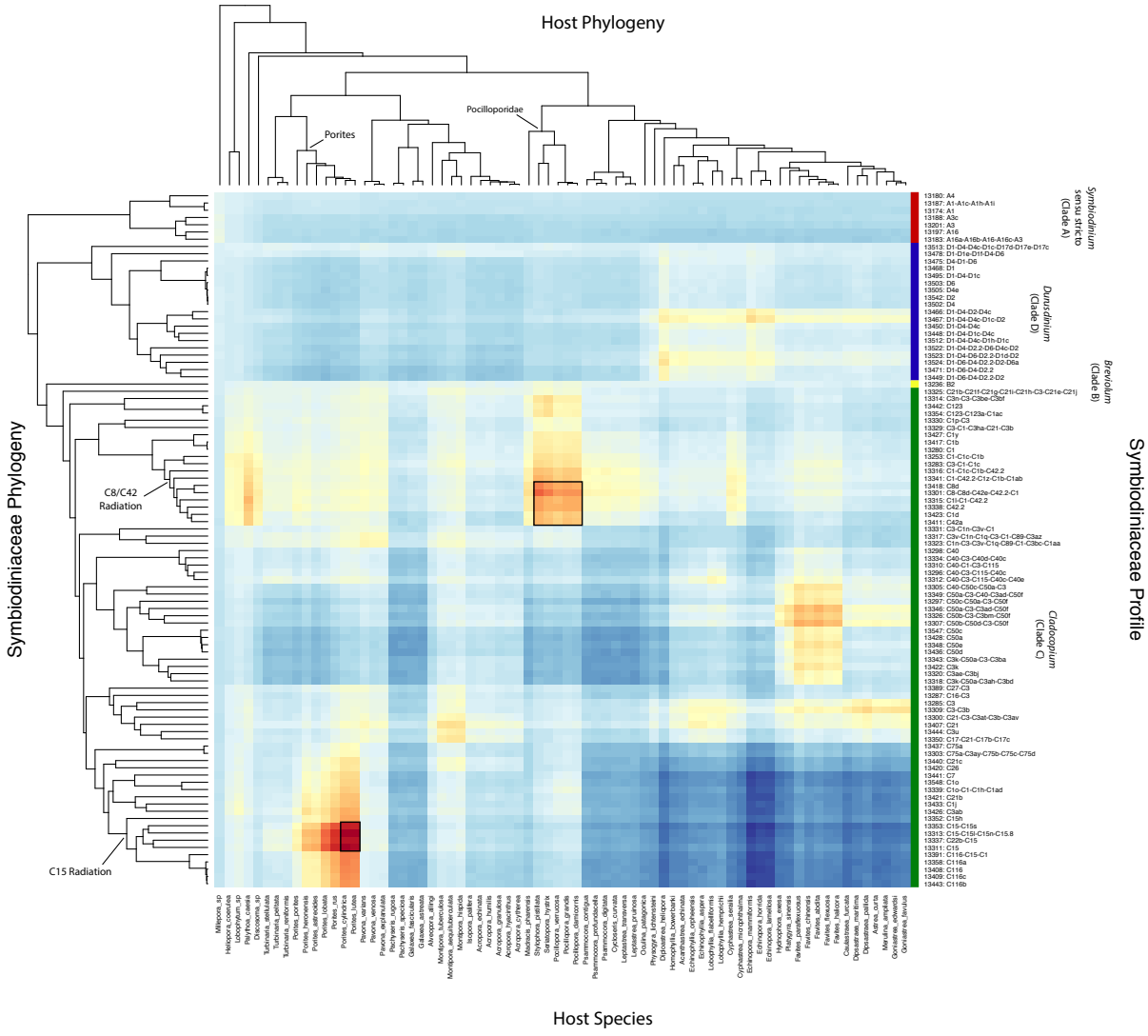


Figure 3.4. Variation in evolutionary rates.

The combined rate of evolution for each host-microbe pair is the product of the mean rate and the rate multipliers of all ancestral nodes. Plotted is the mean of all draws of the sum of all log-scale rate multipliers for the tips of each tree. Red squares correspond to higher rates of modern evolution, and blue to lower rates. Black boxes mark regions of the interacting phylogenies with significantly increased rates of historical evolution of host specificity compared to the grand mean. The two regions correspond to the evolution of the C15 radiation in *Porites* corals, and the evolution of the C8/C42 radiation in the Pocilloporidae. The host phylogeny plotted is the first random draw used in the analysis;

the branch lengths of both phylogenies are the mean posterior chronological branch lengths estimated during model fitting.

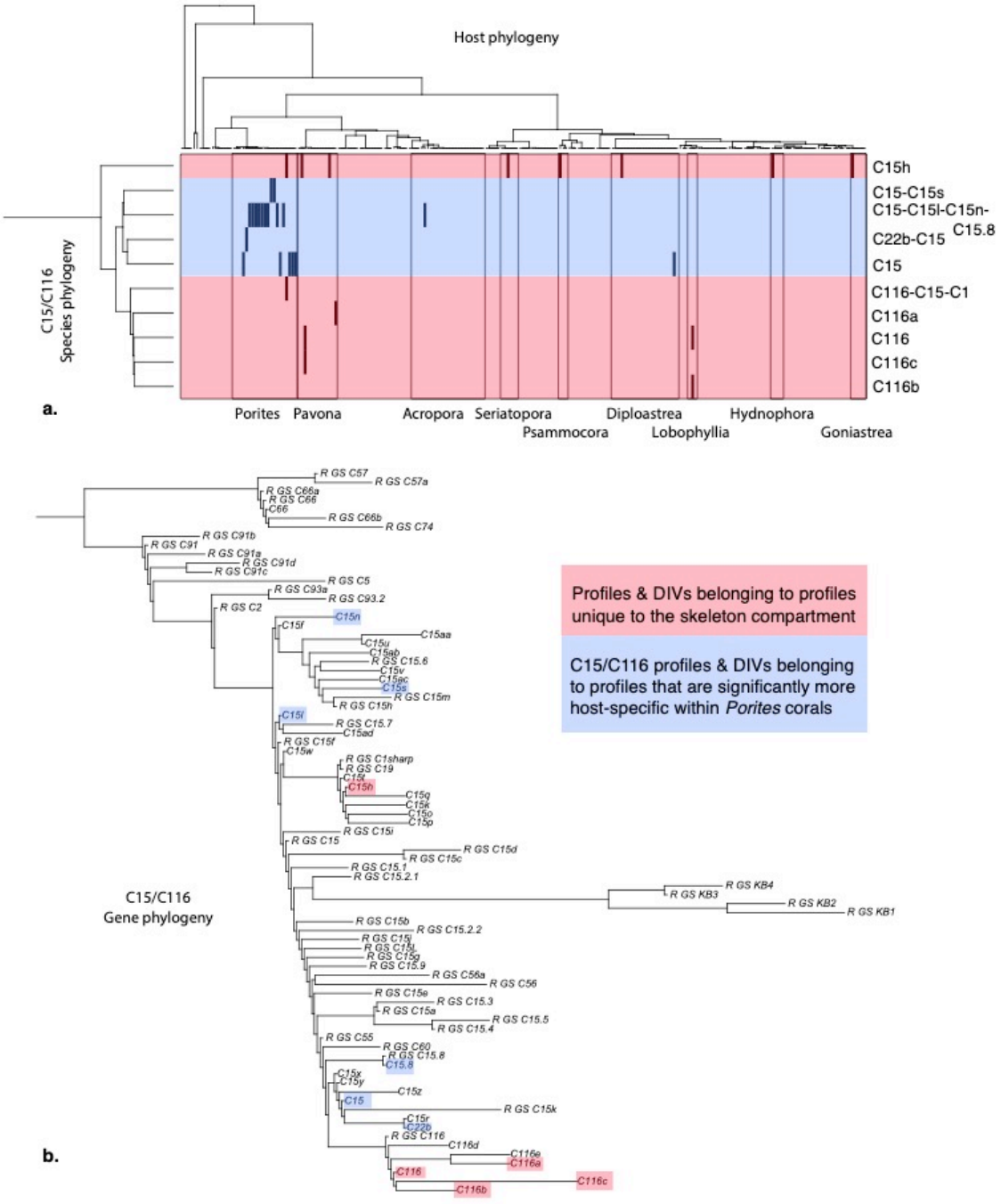


Figure 3.5. Phylogenetic placement of distinct members of the C15 radiation.

(a) C15/C116 profile distribution among coral skeletal samples. Profiles highlighted in red were not found in any tissue or mucus samples, while profiles in blue were found to have a higher degree of host specificity within *Porites* corals. The rough 'species tree' of Symbiodiniaceae profiles generated in this study suggests that C15 types that are highly specific to *Porites* corals are nested within a clade with skeleton-specific associations. The skeleton-specific profiles were found in a somewhat diverse assortment of host species (*Porites* and *Pavona* in Complex clade corals, and *Lobophyllia*, *Diploastrea*, *Seriatopora*, *Psammocora*, *Hydnophora*, and *Goniastrea* in Robust clade corals). However, the discriminating intergenomic sequence variants that define the profiles are less clearly assorted (b).

	Skeleton	Tissue	Mucus	Whole	Total samples
Samples with C116 / C15h	11	0	0	0	11
Samples without C116 / C15h	212	204	164	7	587
Total samples	223	204	164	7	598

Table 3.1. *Cladocopium* C116 and C15h are associated with coral skeletal samples, but not tissue or mucus samples.

References

Arif, C. et al. Assessing *Symbiodinium* diversity in scleractinian corals via next-generation sequencing-based genotyping of the ITS2 rDNA region. *Mol. Ecol*. 23, 4418–4433 (2014).

Aronesty, E. Comparison of Sequencing Utility Programs. *The Open Bioinformatics Journal* 7, 1–8 (2013).

Barneah, O. et al. Three party symbiosis: acoelomorph worms, corals and unicellular algal symbionts in Eilat (Red Sea). *Mar. Biol*. 151, 1215–1223 (2007).

Björk, J. R., Hui, F. K. C., O'Hara, R. B. & Montoya, J. M. Uncovering the Drivers of Animal-Host Microbiotas with Joint Distribution Modeling. *bioRxiv* 137943 (2017). doi:10.1101/137943

Bouckaert, R. R. et al. BEAST 2: A Software Platform for Bayesian Evolutionary

Analysis. *PLoS Comput. Biol.* 10, e1003537 (2014).
Bouckaert, R. R. & Drummond, A. J. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evol. Biol*. 17, 42 (2017).
Brucker, R. M. & Bordenstein, S. R. The capacious hologenome. *Zoology* **116,** 260–261 (2013).
Butler, M. A. & King, A. A. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *Am. Nat*. 164, 683–695 (2004).
Douglas, A. E. & Werren, J. H. Holes in the Hologenome: Why Host-Microbe Symbioses Are Not Holobionts. *mBio* 7, e02099–15 (2016).
Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol*. 4, e88 (2006).
Drummond, A. J. & Suchard, M. A. Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* 8, 114 (2010).
Eren, A. M. et al. Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* 9, 968–979 (2014).
Fabina, N. S., Putnam, H. M., Franklin, E. C., Stat, M. & Gates, R. D. Transmission Mode Predicts Specificity and Interaction Patterns in Coral-*Symbiodinium* Networks. *PLoS ONE* 7, e44970 (2012).
Fay, S. A. & Weber, M. X. The Occurrence of Mixed Infections of *Symbiodinium* (Dinoflagellata) within Individual Hosts. *J Phycol* 48, 1306–1316 (2012).
Fukami, H. *et al.* Mitochondrial and nuclear genes suggest that stony corals are monophyletic but most families of stony corals are not (Order Scleractinia, Class Anthozoa, Phylum Cnidaria). *PLOS ONE* **3,** e3222 (2008).
Gaulke, C. A. *et al.* Ecophylogenetics Clarifies the Evolutionary Association between Mammals and Their Gut Microbiota. *mBio* **9,** e01348–18 (2018).
Gates, R. D. & Ainsworth, T. D. The nature and taxonomic composition of coral symbiomes as drivers of performance limits in scleractinian corals. *J. Exp. Mar. Biol*. Ecol. 408, 94–101 (2011).
Gong, S. et al. Flexible Symbiotic Associations of *Symbiodinium* With Five Typical Coral Species in Tropical and Subtropical Reef Regions of the Northern South China Sea. *Front*. *Microbiol*. 9, 4418 (2018).
Goodrich B, Gabry J, Ali I & Brilleman S. rstanarm: Bayesian applied regression modeling via Stan. R package version 2.17.4. (2018).
Grafen, A. The uniqueness of the phylogenetic regression. *J. Theor. Biol.* **156,** 405–423 (1992).
Gronau, Q. F. bridgesampling: An R Package for Estimating Normalizing Constants. *arXiv* (2017).
Groussin, M. et al. Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. *Nat. Commun*. 8, 14319 (2017).
Hadfield, J. D. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software* **33,** 1–22 (2010).
Hadfield, J. D., Krasnov, B. R., Poulin, R. & Nakagawa, S. A tale of two phylogenies: comparative analyses of ecological interactions. Am. Nat. 183, 174–187 (2014).
Huang, D. & Roy, K. The future of evolutionary diversity in reef corals. *Phil. Trans. R. Soc. B* 370, 20140010–20140010 (2015).

Ives, A. R. & Godfray, H. C. J. Phylogenetic Analysis of Trophic Associations. *Am. Nat.* 168, E1–E14 (2006).
Jhwueng, D.-C. & Maroulas, V. Adaptive Trait Evolution in Random Environment. *arXiv.org stat.AP*, (2015).
Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* 30, 772–780 (2013).
Krueger, T. & Gates, R. D. Cultivating endosymbionts — Host environmental mimics support the survival of *Symbiodinium* C15 *ex hospite*. *J. Exp. Mar. Biol. Ecol*. 413, 169–176 (2012).
Kruuk, L. E. B. & Hadfield, J. D. How to separate genetic and environmental causes of similarity between relatives. *J. Evol. Biol.* **20,** 1890–1903 (2007).
LaJeunesse, T. C. et al. Systematic Revision of Symbiodiniaceae Highlights the Antiquity and Diversity of Coral Endosymbionts. CURBIO 0 (2018).
Lesser, M. P., Stat, M. & Gates, R. D. The endosymbiotic dinoflagellates (*Symbiodinium* sp.) of corals are parasites and mutualists. *Coral Reefs* 32, 603–611 (2013).
Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71,** 8228–8235 (2005).
Mazel, F. *et al.* Is Host Filtering the Main Driver of Phylosymbiosis across the Tree of Life? *mSystems* **3,** e00097–18 (2018).
Monnahan, C. C., Thorson, J. T. & Branch, T. A. Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution* 8, 339–348 (2017).
Moran, N. A. & Sloan, D. B. The Hologenome Concept: Helpful or Hollow? *PLoS Biol*. 13, e1002311 (2015).
Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20, 289–290 (2004).
Plummer, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. in *Proceedings of the International Workshop on Distributed Statistical Computing* (2003).
Pochon, X., Pawlowski, J., Zaninetti, L. & Rowan, R. High genetic diversity and relative specificity among *Symbiodinium*-like endosymbiotic dinoflagellates in soritid foraminiferans. Mar. Biol. 139, 1069–1078 (2001).
Pollock, F. J. et al. Coral-associated bacteria demonstrate phylosymbiosis and cophylogeny. *Nat. Commun*. 9, 4921 (2018).
Pratlong, M., Rancurel, C. & Pontarotti, P. Monophyly of Anthozoa (Cnidaria): why do nuclear and mitochondrial phylogenies disagree? *Zool. Scripta* (2016). doi:10.1111/zsc.12208
Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS ONE 5*, e9490 (2010).
Revell, L. J. Size-correction and principal components for interspecific comparative studies. *Evolution* **63,** 3258–3268 (2009).
Revell, L. J. Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution* **1,** 319–329 (2010).
Revell, L. J., Mahler, D. L., Peres Neto, P. R. & Redelings, B. D. A new phylogenetic method for identifying exceptional phenotypic diversification. *Evolution* **66,** 135–146 (2012).

Ronquist, F. et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol*. 61, 539–542 (2012).
Sampayo, E. M., Dove, S. & LaJeunesse, T. C. Cohesive molecular genetic data delineate species diversity in the dinoflagellate genus *Symbiodinium*. *Mol. Ecol*. 18, 500–519 (2009).
Santos, S. R., Shearer, T. L., Hannes, A. R. & Coffroth, M. A. Fine-scale diversity and specificity in the most prevalent lineage of symbiotic dinoflagellates (*Symbiodinium*, Dinophyceae) of the Caribbean. *Mol. Ecol.* 13, 459–469 (2004).
Schloss, P. D. et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl. Environ. Microbiol*. 75, 7537–7541 (2009).
Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PLoS ONE* 6, e27310 (2011).
Silverstein, R. N., Correa, A. M. S. & Baker, A. C. Specificity is rarely absolute in coral-algal symbiosis: implications for coral response to climate change. *Proc. R. Soc. B* 279, 2609–2618 (2012).
Stat, M., Carter, D. A. & Hoegh-Guldberg, O. The evolutionary history of *Symbiodinium* and scleractinian hosts—Symbiosis, diversity, and the effect of climate change. *Perspect. Plant Ecol*. 8, 23–43 (2006).
Stat, M., Morris, E. & Gates, R. D. Functional diversity in coral-dinoflagellate symbiosis. *Proc. Natl. Acad. Sci. U.S.A.* 105, 9256–9261 (2008).
Stat, M., Pochon, X., Cowie, R. O. M. & Gates, R. D. Specificity in communities of Symbiodinium in corals from Johnston Atoll. *Mar. Ecol. Prog. Ser*. 386, 83–96 (2009).
Venera-Ponton, D. E., Diaz-Pulido, G., Rodriguez-Lanetty, M. & Hoegh-Guldberg, O. Presence of *Symbiodinium* spp. in macroalgal microhabitats from the southern Great Barrier Reef. *Coral Reefs* 29, 1049–1060 (2010).
Zaneveld, J. R. R. et al. Overfishing and nutrient pollution interact with temperature to disrupt coral reefs down to microbial scales. *Nat. Commun*. 7, 11833 (2016).

General Conclusion

The importance of microbes to the health of multicellular organisms has been recognized since they were first linked to disease as many as 500 years ago (Fracastoro 1961), and identified as the source of nitrogen fixation in legumes more than 100 years ago (Beijerinck 1961). However, only recently have technological developments begun to shed light on the ubiquity of host-microbe interactions, and the subtler ways in which the vast majority of the microbes in the environment affect the health of their macrobial counterparts. Most interactions between hosts and microbes are far less intimate than the textbook examples of pathogens and symbionts, but dynamic microbial communities, or microbiomes, are still hypothesized to provide a variety of services to their host, including processing of metabolites and defense against predators and pathogens (Zaneveld *et al.* 2016, Knowlton & Rohwer 2003). Even microbes that are mostly commensal in terms of their immediate effects on health are likely to influence the evolution of their hosts, as their natural diversity and ubiquitous presence create both risks and opportunities that constrain or enhance their hosts' evolutionary potential (Lesser *et al.* 2007). These dynamics are similar to those of the ecology of macrobial communities, as host-associated microbiomes are composed of a wide variety of organisms with diverse levels of specificity, interdependence, and coadaptation.

However, the processes that shape host-associated microbiomes also share characteristics with those that shape genomes. In part this is because the transmission of microbiome members may be more strongly constrained by vertical inheritance than most communities of plants and animals. But because hosts act as ecosystem engineers and are themselves subject to gradual evolutionary change, horizontally acquired members of the microbiome are more likely to successfully colonize the ecosystems of hosts that are closely related to those from which they came (Mazel *et al.* 2018). As a result of both processes (barriers to dispersal and environmental filtering), the microbiomes of closely related hosts tend to have similar compositions; or, in other words, their microbiomes display *phylogenetic signal*. Although there is a rich history of studying the phylogenetic signal of traits (e.g. see Zheng *et al.* 2009), the term *phylosymbiosis* has been coined to describe this pattern in microbiomes, recognizing that a host's interactions with microbial

communities straddles a blurry boundary between phylogenomics and ecology (Brucker & Bordenstein 2013).

There is an ongoing debate about the relatively newly coined terms *holobiont*, *hologenome*, and *phylosymbiosis*. These terms all encompass a growing appreciation for the ubiquity of intimate associations between multicellular organisms and the communities of microbes that live with them (Margulis 1993, Rohwer *et al.* 2002, Gilbert *et al.* 2012). The first two terms refer respectively to the combination of a host and its microbiome, and to the entirety of the genomes of all members of the holobiont. The third refers to the pattern where the composition of the microbiome reflects the phylogeny of its host (Brucker & Bordenstein 2014). The utility of these concepts has been questioned, in part because their use seems to encourage misconceptions about the degree of reciprocal influence each member of the community has on the fitness and evolution of the others (Moran & Sloan 2015, Douglas & Werner 2016). In fact, these patterns do not require any influence of the microbiome on the fitness or evolution of their host, and are consistent with both neutral (physical or geographic barriers to dispersal) and selective (environmental filtering) effects of hosts on the fitness of their microbes (Mazel *et al.* 2018). In reality, there is a spectrum of interactions among macrobial and microbial organisms, and the concept of a holobiont can be seen as an attempt to discretize this spectrum into categories: does a microbe interact meaningfully with a host, or does it not? Whether such a discretization is warranted may depend on whether the spectrum shows a gradual range of host association strength, or if there is a steep drop-off in interaction strength where a boundary could be drawn between ‘meaningful’ and ‘non-meaningful’ microbial interactions.

In my dissertation, I first established that phylosymbiotic patterns exist in corals, discovered the many caveats of interpretation of these patterns, and developed a new statistical method to help get around some of those caveats. I provide a framework that can combine many separate processes into a single holistic model. Thus many separate analyses, such as of beta-diversity (e.g. with PERMANOVA using separate distance matrices for ‘phylogenetic’ and ‘non-phylogenetic’ effects), alpha-diversity (e.g. through independent regressions of metrics such as Faith’s Phylogenetic Diversity, total richness,

and total evenness), and separate univariate regressions of microbial taxa at numerous scales, can be reduced to a single analysis with many forms of summary; each of which controls for the effects of the others. This framework provides clearer and more precise interpretations for each effect, and can help tease apart the many processes that contribute to shaping the microbiome.

Bibliography

Barnosky, A. D. et al. Has the Earth's sixth mass extinction already arrived? Nature 2011 471:7336 471, 51–57 (2011).

Beijerinck, M. W. in Milestones in Microbiology (ed. Brock, T. D.). 200–204 (Prentice-Hall International, Inc., London, 1961).

Bowen, B. W., Rocha, L. A., Toonen, R. J. & Karl, S. A. The origins of tropical marine biodiversity. Trends Ecol. Evol. (2013). doi:10.1016/j.tree.2013.01.018

Brooks, A. W., Kohl, K. D., Brucker, R. M., van Opstal, E. J. & Bordenstein, S. R. Phylosymbiosis: Relationships and Functional Effects of Microbial Communities across Host Evolutionary History. PLoS Biol. 14, e2000225 (2016).

Brucker, R. M. & Bordenstein, S. R. The hologenomic basis of speciation: gut bacteria cause hybrid lethality in the genus Nasonia. Science 341, 667–669 (2013).

Darwin, C. On the Origin of Species by Means of Natural Selection (Murray, London). 1859.

Darwin, C. The structure and distribution of coral reefs. (1889).

Douglas, A. E. & Werren, J. H. Holes in the Hologenome: Why Host-Microbe Symbioses Are Not Holobionts. mBio 7, e02099–15 (2016).

Fracastoro, G. in Milestones in Microbiology (ed. Brock, T. D.). 83–89 (Prentice-Hall International, Inc., London, 1961).

Fukami, H. Short review: molecular phylogenetic analyses of reef corals. Galaxea, Journal of Coral Reef Studies 10, 47–55 (2008).

Gilbert, S. F., Sapp, J. & Tauber, A. I. A Symbiotic View of Life: We Have Never Been Individuals. Q. Rev. Biol. 87, 325–341 (2012).

Grigg, R. W. Darwin Point: A threshold for atoll formation. Coral Reefs 1, 29–34 (1982).

Groussin, M. et al. Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. Nat. Commun. 8, 14319 (2017).

Hadfield, J. D., Krasnov, B. R., Poulin, R. & Nakagawa, S. A tale of two phylogenies: comparative analyses of ecological interactions. Am. Nat. 183, 174–187 (2014).

Hafner, M. S. & Nadler, S. A. Phylogenetic trees support the coevolution of parasites and their hosts. 332, 258–259 (1988).

Huang, D. & Roy, K. The future of evolutionary diversity in reef corals. Phil. Trans. R. Soc. B 370, 20140010–20140010 (2015).

Hughes, T. P. et al. Global warming and recurrent mass bleaching of corals. 543, 373–377 (2017).

Hughes, T. P. et al. Coral reefs in the Anthropocene. 546, 82–90 (2017).

Hughes, T. P., Kerry, J. T. & Simpson, T. Large-scale bleaching of corals on the Great Barrier Reef. Ecology 99, 501–501 (2018).

Ives, A. R. & Godfray, H. C. J. Phylogenetic Analysis of Trophic Associations. Am. Nat. 168, E1–E14 (2006).

Knowlton, N. & Rohwer, F. L. Multispecies Microbial Mutualisms on Coral Reefs: The Host as a Habitat. Am. Nat. 162, S51–S62 (2003).

Lesser, M. P., Bythell, J. C., Gates, R. D., Johnstone, R. W. & Hoegh-Guldberg, O. Are infectious diseases really killing corals? Alternative interpretations of the experimental and ecological data. J. Exp. Mar. Biol. Ecol. 346, 36–44 (2007).

Margulis, L. Origins of species: acquired genomes and individuality. Biosystems 31,

121–125 (1993).
Mazel, F. et al. Is Host Filtering the Main Driver of Phylosymbiosis across the Tree of Life? mSystems 3, e00097–18 (2018).
Medzhitov, R. Recognition of microorganisms and activation of the immune response. (2007).
Moran, N. A. & Sloan, D. B. The Hologenome Concept: Helpful or Hollow? PLoS Biol. 13, e1002311 (2015).
Pandolfi, J. M. et al. Global Trajectories of the Long-Term Decline of Coral Reef Ecosystems. Science 301, 955–958 (2003).
Paton, A. J. et al. Towards Target 1 of the Global Strategy for Plant Conservation: a working list of all known plant species - progress and prospects. Taxon 57, 602–611 (2008).
Pollock, F. J. et al. Coral-associated bacteria demonstrate phylosymbiosis and cophylogeny. Nat. Commun. 9, 4921 (2018).
Rohwer, F. L., Seguritan, V., Azam, F. & Knowlton, N. Diversity and distribution of coral-associated bacteria. Mar. Ecol. Prog. Ser. 243, (2002).
Van Oppen, M. J. H. & Gates, R. D. Conservation genetics and the resilience of reef-building corals. Mol. Ecol. 15, 3863–3883 (2006).
Werner, G. D., Cornwell, W. K., Sprent, J. I. & Kattge, J. A single evolutionary innovation drives the deep evolution of symbiotic N2-fixation in angiosperms. Nat. Commun. 5, (2014).
Wertheim, J. O., Sanderson, M. J., Worobey, M. & Bjork, A. Relaxed Molecular Clocks, the Bias–Variance Trade-off, and the Quality of Phylogenetic Inference. *Syst. Biol.* 59, 1–8 (2010).
Zanne, A. E. et al. Three keys to the radiation of angiosperms into freezing environments. 506, 89–92 (2013).
Zaneveld, J. R. R. et al. Overfishing and nutrient pollution interact with temperature to disrupt coral reefs down to microbial scales. Nat. Commun. 7, 11833 (2016).
Zaneveld, J. R. R., McMinds, R. & Vega Thurber, R. L. Stress and stability: applying the Anna Karenina principle to animal microbiomes. Nat. Microbiol 2, 17121 (2017).
Zheng, L. et al. New multivariate tests for phylogenetic signal and trait correlations applied to ecophysiological phenotypes of nine Manglietia species. Funct Ecol 23, 1059–1069 (2009).

Appendices

Supplementary text for chapter 2: Coral-associated bacteria demonstrate phylosymbiosis and cophylogeny

Supplementary Note 1

Sequencing depths Following quality control, removal of chimeras and singleton sequences, and exclusion of environmental and outgroup libraries from the main analysis (except where otherwise noted), 649 scleractinian coral 16S rRNA gene amplicon libraries remained. Of these, 30 samples had fewer than 1000 sequences/sample, 80 samples had 1000-5000 sequences/sample; 112 samples had 5000-10000 sequences; 141 samples had 10000-15000 sequences; 294 samples had 15000-20000 sequences; 131 samples had >20000 sequences. To prevent unequal sequencing depths from influencing the analysis, we rarified to 1000/sequences per sample, which gave 619 total coral mucus, tissue, and skeleton samples. However, we also tested 5000, 10000, 15000, or 20000 sequences per sample for alpha-diversity and core-microbiome analysis. For differential abundance testing, we used 1000 sequences/sample rarefaction, or a parametric method within the phylogenetic GLMMs to account for sequencing depth variation.

Supplementary Note 2

A brief overview: dominant phyla in the coral microbiome Before comparing differences between compartments at more detailed taxonomic levels, we summarized the proportional representation of bacterial and archaeal phyla in the coral microbiome, averaged across all samples. At the coarsest scale, Proteobacteria were by far the most abundant bacterial or archaeal phylum in the dataset (56.5% of total reads), ~4-fold more abundant than the next most abundant phylum (Bacteroidetes, 14.2%), and ~45-fold more abundant than the most abundant archaeal phylum (Crenarchaeota, 1.3%). Like many host-associated microbial communities, coral microbiomes were highly uneven: 10 phyla accounted for 90% of the microbiome, averaging across all samples (Proteobacteria, 56.5%; Bacteroidetes, 14.2%; Firmicutes, 5.7%; Cyanobacteria, 4.4%; Actinobacteria, 3.5%; Planctomycetes, 1.7%; Crenarchaeota, 1.3%; Chlamydiae, 1.1%; Chloroflexi, 1.1%; Verrucomicrobia, 0.8%), with 59 other phyla, each at <1% mean abundance, making up the remainder. Even after removing identifiable mitochondria and

chloroplasts, an average of ~4.9% of reads in each sample could not be assigned to a phylum.

Supplementary Note 3

Microbiome richness differs between coral mucus, tissue, and skeleton Microbiomes from distinct anatomical compartments differed significantly in OTU richness (i.e., observed OTUs per 1000 reads) ($df = 2$, F-value = 19.55, $p < 0.001$; Fig. 1a). The coral mucus layer hosted significantly lower richness (119.6 ± 6.8l OTUs per 1000 reads) than coral tissue (166.5 ± 11.1; $p = 0.0015$) and skeleton (208.7 ± 11.9; $p < 0.001$) (Fig. 1a). OTU richness of endolithic communities within coral skeletons was 75% higher than coral mucus ($p < 0.001$) and 25% higher than tissue ($p = 0.010$) (Fig. 1a). These differences were also significant at rarefaction depths of 5000, 10000, or 15000. At 20000 sequences per sample, all compartment differences were significant except mucus vs. tissue, likely due to removal of most tissue samples at this depth (n = 17 vs. n = 199 at 1000 sequences per sample).

Supplementary Note 4

Compositional differences between compartments are robust to choice of distance metric As expected, given the complexity of underlying variation in coral phylogeny, environment, and anatomy, the first 3 PC axes captured only a small fraction of overall variation in community distances (PC1 8.9%; PC2 4.3%; PC3 2.1%). However, statistical analyses of all multivariate dissimilarity measures examined (Bray-Curtis, Weighted UniFrac and Unweighted UniFrac) revealed significant differences in microbial community composition among coral compartments (all Adonis $p < 0.001$, Fig. 1b). Microbial communities from different compartments separated along the second principal coordinates axis, with mucus samples clustering towards the positive end of this axis, transitioning to tissue and then to skeleton microbiomes near the negative end of the axis (Fig. 1b).

Supplementary Note 5

Coral mucus, tissue, and skeleton host distinct core microbiomes 16S rRNA V4 amplicons sequenced from scleractinian coral mucus, tissue and skeleton (n = 614) were

evaluated for differences in 'core' microbiome membership, community composition and diversity, and the influence of environmental and host conditions on these parameters.

Core microbiome membership was assessed at multiple taxonomic levels by identifying taxa that were present in more than 70% of rarefied samples within a given compartment. At the OTU level, no specific bacterial or archaeal OTUs were core members of any compartment at 1000 reads per sample, with the exception of a single OTU of water-column *Synechococcus* consistently associated with mucus microbiomes. Note that at 1000 reads per sample, a relatively rare microbe with 0.5% abundance will be sequenced in 2 or more reads ~95.9% of the time (cumulative binomial, p=0.005, n=2, trials=1000). Microbes that are biologically part of the core microbiome, but not observed here are thus likely to have low abundances. This shallow depth, which includes the greatest number of biological samples, was therefore deemed sufficient for most of our analyses (which interrogate the ecology and evolution of the main lineages of bacteria and archaea in each compartment).

We also calculated core microbiomes at higher read depths. At 5000 seqs/sample, a microbe with an abundance of 0.05% will be detected with 2 reads in ~71.3% of samples (cumulative binomial p=0.0005, n=2, trials=5000). At this depth, results were identical as at 1000 seqs/sample, except that *Endozoicomonas* OTU 739464 was consistently associated with tissue microbiomes.To test whether broader taxonomic ranks were conserved, we also calculated which bacterial orders were prevalent across all samples with 1000 reads or more. All compartments shared 6 core microbiome members: Bacteroidetes from the order Flavobacteriales*;* unclassified α-Proteobacteria and α-Proteobacteria from orders Rhizobiales and Rhodobacterales; and γ-Proteobacteria from the orders Oceanospirallales and Alteromonadales*.* In addition, compartments had unique core microbial orders: the Cyanobacteria order Synechococcales and the α-Proteobacteria order Rickettsiales (both common in coastal waters) were core in mucus; Clostridiales and Rhodospirillales were core in skeleton; and Cytophagales were core in both tissue and skeleton.

Together, these results indicate a pattern of conservation in the orders of bacteria present in coral microbiomes, but variability at finer taxonomic scales. While the core

microbiome (i.e. OTUs with >70% prevalence) of coral mucus and tissue held only *Synechococcus* in mucus and *Endozoicomonas* OTU 739464 in tissue, several bacterial orders were consistently present in coral mucus, tissue, and skeleton, even across very diverse coral hosts spanning large geographic ranges. Greater consistency of microbial orders than OTUs could be due to partial niche overlap between OTUs within the same order, or may reflect co-diversification of some strains of bacteria with their host.

Supplementary Note 6

Overlapping prevalent microbial orders in Australian and Florida Keys Coral Mucus We tested whether the microbial orders identified as highly consistent in this study were consistent with past results. In a time-series study of the mucus microbiome of corals exposed to simulated overfishing or nutrient pollution, Zaneveld *et al.* calculated core orders in the coral mucus of three genera of corals in plots in the Florida Keys[2]. 100% of the 8 orders of bacteria found in >70% of coral mucus samples in this study were also found to be prevalent (>95% of samples) in coral mucus from Zaneveld et al., 2016. (The converse was not true: only 8/11 orders with >95% prevalence in control corals from Zaneveld et al., 2016 were highly prevalent here). Similarly, OTUs belonging to 7/8 of these named orders (Synechococcales, Oceanospirallales, Alteromonadales, Rickettsiales, Rhodobacterales, Rhizobiales, and Flavobacterales, but not unclassified α-Proteobacteria) were part of the core microbiome in Apprill et al., 2017[1].

Supplementary Note 7

Confirmation of *Candidatus* Amoebophilus as a common coral associate

Apprill *et al.*,[1] previously identified *Candidatus* Amoebophilus as a core member of the microbiome of three coral species. This finding was intriguing because *Candidatus* Amoebophilus are intracellular symbionts of microbial eukaryotes, so we decided to look for it in our data. Using the same definition of 'consistent' association (presence in >50% of samples) and when rarefying at an equal read depth (10,000 sequences/sample) we also find *Candidatus* Amoeobophilus (specifically Greengenes OTU 321533 with taxonomy string 'k__Bacteria; p__Bacteroidetes; c__Cytophagia; o__Cytophagales;

f__[Amoebophilaceae]; g__SGUS912;s__') to be one of 11 OTUs present in the 50% tissue 'core microbiome' at 10000 sequences/sample.

Supplementary Note 8

Coral compartments differ in consistency. We calculated two measures of the consistency of coral compartments: the proportion of reads that were part of 'core' microbial orders (present in 70% of samples), and inter-colony variation in beta-diversity. The proportion of reads belonging to core microbial orders differed significantly among compartments ($df = 2$, F-value = 27.2, $p < 0.001$; Supplementary Figure 2a). The microbiome of coral mucus, which is in closest proximity to surrounding seawater, had higher core microbiome abundance than tissue or skeleton. Microbes belonging to core orders accounted for 64.5 ± 1.7% of reads within mucus microbiomes, 30% more than in tissue (49.7 ± 2.0%; $p < 0.0001$) and 38% more than in skeleton (46.7 ± 1.8%; $p < 0.0001$; Supplementary Figure 2a). Similarly, overall community composition was least variable in mucus across samples (Weighted UniFrac distance, $p < 0.001$; Supplementary Figure 2b). Thus, on both measures, mucus microbiomes were more consistent in composition than tissue or skeleton microbiomes. This may reflect that, despite some species differences, variation in mucus microbiomes between coral species is lower than tissue or skeleton microbiomes.

Supplementary Note 9

In healthy corals, host species has a stronger influence on the microbiome than geography. To assess the relative influence of host and environmental factors on coral microbiomes, the impacts of host (e.g. host genus and disease susceptibility) and environmental (e.g. collection season, reef, and latitude) variables were quantified for mucus (n = 207), tissue (n = 199), and skeleton (n = 208) microbiomes (OTU level) across scleractinian coral genera. Several host and environmental factors significantly influenced microbial community composition in each of the three coral compartments (all Weighted UniFrac Adonis $p < 0.05$ in all compartments; Fig. 1c). Across all compartments, host species had the greatest influence on microbial community composition (raw R^2 0.37-0.48), and in tissue explained nearly half the variance in community composition. Host species was still the most influential variable after

adjusting R^2 values for the degrees of freedom in each variable (adjusted R^2 0.15-0.24). In the context of this study, sampling location (reef name) explained less variance than host species (raw R^2 0.16-0.20) and had consistently lower adjusted R^2 values than host species. However, the relative influence of these and other measured parameters differed among compartments.

Supplementary Note 10

Across all compartments and multivariate dissimilarity measures, more specific taxonomic ranks explain more microbiome clustering. To test the explanatory power of various levels of coral taxonomy, we ranked the Adonis adjusted R^2 value of coral species, genus, clade *sensu* Fukami (family-level group), and complex vs. robust clade membership (broadest division). A table of these values is reported in Supplementary Data 3. The absolute R^2 values differed across compartments, metrics, and rarefaction depths (e.g. higher R^2 values for host factors in tissue). However, their relative rank was remarkably consistent. Across all compartments, three dissimilarity measures (Weighted UniFrac distance, Unweighted UniFrac distance, and Bray Curtis dissimilarity) and two rarefaction depths (1000 sequences/sample or 10000 sequences/sample), more specific taxonomic levels always explained microbiome beta-diversity better than more general ones. The largest fall in explanatory power typically occurred between genus and family-level group, and between family-level group and Complex vs. Robust clade membership. This analysis is largely consistent with signals of phylosymbiosis in Mantel test results, and also shows that the relative effect of coral taxonomy on microbiome membership is robust to common choices for rarefaction depth and dissimilarity measure.

Supplementary Note 11

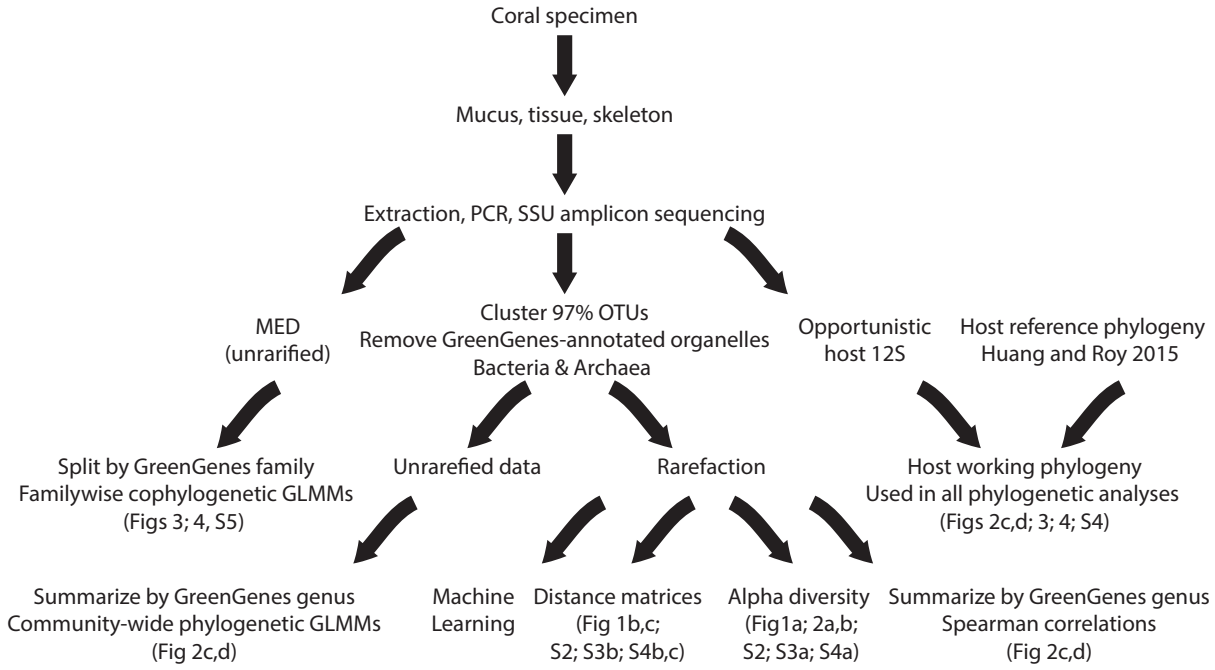
Host factors more strongly influence coral tissue and skeleton microbiomes while environmental conditions more strongly influence mucus communities. The influence of host factors on microbiome composition (i.e., Weighted UniFrac Adonis R^2) was most pronounced in the tissue and skeleton compartments, whereas environmental factors tended to have the strongest influence on mucus communities (Fig. 1c). Similarly, host genus had the strongest influence on tissue microbiomes. The influence of host genus on tissue microbiome composition was 1.53-fold greater than on mucus microbiome composition (mucus raw Adonis $R^2 = 0.249$; tissue raw $R^2 = 0.380$) and 1.14-fold greater in skeleton microbiomes (raw $R^2 = 0.334$; Adonis Bonferroni $p < 0.05$ in all compartments; Fig. 1c). Microbiome composition was also significantly correlated with disease susceptibility (i.e., 10-year genus-level disease prevalence on mid-shelf reefs of the northern Great Barrier Reef; Willis Great Barrier Reef Disease Database v20161016[3]), and like other host-associated factors species-wide disease prevalence was most strongly correlated with microbiome composition in tissue and skeleton (Adonis Bonferroni $p < 0.05$; Fig. 1c). Conversely, reef (i.e., collection site) also influenced coral mucus ($R^2 = 0.209$) 1.15-fold more strongly than coral tissue ($R^2 = 0.184$) and 1.28-fold more strongly than skeleton communities ($R^2 = 0.163$) (Weighted UniFrac Adonis Bonferroni $p < 0.05$ in all compartments; Fig. 1c). Latitude had a small but consistent effect on community composition. Latitude was significantly correlated with microbiome structure in all compartments (all Adonis $p < 0.05$; Fig. 1c), and again this environmental factor had the strongest effect on communities in the mucus. (mucus $R^2 = 0.045$; tissue $R^2 = 0.020$; skeleton $R^2 = 0.018$; Fig. 1c; Supplementary Data S3). Finally, we conducted a separate analysis of the effects of sampling season (i.e., summer vs. winter) using only the subset of samples collected at Lizard Island. Like other environmental parameters, sampling season influenced mucus microbial communities ($R^2 = 0.108$) 3.29-fold more strongly than tissue ($R^2 = 0.033$) and 2.78-fold more strongly than skeleton communities ($R^2 = 0.039$) (Adonis $p < 0.05$ in all compartments; data for the Lizard island subset used in this analysis not shown).

Together, these data show that for healthy Australian corals, host species tended to have a stronger influence than geography or measured environmental parameters in all compartments tested (i.e. including mucus). They also show that host traits tend to have a stronger relative influence on coral tissue and skeleton microbiomes relative to mucus, whereas environmental parameters have a stronger relative influence on mucus microbiomes compared to tissue or skeleton.

Supplementary Note 12

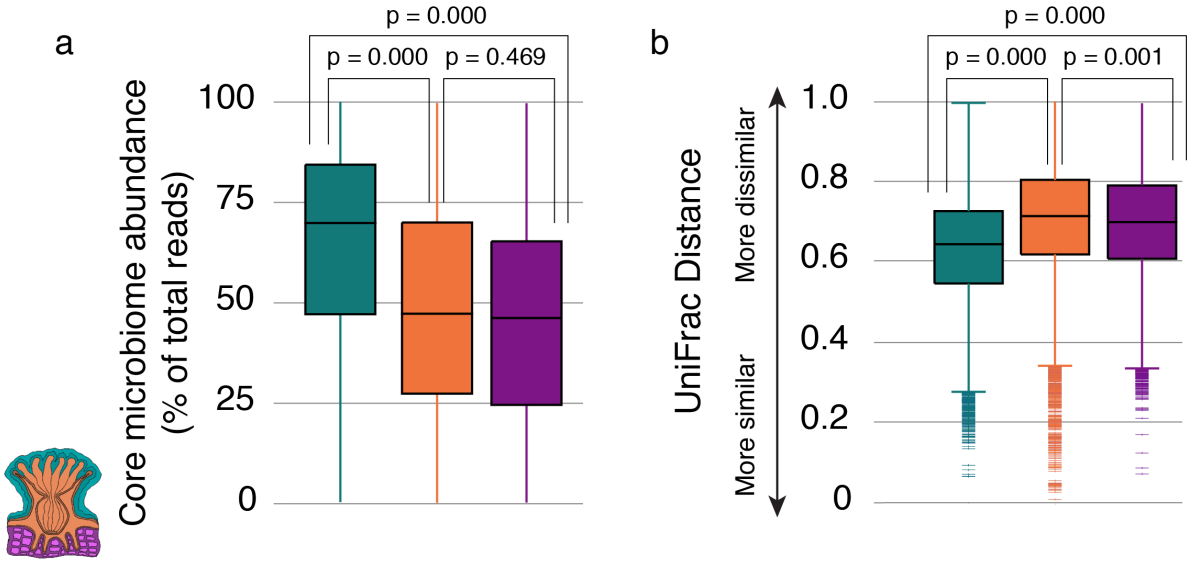
Putatively opportunistic bacteria associated with small corals We tested how relative coral size influenced the microbiome. Several measures of relative coral colony size showed significant correlations with microbiome composition. We tested both absolute coral dimensions as well as those sizes normalized by the largest dimensions recorded for that coral in this study (prop_Colony_maximum_GCMP_recorded) or in either this study or the coral traits database (prop_Colony_maximum_universal). Using either measure, tissue and skeleton microbiomes were significantly associated with coral size (raw Adonis $p < 0.05$). Following stringent Bonferroni correction across all factors and compartments, the association between coral colony size and microbiome composition was significant only in coral skeleton (Adonis permutational $p < 0.05$), though even there the magnitude of the effect was very small (raw $R^2 <= 0.022$). Phylogenetic GLMMs showed that 14.9% (51/343, mucus), 47.6% (214/450, tissue), and 31.2% (151/483, skeleton) of genera were significantly associated with smaller corals; while 7.3% (25/343, mucus), 2.7% (12/450, tissue), and 7.5% (36/483, skeleton) were significantly associated with larger corals. Additionally, pGLMMs showed that *Balneola* was significantly less abundant in large corals in both the tissue (upper 95% CI = -1.51) and skeleton (upper 95% CI = -0.22) microbial communities, and the abundance of unclassified Aurantimonadaceae was significantly lower in the skeletons of large colonies (upper 95% CI = -0.12), although its overall prevalence was too low in tissues to be tested.

Supplementary Figures



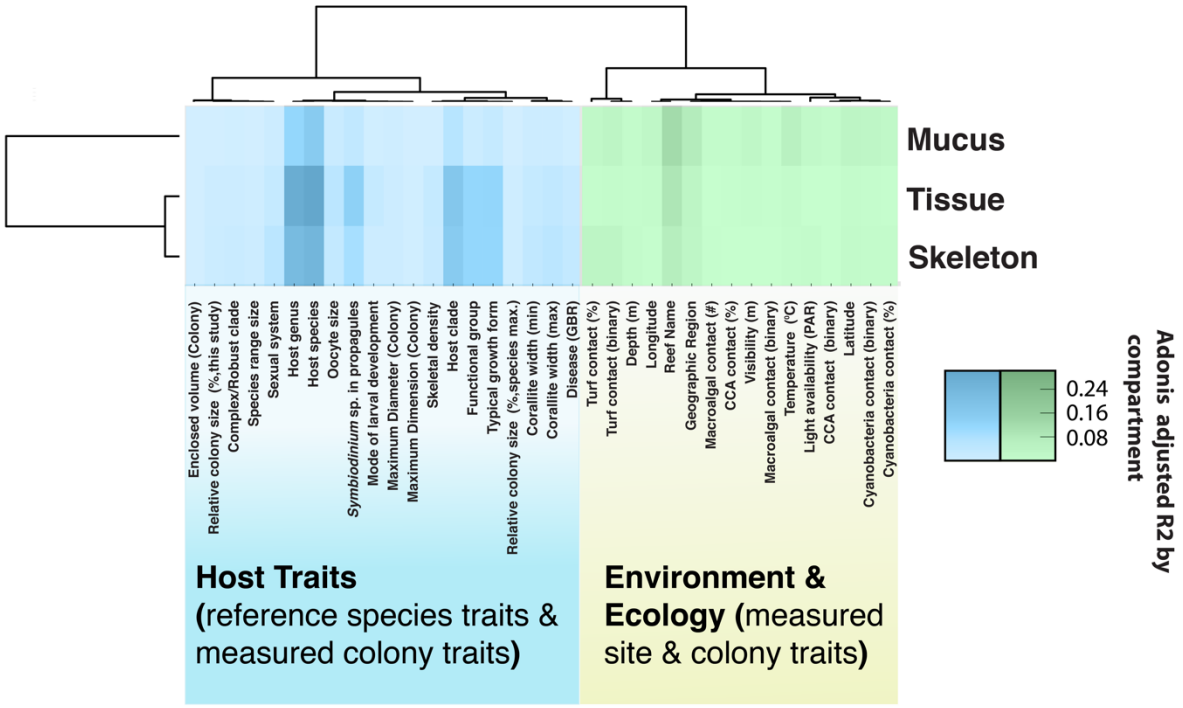
Supplementary Figure 1.1. Data processing workflow.

Several sampling procedures, bioinformatics, and analytical methods were combined and used to evaluate our microbiome and coral host phylogeny data. All samples underwent collections in the field where corals were subjected to mucus, tissue and skeleton separation (see Methods for details). These samples were placed in MoBio PowerSoil kits, frozen, and shipped back to OSU or Penn State for processing. DNA was extracted according to the manufactures recommendations, and 16S V4 amplicon libraries generated (see Methods for details). All amplicon data then underwent quality control parsing (see Methods) prior to further downstream analysis. Amplicons were then subjected to several analytics to address different questions about the influence of the host and environment on coral microbiomes (see flow chart).



Supplementary Figure 1.2. Microbiome consistency among coral compartments.

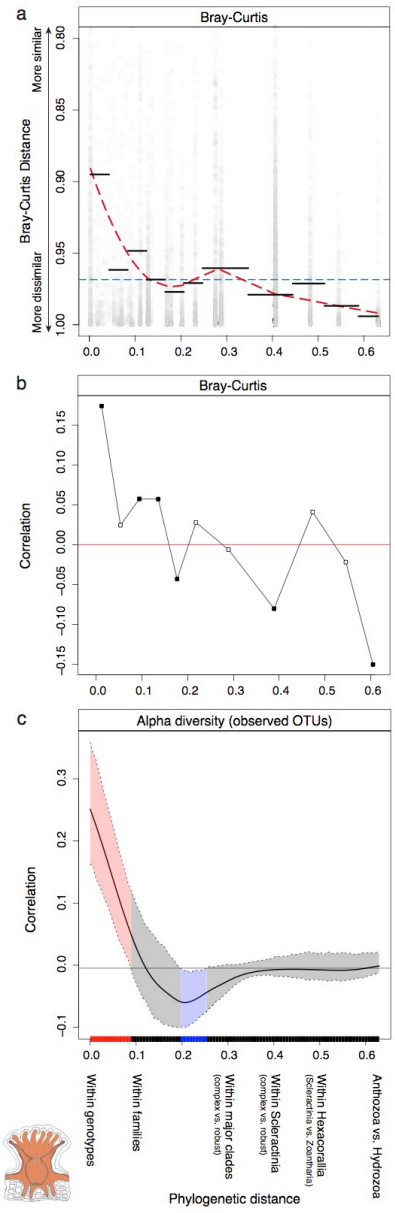
Compartments are denoted by color as mucus (teal), tissue (orange), and skeleton (purple). **a**) Core microbiome abundance. Bars show the proportion of sequence reads for each compartment that belonged to members of the core microbiome for that compartment. Microbial orders were deemed "core" if present in ≥ 70% of samples. **b**) Beta-diversity. Mucus compartments showed less microbiome variability than tissue and skeleton. All p-values reflect Tukey's HSD.



Supplementary Figure 1.3. Influence of coral traits and local environment on microbial community composition.

This figure is presented for comparison with Fig 1.1c. The heatmap visualizes the influence of several host and environmental factors on microbial community composition in coral mucus tissue or skeleton. This figure is identical to **Fig. 1.1c** in the main text, except that it presents raw R^2 values, rather than R^2 values that are Z-score normalized within each factor. Thus, this figure is more useful for seeing which factors are influential in an absolute sense (e.g. even in mucus different host species have different microbiomes), while Fig 1c highlights the relative influence of each parameter across compartments (e.g. host species has a much stronger relative influence on tissue and skeleton microbiomes than on mucus). All the results in the figure are based on results from the Weighted UniFrac multivariate dissimilarity measure on data tables rarified to 1000 sequences/sample (see Supplementary Data S4 for alternative choices of

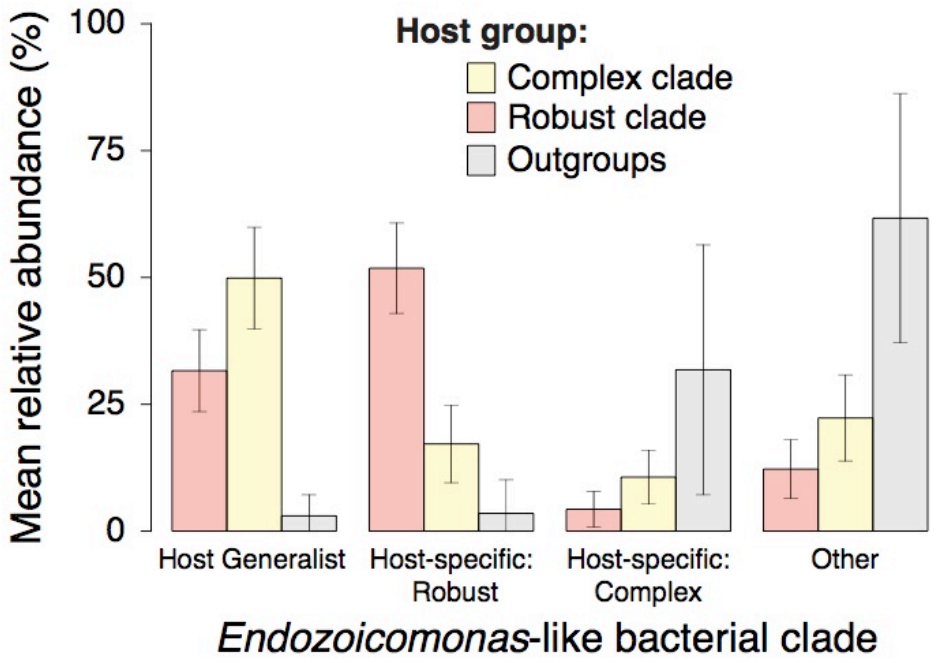
dissimilarity measure or rarefaction depth). Light cells represent lower *Adonis* adjusted R^2 values for that factor (i.e. traits that have lesser influence on microbial community composition in a given compartment), whereas darker colors represent traits with a stronger influence. Traits were automatically clustered according to their R^2 values across compartments, and compartments were clustered according to the similarity of R^2 values within them. Host vs. environmental traits were manually colorized to highlight the split between host and environmental traits that emerged from clustering. Significance values for all trait x compartment combinations are available in Supplementary Data 4, and Bonferroni-corrected significance values for each combination are marked in Fig. 1c.



Supplementary Figure 1.4. Phylogenetic correlograms of tissue microbiome diversity.

a) Scatterplot of phylogenetic distances versus Bray-Curtis dissimilarity. The grand mean of all pairwise community dissimilarities is shown as a dashed blue line, and the mean of community dissimilarities within each phylogenetic distance class is plotted as a horizontal black line throughout that class. A smoothed (loess) curve showing the overall trend in community dissimilarity is displayed as a dashed red line. **b)** Phylogenetic

Mantel correlogram is shown for Bray-Curtis microbiome dissimilarity, with Mantel *r* versus coral host phylogenetic distance. Solid black boxes indicate phylogenetic distance classes within which pairwise microbial community dissimilarities are significantly different (i.e., significantly more similar or significantly less similar) from dissimilarities in all other distance classes (Mantel's *r*, $p < 0.05$), whereas open boxes indicate distance classes where dissimilarities are not significant. **c)** Correlogram showing autocorrelation in microbiome richness (Moran's *I* calculated from observed OTU values) versus coral host phylogenetic distance. Small phylogenetic distances indicate coral species of recent divergence. Red bars on the x-axis correspond to phylogenetic distances where microbial community parameters are significantly more similar between samples at a given distance class than between samples at all other phylogenetic distances, and blue bars correspond to distances where parameters are significantly less similar.



Supplementary Figure 1.5. Distribution of *Endozoicomonas*-like bacteria across Robust vs. Complex corals.

The abundances of each *Endozoicomonas*-like bacterial clade (i.e., Host Generalist, Host-specific: Robust, and Host-specific: Complex, and others) in coral tissue samples are shown, relative to the total abundance of *Endozoicomonas*-like bacteria. Error bars indicate standard error ($n_{complex}$ = 90; n_{robust} = 124, $n_{outgroups}$ = 10).

Supplementary Data

Supplementary Data 1. **Sample Summary.** This Excel file summarizes samples collected and reefs visited in this study. This information is also available from the QIIME mapping file for samples (**Supplementary Data 2)**, but is summarized here for easier reference. **a.** Samples collected, subdivided by geographic region, coral species and coral compartment. **b.** Reefs visited as part of the study, along with longitude and latitude.

Supplementary Data 2. **Sample Metadata.** This Excel file is the QIIME mapping file containing all metadata used throughout the analysis.

Supplementary Data 3. **Prevalent 'Core' Microbes.** This Excel file summarizes prevalent microbes associated with coral microbiomes (e.g. 'core microbiomes', *sensu lato*) surveyed in this study, quantification of the effects of rarefaction depth on which OTUs are prevalent, and comparisons against two literature references. **a.** Graphical summary of microbial OTUs that had >70% prevalence at 1000, 5000, 10000, 15,000, or 20,000 sequences per sample. **b.** Machine readable data table of the prevalence and taxonomy of OTUs from panel a. **c.** Prevalence of microbial orders in the coral microbiome at 1000 seqs/sample, and a comparison of prevalent microbes in coral mucus with Zaneveld et al., 2016. **d.** Comparison with results from Apprill et al., 2016, conducted under similar rarefaction depth (10,000 seqs/sample) and prevalence threshold (50%).

Supplementary Data 4. Beta-diversity (multivariate dissimilarities). This Excel file provides a detailed accounting of factors influencing microbial β-diversity (multivariate dissimilarities or community composition) in each compartment according to several β-diversity metrics, and across rarefaction depths. **a.** Factors influencing microbiome β-diversity by compartment at 1,000 sequences per sample. **b**. Factors influencing microbiome β-diversity by compartment at 10,000 sequences per sample. **c.** Factors consistently and strongly associated with coral microbiome β-diversity by compartment at 1,000 sequences per sample. To find which factors were most consistently associated with microbiome beta-diversity, we calculated factors that were a) significant b) had adjusted R2 >= 0.05 for all distance metrics analyzed. **d.** Summary of how taxonomic

ranks structure beta-diversity. Data from a and b are combined to illustrate that more specific taxonomic ranks for corals provide more information about microbial beta-diversity than more general ranks, regardless of distance metric and rarefaction depth chosen.

Supplementary Data 5. Alpha-diversity or richness. This Excel file provides a detailed accounting of microbiome richness. **a.** Results for permutational T-tests comparing coral microbiome vs. environmental community richness. Bacterial and Archaeal Diversity of Corals vs. Water and Sediment. Data reflect richness per 1000 reads of coral mucus, tissue, or sediment vs. reef water or sediment. **b.** Results for permutational T-tests comparing microbiome richness of corals vs. outgroups surveyed (blue corals, matt anemones, hydrozoans, etc).

Supplementary Data 6. Microbes correlated with host and environmental parameters. This Excel file summarizes microbes that were correlated with host and environmental parameters using either Spearman or Phylogenetic GLMM analysis. **a.** Spearman results summary. Summary of the number of bacterial genera significantly correlated with selected host or environmental metadata in each compartment, assessed by FDR-controlled Spearman regressions. For each factor, the Greeengenes taxonomy and R value for the top 3 genera positively or negatively correlated with that factor are listed. Additional statistical tests assess whether positive or negative associations are enriched for a given factor. **b.** Phylogenetic GLMM summary. Summary of numbers of genera associated with a subset of host and environmental factors, as identified in phylogenetic GLMMs. **c.** Full phylogenetic GLMM results. A comprehensive list of genera associated with a subset of host and environmental parameters, subdivided by tissue compartment, along with p values, estimated effect sizes, and 95% confidence intervals. For categorical data, the category value with which a microbial genus is associated is also reported.

Supplementary Data 7. Random Forest results. This Excel file describes coral host features that can be predicted from the microbiome. To measure the strength of association between the microbiome and coral physiology, we attempted to build supervised classification models using Random Forests analysis, and then back-predict

certain host features. This addresses the question: "given microbial data alone, how much can you say about the coral host?". **a.** Results of Random Forest models of the coral microbiome, subdivided by compartment and the host trait predicted. **b.** Summary of host factors that can be accurately predicted from the coral microbiome, and which compartments predict them. Raw accuracies and error ratios are presented, and results with accuracy >70% and error ratios >1.0 are highlighted.

Supplementary Data 8. Mantel test results. This Excel file summarizes results from Mantel tests and Mantel correlograms (Methods). These permutation-based tests assess the degree of correlation between two distance matrices (e.g. geographic distance and genetic distance, etc). Here they were applied to test the extent to which host evolutionary distances corresponded to differences in microbiome composition, as reflected by between-sample beta-diversity distances. We calculated this measure for both a non-phylogenetic measure (Bray-Curtis divergences) and a phylogenetic beta-diversity distance metric (Weighted UniFrac distances).

Supplementary References

1. Apprill, A., Weber, L. G. & Santoro, A. E. Distinguishing between Microbial Habitats Unravels Ecological Complexity in Coral Microbiomes. mSystems 1, e00143-16 (2016).
2. Zaneveld, J. R. et al. Overfishing and nutrient pollution interact with temperature to disrupt coral reefs down to microbial scales. Nat. Commun. 7, ncomms11833 (2016).
3. Willis, B. L., Page, C. A. & Dinsdale, E. A. Coral disease on the Great Barrier Reef. in Coral health and disease (eds. Rosenberg, E. & Loya, Y.) 69–104 (Springer, 2004).
4. Darling, E. S., Alvarez-Filip, L., Oliver, T. A., McClanahan, T. R. & Côté, I. M. Evaluating life-history strategies of reef corals from species traits. Ecol. Lett. 15, 1378–1386 (2012).

Supplementary text for chapter 3: Novel generalized linear model framework identifies a clade of Symbiodiniaceae that is associated with coral skeletal samples, but not tissue or mucus

Supplementary Table 3.1. Summary of coral colonies sampled by location and host genus

Ocean Ocean Area Reef Host Genus	Number of colonies
Pacific	**183**
Coral Sea	**133**
Big Vickie	7
Echinopora	1
Galaxea	1
Hydnophora	1
Isopora	1
Lobophytum	1
Pachyseris	1
Pocillopora	1
Broadhurst	8
Acropora	1
Diploastrea	1
Echinopora	1
Galaxea	1
Lobophyllia	1
Pavona	1
Pocillopora	1
Porites	1
Day	2
Acropora	1
Diploastrea	1
Fork	8
Acropora	4
Diploastrea	4
Horseshoe	14
Acropora	1

Diploastrea	1
Galaxea	1
Leptastrea	1
Lobophyllia	2
Millepora	2
Pachyseris	1
Palythoa	1
Pavona	1
Pocillopora	1
Turbinaria	2
Kelso	2
Acropora	2
Lagoon Bommie	3
Echinopora	1
Pachyseris	1
Physogyra	1
Lagoon entrance	2
Lobophyllia	1
Porites	1
Linnett - Site 1	1
Diploastrea	1
Little Kelso	5
Acropora	1
Diploastrea	4
Martin - Site 1	2
Diploastrea	2
Martin - Site 2	2
Diploastrea	2
Maxwell - Site 2	2
Acropora	2
Trawler	69
Acropora	4
Caulastraea	1
Diploastrea	3
Echinophyllia	1
Echinopora	4
Favites	2
Fungid	5

Galaxea	3
Heliopora	1
Isopora	1
Lobophyllia	5
Millepora	2
Montipora	1
Pachyseris	5
Palythoa	3
Pavona	6
Physogyra	4
Porites	6
Psammocora	2
Stylophora	5
Turbinaria	5
Yonge - Site 1	3
Acropora	1
Diploastrea	2
Yonge - Site 2	3
Acropora	2
Diploastrea	1
South Pacific	**16**
LTER_1_Backreef	7
Acanthastrea	1
Acropora	1
Leptastrea	1
Lobophyllia	1
Montipora	1
Pavona	1
Pocillopora	1
LTER_1_Fringing	9
Acropora	1
Discosoma	1
Fungid	1
Leptastrea	1
Pavona	1
Pocillopora	1
Porites	2
Psammocora	1

Tasman Sea	**34**
Comets Hole	3
Alveopora	3
Far Flats	13
Astrea	3
Dipsastraea	3
Favites	3
Goniastrea	3
Homophyllia	1
North Bay	18
Acropora	3
Cyphastrea	3
Isopora	3
Pocillopora	2
Porites	3
Seriatopora	3
Stylophora	1
Indian	**61**
Eastern Indian	**49**
Bills Bay	43
Acropora	2
Cyphastrea	3
Dipsastraea	1
Echinophyllia	1
Echinopora	3
Favites	5
Galaxea	3
Goniastrea	2
Hydnophora	3
Lobophyllia	1
Merulina	3
Montipora	4
Pavona	3
Platygyra	3
Pocillopora	3
Seriatopora	3
Bills Bommie	6
Lobophyllia	2

Porites	4
Red Sea	**2**
Al Fahal	2
Pocillopora	1
Porites	1
South China Sea	**1**
Raffles Lighthouse	1
Pocillopora	1
Western Indian	**9**
Boucan Canot	2
Pocillopora	1
Porites	1
Pass	2
Pocillopora	1
Porites	1
Saint-Leu	1
Pocillopora	1
Trou d_Eau	3
Pocillopora	1
Porites	2
Volcano	1
Pocillopora	1
Atlantic	**8**
Caribbean	**2**
West_of_channel	2
Porites	2
Mediterranean	**6**
Michmoret	6
Madracis	2
Oculina	4
Grand Total	**252**