Constraining the interpretation of 2-methylhopanoids through genetic and phylogenetic methods

Thesis by

Jessica Nicole Ricci

In Partial Fulfillment of the Requirements for the degree
of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California
2015
(Defended May 21, 2015)

ACKNOWLEDGEMENTS

There are many people I would like to thank for their support during my time as a graduate student. Firstly, I am grateful to my Ph.D. advisor Dianne Newman for allowing me to join her lab and for her helpful advice over the last five years. I have interacted with many wonderful people in the Newman lab who have helped me with my research enormously, especially Gargi, Chia, Dave, Caj, Alice, Megan, Suzanne, Nick, and Elise. I would also like to thank my committee members Paul Sternberg, Alex Sessions, Sarkis Mazmanian, and Jared Leadbetter for their useful advice and suggestions. A special thanks goes to Woodward Fischer, Victoria Orphan, and members of their labs for many valuable and stimulating discussions.

My experience at Caltech has been enriched by a number of teaching-related activities. Thank you to Cassandra, Tina, Noelle, and Daniel. I have appreciated the countless seminars, workshops, and events you have organized that have without a doubt made me a better teacher. I am particularly grateful to Ryan Skophammer, who showed me that teaching K-12 is not a scary as I thought it would be.

Lastly, I would like to thank my family for always encouraging me to pursue my interests. Mom, Dad, and John Paul, thank you for supporting me during my graduate career and beyond. To my partner, Alan, without your unwavering encouragement this would not have been possible. Thank you to my cat, Oscar, who I can always count on to greet me with lots of purring after a long day.

ABSTRACT

Hopanoids are a class of sterol-like lipids produced by select bacteria. Their preservation in the rock record for billions of years as fossilized hopanes lends them geological significance. Much of the structural diversity present in this class of molecules, which likely underpins important biological functions, is lost during fossilization. Yet, one type of modification that persists during preservation is methylation at C-2. The resulting 2-methylhopanoids are prominent molecular fossils and have an intriguing pattern over time, exhibiting increases in abundance associated with Ocean Anoxic Events during the Phanerozoic. This thesis uses diverse methods to address what the presence of 2-methylhopanes tells us about the microbial life and environmental conditions of their ancient depositional settings. Through an environmental survey of *hpnP*, the gene encoding the C-2 hopanoid methylase, we found that many different taxa are capable of producing 2-methylhopanoids in more diverse modern environments than expected. This study also revealed that *hpnP* is significantly overrepresented in organisms that are plant symbionts, in environments associated with plants, and with metabolisms that support plant-microbe interactions; collectively, these correlations provide a clue about the biological importance of 2-methylhopanoids. Phylogenetic reconstruction of the evolutionary history of *hpnP* revealed that 2-methylhopanoid production arose in the Alphaproteobacteria, indicating that the origin of these molecules is younger than originally thought. Additionally, we took genetic approach to understand the role of 2-methylhopanoids in Cyanobacteria using the filamentous symbiotic *Nostoc punctiforme*. We found that hopanoids likely aid in rigidifying the cell membrane but do not appear to provide resistance to osmotic or outer membrane stressors, as has been shown in other organisms. The work presented in this thesis supports previous findings that 2-methylhopanoids are not biomarkers for oxygenic photosynthesis and provides new insights by defining their distribution in modern environments, identifying their evolutionary origin, and investigating their role in Cyanobacteria. These efforts in modern settings aid the formation of a robust interpretation of 2-methylhopanes in the rock record.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter I

Chapter II

Chapter III

Chapter IV

LIST OF TABLES

Chapter II

Chapter III

Chapter IV

Preface

OVERVIEW OF CHAPTERS

Chapter 1 introduces hopanoids and 2-methylhopanoids as geologically important molecular fossils and compares them to other fossils and biomarkers in terms of longevity and information richness. The putative biological functions of hopanoids and their methylated counterparts, as well as the 2-methylhopane fossil record, are also discussed. Portions of Chapter 1 will be incorporated into an article entitled "Cellular Biological Approaches to Interpreting Molecular Fossils" to be submitted to Annual Reviews in Earth and Planetary Sciences.

Chapter 2 explores the environmental distribution of *hpnP*, the C-2 hopanoid methylase. *hpnP* sequences from diverse organismal sources were found in a variety of locales, but the presence of the gene was found to significantly correlate with plant-associated microorganisms, environments, and metabolisms. From our findings, we propose a niche where *hpnP* is enriched: sessile microbial communities with low oxygen, low fixed nitrogen, and high osmolarity. This study was published in the ISME Journal: Ricci JN, Coleman ML, Welander PV, Sessions AL, Summons RE, Spear JR, and Newman D (2014) Diverse capacity for 2-methylhopanoid production correlates with a specific ecological niche. *The ISME Journal*, **8**, 675-684. Supplemental datasets for this chapter can be found in **Appendix A**.

Chapter 3 investigates the evolutionary history of HpnP through phylogenetics, revealing Alphaproteobacteria to be the originators of 2-methylhopanoids. Initial evolution of *hpnP* in this phylum implies that these molecules arose later than previously thought and provides insight into the ancestral function of 2-methylhopanoids. This work was originally published in Geobiology: Ricci JN, Michel AJ, Newman DK (2015) Phylogenetic analysis of HpnP reveals the origin of 2-

methylhopanoid production in Alphaproteobacteria. *Geobiology*, **13**, 267-277. The supplementary dataset for this study can be found in **Appendix B**.

Chapter 4 examines the biological roles of hopanoids and 2-methylhopanoids in *Nostoc punctiforme*, a filamentous symbiotic cyanobacterium. We report that a hopanoid-lacking mutant displays growth phenotypes at extreme temperature consistent with hopanoids rigidifying the cell membrane. Yet hopanoids did not play significant roles in symbiotic efficiency, resistance to outer membrane destabilization, or osmotic stress tolerance, as has been shown in other species. This chapter was a collaborative effort with Michael Summers (California State University, Northridge).

Chapter 5 summarizes the findings of this thesis with a focus on how they have improved the understanding of 2-methylhopanoids as biomarkers. It then turns an eye to the future by outlining further studies that build upon the conclusions presented in prior chapters.

Chapter 1

INTERPRETING HOPANOIDS AS MOLCULAR FOSSILS

Parts of this chapter will be reformatted for an article entitled “Cellular Biological Approaches to Interpreting Molecular Fossils” to be submitted to Annual Reviews in Earth and Planetary Sciences. Approximately 50% of the text in this chapter was contributed by Ann Pearson (Harvard University).

Fossil types

Fossils, the preserved remains and traces of ancient organisms, can be used to unravel the history of life on Earth. Some fossils are the physical remains of organisms that have mineralized, while others are imprints, traces, or burrows that life has left behind. The morphology of both these fossil types can be compared to modern species to decipher their evolutionary relationships. Organisms can also leave chemical fossils by disturbing isotope records or depositing biomolecules (i.e., DNA, RNA, proteins, and lipids) that can be preserved as biomarkers or molecular fossils. Biomarkers can serve as a fossil record of organisms that are unlikely to leave other types of fossils due to their size and lack of hard bodies, such as bacteria. They carry information about their source organism in their sequence of nucleotides or amino acids or within their lipid structure, which often have more diagnostic characteristics than traditional morphological fossils. With the large availability of sequences to compare to, it is easy to decode the rich information stored in DNA, RNA, and proteins.

Unfortunately, with the exception of a few rare cases, DNA, RNA, and protein degrade quickly on geologic timescales (Brocks & Banfield 2009). RNA is an extremely short-lived molecule with a half-life on the timescale of hours. DNA and protein have been recovered from significantly older samples, but these instances are limited to particular environments and molecules (Figure 1.1). For example, 600,000-year-old plant DNA has been retrieved from sediments, but the DNA half-life in aqueous environments is thought to be on the order of weeks (Willerslev et al. 2003; Pedersen et al. 2015). Collagen protein sequences have been recovered from dinosaurs ranging from 68–80 million years (Myr) ago (Asara et al. 2007; Schweitzer et al. 2009), yet no other proteins were found.

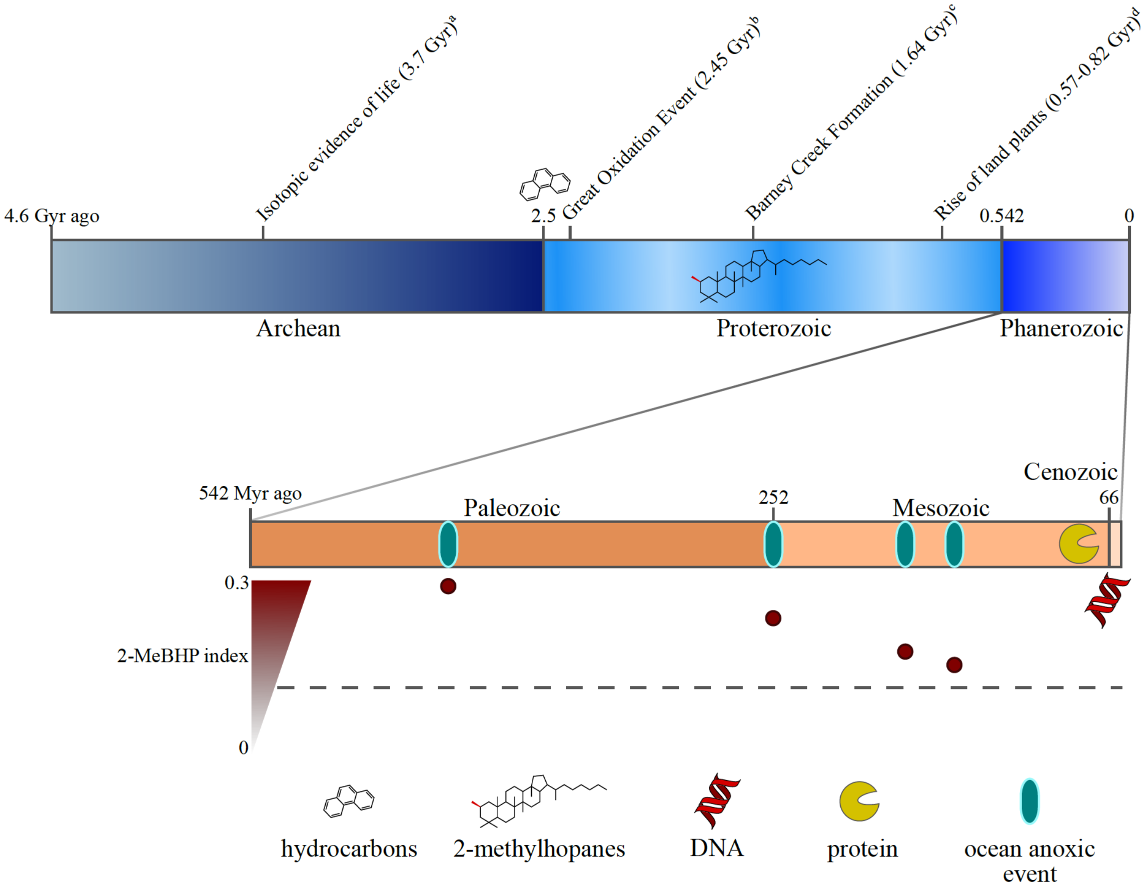


Figure 1.1 | Timeline of select geological events and oldest biomarker fossils discussed in the text. Lipid, DNA, and protein symbols indicate the oldest unknown occurrences of these molecules in the rock record (Brocks et al. 2003, 2005; Willerslev et al. 2003; Schweitzer et al. 2009). Phanerozoic 2-methylhopane (2-MeBHP) indices measured during OAEs are represented with red dots, while the average 2-methylhopane index across this timeframe is denoted by the gray dashed line (Knoll et al. 2007). [a]Rosing, 1999 [b]Bekker et al. 2004 [c]Brocks et al. 2005 [d]Clarke et al. 2011.

Unique among biomarkers, lipids are preserved in the geologic record for extended lengths of time. Diagenesis, the process of sediment becoming rock, may involve isomerization and reduction, but often the original hydrocarbon skeleton remains recognizable – if not identical – relative to the original biosynthetic product as it would have existed in the source organism. This remarkable property of lipids is most directly manifested in the form of extractable petroleum deposits, the oldest of which date to 1.64 billion years (Gyr) ago in the Mesoproterozoic and are undisputedly the remnants of ancient ecosystems (Peters et al. 2005) (Figure 1.1). More broadly, preserved hydrocarbons also exist throughout the immature sedimentary rocks of the Proterozoic and early Phanerozoic dating to 2.5 Gyr ago (Brocks et al. 2003) (Figure 1.1). In most places, their concentrations are far too low for economic viability, yet remain well above the threshold needed for biogeochemical analyses; it is from these rocks that we gain much of our insight into the poorly-fossiliferous young Earth. In contrast, no syngenetic lipids have been found in Archean rocks to date (Brocks et al. 1999; Rasmussen et al. 2008; Brocks 2011).

The most valuable lipid biomarkers are those that can be unequivocally correlated to their biological sources. They provide a window into Earth history that is at least as long as that provided by other more traditional types of fossils. Lipid fossils have helped establish the existence of multicellular animals prior to 635 Myr ago (Love et al. 2009) and the presence of the oxygen-dependent sterol biosynthetic pathway by at least 1640 Myr ago (Brocks et al. 2005). These dates are complementary to, and extend the information provided by, microscopic fossils and trace fossils alone.

Introduction to hopanoids and 2-methylhopanoids

Among the longest-lived molecular fossils are compounds from the class of triterpenoid lipids known as hopanoids (Figure 1.2). Their diagenetic products, hopanes, are ubiquitous in modern and ancient systems. It has been estimated that billions of kilograms of hopanes are stored in sedimentary rocks and oil reservoirs, a mass that is almost as high as the combined carbon mass in all living organisms (Ourisson & Albrecht 1992). For many years, hopanes were regarded as 'orphan' lipids because their dominant sources in sediments were unknown. Discovery of the parent compounds in *Acetobacter xylinium* led to the proposal that bacterial hopanoids are functional analogues to eukaryotic sterols, such as cholesterol (III), which assist in rigidifying membranes (Forster et al. 1973; Rohmer & Ourisson 1976; Rohmer et al. 1979). Although hopanoids are present in approximately 10% of bacteria and have an irregular phylogenetic distribution (Pearson et al. 2007; Frickey & Kannenberg 2009), they are widespread in diverse contemporary environments (Talbot et al. 2003; Farrimond et al. 2004; Talbot & Farrimond 2007; Xu et al. 2009; Pearson et al. 2009; Zhu et al. 2011; Sáenz et al. 2011a) and throughout geologic time (Peters et al. 2005), bearing the potential to provide insight into past ecological and environmental conditions.

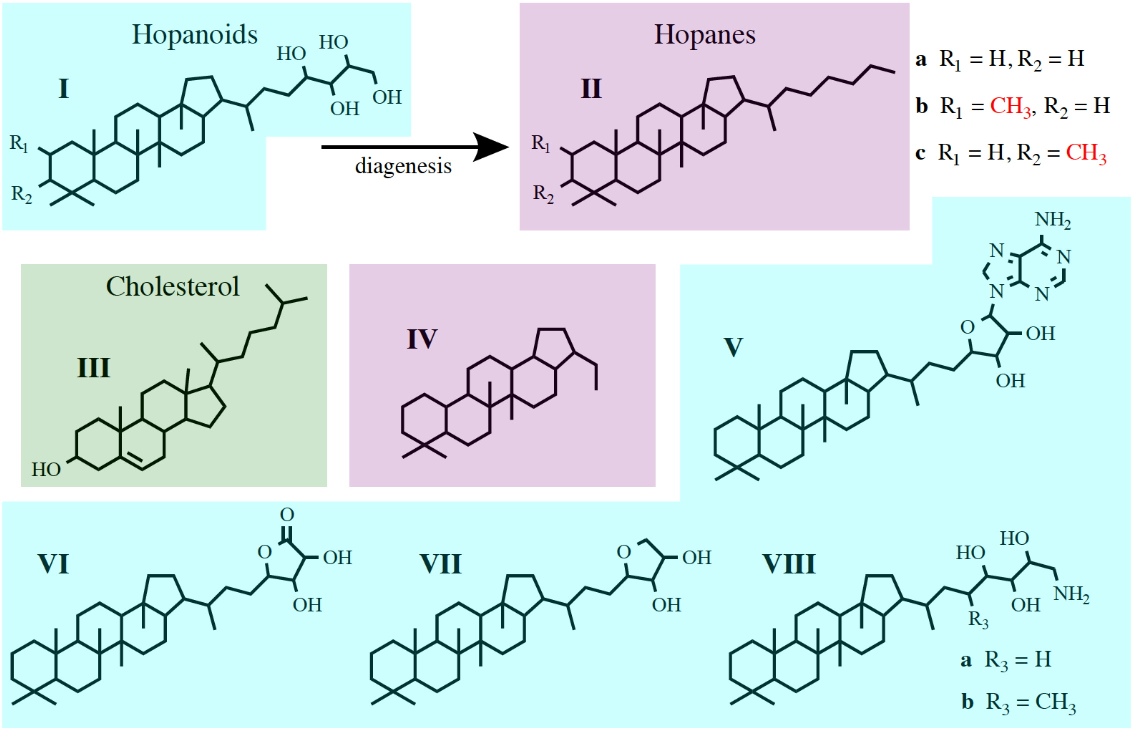


Figure 1.2 | Structures of select hopanoids found in extant organisms, modern environments, and ancient sedimentary depositions. During diagenesis, hopanoids become fossilized hopanes and loss many of their differentiating structures, except C-2 and C-3 methylations. Abbreviations include Ia bacteriohopanetetrol, Ib 2-methylbacteriohopanetetrol, Ic 3-methylbacteriohopanetetrol, IIa C_{35} hopane (i.e., pentakishomohopane), IIb C_{35} 2-methylhopane, IIc C_{35} 3-methylhopane, III cholesterol, IV 25,28,30-trisnorhopane, V adenosylhopane, VI hopaneribonolactone, VII 32,35-anhydrobacteriohopanetetrol, VIIIa aminobacteriohopanetetrol, and VIIIb aminobacteriohopanepentol.

During diagenesis, the process of sediment becoming rock, hopanoid parent molecules lose many of their differentiating features, such as hydroxyl and amine groups, leaving behind only their hydrocarbon skeletons. In contrast, modifications that can be preserved include methylations to the hydrocarbon backbone at C-2 or C-3 (Figure 1.2). Methylated hopanoids are important biomarkers because methylation allows different hopanes to be distinguished from each other in the fossil record. 2-Methylhopanes (e.g., C_{35} 2-methylhopane) (Figure 1.2 IIb), the preserved form of 2-methylhopanoids (e.g., 2-methylbacteriohopanetetrol) (Figure 1.2 Ib), are among the oldest syngenetic biomarkers dating to 1.64 Gyr ago in the Barney Creek Formation, an anoxic sulfidic permanently stratified marine depositional setting (Brocks et al. 2005) (Figure 1.1). Furthermore, the 2-methylhopane index, the ratio of 2-methylhopanes to methylated and unmethylated hopanes (C_{30} hopanes are used to calculate the index in the rock record but are derived from C_{35} hopanoids), shows intriguing patterns of change over time (Summons et al. 1999; Knoll et al. 2007). Relatively high 2-methylhopane indices are found throughout the Proterozoic (Summons et al. 1999), while the highest measurements co-occur with Ocean Anoxic Events (OAEs), including the Permo-Triassic boundary (Xie et al. 2005) and during Cretaceous OAEs (Kuypers et al. 2004) (Figure 1.1). OAEs are characterized by depletion of oxygen in the oceans but are also associated with global disruptions in nutrient cycling and mass extinctions. This correlation raises the question of what 2-methylhopanes, as well as hopanes in general, tell us about OAEs and about other times in Earth history.

Proposed interpretations of hopanoids and 2-methylhopanoids

In general, hopanoids in modern sediments and the geologic record have been used as indicators for bacterial activity or bacterial input to sediments. Numerous minor structural variations within the molecular class can be linked to more particular functional, taxonomic, or environmental interpretations. Certain hopanoids have been suggested to be diagnostic tools for sedimentary diagenesis, tracers of soil, terrestrial, or estuarine carbon, indicators of methanotrophic activity, proxies for oxygen, and associated with nitrogen fixation. We discuss the rational behind these putative interpretations below.

Indicators of Sediment Diagenesis

Diagenetic modification of hopanoids leads to a rich variety of hopanes in the geologic record. While degradation can preserve critical aspects of the original skeletal structure (e.g., 2-methylhopane), other parts of the molecule are often altered in ways that are both predictable and diagnostic. The ratio of total hopanes to total steranes has been used to interpret changes in the fraction of bacterial versus algal input to sediment and petroleum (Moldowan et al. 1985; Sinninghe Damsté & Schouten 1997; Andrusevich et al. 2000). An increased abundance of norhopanes (e.g., 25,28,30-trisnorhopane) (Figure 1.2 IV) indicates a direct hydrocarbon contribution to sediments, because these compounds are products of biodegradation and are not produced during thermal maturation (Noble et al. 1985). Hopane chain length proxies such as the C_{35} homohopane index, defined as $[C_{35}/(SC_{31}\text{-}C_{35})]$, record the extent of degradation of hopanoids. A higher index, denoting less degradation, is believed to reflect greater anoxia (Peters & Moldowan 1991). Together these various proxies are used as tools to interpret the environmental conditions under which ancient sediments were deposited. For example, a high homohopane index accompanies other indicators of a biotic, oxygen-poor crisis associated with the Permo-Triassic extinction (Hays et al. 2012).

Tracer of Terrestrial Environments

In recent sediments, intact hopanoids may be useful tracers for carbon source inputs. Although few hopanoids are presently thought to be environmentally or metabolically specific, adenosylhopane (Figure 1.2 V) may be an exception. It is abundant in terrestrial and lacustrine environments, whereas it is absent in sediments with negligible terrigenous input (Talbot & Farrimond 2007;

Cooke et al. 2008; Xu et al. 2009; Pearson et al. 2009). However, because adenosylhopane is the first biosynthetic intermediate in the formation of all hopanoids with extended side-chains (Bradley et al. 2010; Welander et al. 2012), the accumulation of adenosylhopane by definition cannot be taxonomically diagnostic. It remains unknown why this apparently "arrested synthesis" would be more prevalent in soil-dwelling microbes, but it nonetheless appears to be a good tracer for terrestrial input. Similarly, the side-chain structures of hopaneribonolactone (Figure 1.2 VI) and 32,35-anhydrobacteriohopanetetrol (Figure 1.2 VII) have been found in oxidizing and reducing environments, respectively (Bednarczyk et al. 2005; Talbot et al. 2005; Sáenz et al. 2011b). Bradley et al. (2010) suggested that both compounds are generated abiotically through oxidative or reductive cleavage of an as-yet unidentified precursor. The ratio of hopaneribonolactone to 32,35-anhydrobacteriohopanetetrol might provide a useful environmental or physiological signal, albeit independent of any relationship to taxonomy.

Markers of Methanotrophy

Among hopanoids that may have taxonomic associations, the best examples may be those that contain amine groups. Aminobacteriohopanetetrol (Figure 1.2 VIIIa) has only been found in methanotrophs and *Desulfovibrio* spp. (Blumenberg et al. 2006), and it with aminobacteriohopanepentol (Figure 1.2 VIIIb) are observed in areas where aerobic methane oxidation is an important component of the carbon cycle (Talbot et al. 2003; Zhu et al. 2010; van Winden et al. 2012). However, the best confirmation of a methanotrophic signal is the simultaneous evidence of depleted $\delta^{13}C$ values that can be found in sedimentary hopanoids, many of them being the potential diagenetic products of these functionalized precursors. Such examples are primarily associated with lacustrine systems (Freeman et al. 1990) or seep zones in which methane enters oxygenated marine bottom waters (Elvert & Niemann 2008). Interestingly, this signal also occurs in anoxic sediments associated with anaerobic oxidation of methane (Thiel et al. 2003; Pancost et al.

2010). Differentiation between aerobic and anaerobic oxidation of methane therefore has relied on the presence of 3-methylhopanoids, markers for aerobic and/or acetic acid bacteria (Zundel & Rohmer 1985) (Figure 1.2 Ic IIc) that perhaps have a wider taxonomic and ecological distribution (Welander & Summons 2012).

Proxies for Oxygen

Early studies of bacterial culture collections found evidence for hopanoid production only in bacteria growing aerobically, suggesting that geologic hopanes were good general indicators for oxygenated environments (Rohmer et al. 1984). Now, hopanoids have been reported in a wide range of bacterial species growing in anaerobic conditions and have been found in diverse environments, suggesting there is not a requisite connection between hopanoids and aerobic ecosystems or habitats (Fischer et al. 2005; Härtner et al. 2005; Blumenberg et al. 2006; Eickhoff et al. 2013).

Similarly, the distinctive 2-methylhopane skeleton was initially determined to be common in freshwater and mat-dwelling cyanobacteria (Summons et al. 1999) (Figure 1.2 IIb). Thus, when 2-methylhopanes were detected in 2.7 Ga strata, they were interpreted as evidence for the very ancient evolution of oxygenic photosynthesis (Brocks et al. 1999). Two lines of evidence shadow doubt on this conclusion: the syngeneity of these Archean biomarkers has been called into question (Rasmussen et al. 2008; French et al. 2015) and uncertainty that exists about the strength of the taxonomic correlation with Cyanobacteria. In particular, the anoxygenic phototrophic alphaproteobacterium, *Rhodopseudomonas palustris* TIE-1 was shown to make 2-methylhopanoids in equivalent abundance to many cyanobacteria (Rashby et al. 2007). Additionally, the identification of a gene that codes for the C-2 methylase, *hpnP*, expanded the known taxonomic diversity of 2-methylhopanoids producers to include many more Alphaproteobacteria and an acidobacterium. Although *hpnP* was discovered through genetic means, a phenotype for its deletion was not found, leaving open the question of 2-methylhopanoids' biological role (Welander et al. 2010); as a whole, these data rule out 2-methylhopanoids as specific biomarkers for Cyanobacteria and oxygenic photosynthesis.

Hopanoids and Nitrogen Fixation

Hopanoids have been proposed to be able to protect the nitrogenase against damage by oxygen during nitrogen fixation. For example, in *Frankia* spp., nitrogen-fixing root-nodule symbionts, the number of hopanoid rich laminated membrane layers around their nitrogen-fixing vesicles increase as oxygen partial pressure increases (Berry et al. 1993; Nalin et al. 2000). Additionally, a *Bradyrhizobium* sp. BTail strain that is unable to make hopanoids, including 2-methylhopanoids, is unable to form effective symbiosis with its plant partner, *Aeschynomene evenia*. The authors suggest that this defect is due to a weakened cell membrane, causing a decrease in nitrogen fixation efficiency (Silipo et al. 2014). In a separate study, particular hopanoids, especially 2-methylhopanoids, have been shown to rigidify the membrane, which could be a potential mechanism for protecting the nitrogenase in the *Frankia* spp. and *Bradyrhizobium* sp. BTail symbioses (Wu et al. 2015). To further support the association between 2-methylhopanoids and nitrogen fixation, casual observations suggest that many 2-methylhopanoid producers fix nitrogen, including numerous members of the Cyanobacteria and Alphaproteobacteria. Work in the cyanobacterium *Nostoc punctiforme* has shown that heterocysts (nitrogen-fixing cells) have the highest ratio of 2-methylhopanoids among the cell types tested; in contrast akinetes (spore-like cells) have the largest abundance of hopanoids, including C-2 methylated members (Doughty et al. 2009) (Figure 1.3). Further investigation is needed to refine these noteworthy results.

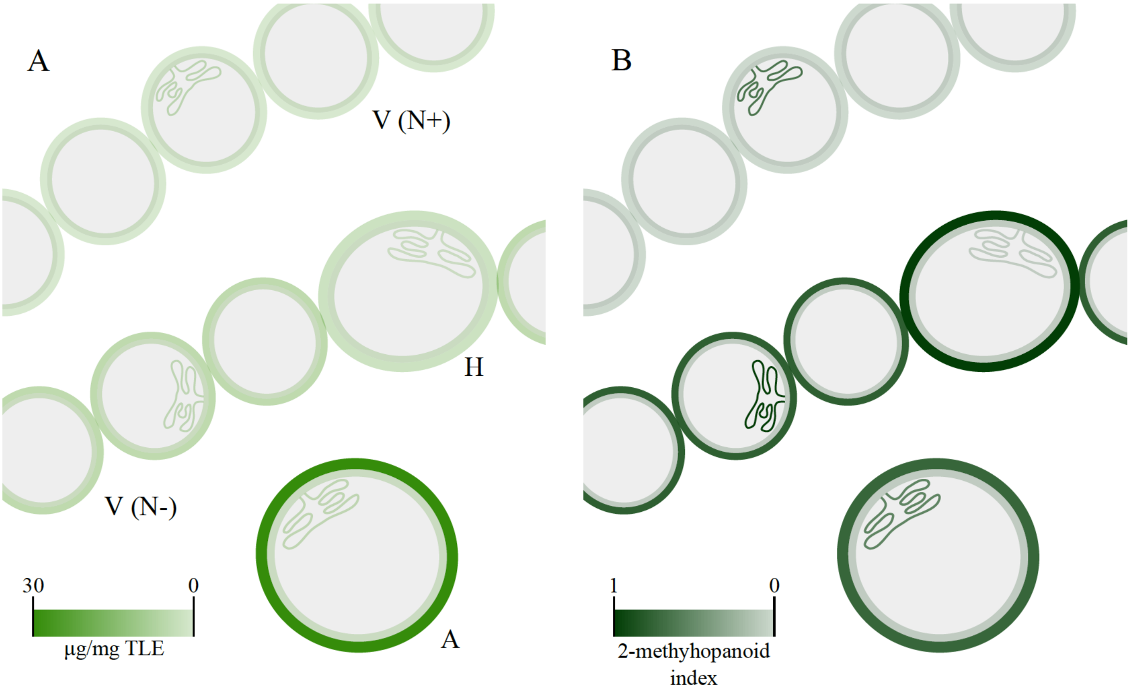


Figure 1.3 | Quantification of 2-methylhopanoids in specific membranes of *N. punctiforme*. The highest amount of 2-methylhopanods was found in the outer membranes of akinetes, while the highest 2-methylhopanoid index was measured in the outer membranes of heterocysts and the thylakoid membranse of vegetative cell grown without fixed nitrogen. The concentration of 2-methylhopanoids was measured as (A) μg of 2-methylhopanoids per mg of total lipid extract and (B) the 2-methylhopanoid index (2-methylhopanoids / 2-methylhopanoids + desmethylhopanoids). Data on the abundance of hopanoids in *N. punctiforme* are from Doughty et al. (2009). Abbreviations include V (N+) vegetative cells grown in the presence of fixed nitrogen, V (N-) vegetative cells grown without fixed nitrogen, H heterocysts, and A akinetes.

The correlation between 2-methylhopanoids and nitrogen fixation in modern organisms is connected to the rock record through OAEs, episodes correlated with high 2-methylhopane indices. Isotope data suggest that the nitrogen cycle was disrupted during OAEs, enhancing the need for fixed nitrogen, and causing a proliferation of nitrogen fixing bacteria (Kuypers et al. 2004; Knoll et al. 2007). The positive relationship between high 2-methylhopane indices and augmented nitrogen fixation, as well as the association between 2-methylhopanoids and nitrogen fixation made in contemporary biological systems, calls for future study in this area.

Questions addressed in this thesis

The questions addressed in the following chapters arise from the state of knowledge about 2-methylhopanoids as presented here and seek to form a robust interpretation for 2-methylhopanes in the rock record. These specific questions and brief rationales of how they are answered are found below:

- Who makes 2-methylhopanoids in modern environments? Are there habitats with more 2-methylhopanoid producers? To answer these questions, we take a two-pronged approach: surveying *hpnP* in metagenomes and amplifying *hpnP* by PCR in targeted environments. This allows us to identify which organisms are mostly likely to be producing 2-methylhopanoids based on their genetic capacity. (Chapter 2)

- What is the evolutionary history of 2-methylhopanoids? In which group of organisms did this capacity evolve? In this chapter, we reconstruct the phylogenetic tree of *hpnP* to identify the taxa where the gene initially evolved. The *hpnP* phylogeny is also used to date the emergence of 2-methylhopanoids relative to other geological and biological events. (Chapter 3)

- What are the roles of hopanoids and 2-methylhopanoids in the cyanobacterium *N. punctiforme*, a filamentous plant symbiont? Here, we use genetic methods to knock out hopanoid biosynthetic genes of interest. Defects in the resulting mutants are identified in the presence of particular stresses to provide insight into the biological function of hopanoids in *N. punctiforme*. These findings can be applied more generally to Cyanobacteria, in which the role of hopanoids has been poorly studied. (Chapter 4)

References

Andrusevich VE, Engel MH, Zumberge JE (2000) Effects of paleolatitude on the stable carbon isotope composition of crude oils. *Geology*, **28**, 847–850.

Asara JM, Schweitzer MH, Freimark LM, Phillips M, Cantley LC (2007) Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry. *Science*, **316**, 280–285.

Bednarczyk A, Hernandez TC, Schaeffer P, Adam P, Talbot HM, Farrimond P, Riboulleau A, Largeau C, Derenne S, Rohmer M, Albrecht P (2005) 32,35-Anhydrobacteriohopanetetrol: An unusual bacteriohopanepolyol widespread in recent and past environments. *Organic Geochemistry*, **36**, 673–677.

Bekker A, Holland HD, Wang P-L, Rumble D, Stein HJ, Hannah JL, Coetzee LL, Beukes NJ (2004) Dating the rise of atmospheric oxygen. *Nature*, **427**, 117–120.

Berry AM, Harriott OT, Moreau RA, Osman SF, Benson DR, Jones AD (1993) Hopanoid lipids compose the Frankia vesicle envelope, presumptive barrier of oxygen diffusion to nitrogenase. *Proceedings of the National Academy of Sciences of the United States of America*, **90**, 6091–6094.

Blumenberg M, Krüger M, Nauhaus K, Talbot HM, Oppermann BI, Seifert R, Pape T, Michaelis W (2006) Biosynthesis of hopanoids by sulfate-reducing bacteria (genus *Desulfovibrio*). *Environmental Microbiology*, **8**, 1220–1227.

Bradley AS, Pearson A, Sáenz JP, Marx CJ (2010) Adenosylhopane: The first intermediate in hopanoid side chain biosynthesis. *Organic Geochemistry*, **41**, 1075–1081.

Brocks JJ (2011) Millimeter-scale concentration gradients of hydrocarbons in Archean shales: Live-oil escape or fingerprint of contamination? *Geochimica et Cosmochimica Acta*, **75**, 3196–3213.

Brocks JJ, Banfield J (2009) Unravelling ancient microbial history with community proteogenomics and lipid geochemistry. *Nature Reviews Microbiology*, **7**, 601–609.

Brocks JJ, Logan GA, Buick R, Summons RE (1999) Archean Molecular Fossils and the Early Rise of Eukaryotes. *Science*, **285**, 1033–1036.

Brocks JJ, Love GD, Summons RE, Knoll AH, Logan GA, Bowden SA (2005) Biomarker evidence for green and purple sulphur bacteria in a stratified Palaeoproterozoic sea. *Nature*, **437**, 866–870.

Brocks JJ, Summons RE, Buick R, Logan G a. (2003) Origin and significance of aromatic hydrocarbons in giant iron ore deposits of the late Archean Hamersley Basin, Western Australia. *Organic Geochemistry*, **34**, 1161–1175.

Clarke JT, Warnock RCM, Donoghue PCJ (2011) Establishing a time-scale for plant evolution. *New Phytologist*, **192**, 266–301.

Cooke M, Talbot HM, Wagner T (2008) Tracking soil organic carbon transport to continental margin sediments using soil-specific hopanoid biomarkers: A case study from the Congo fan (ODP site 1075). *Organic Geochemistry*, **39**, 965–971.

Doughty DM, Hunter R, Summons RE, Newman DK (2009) 2-Methylhopanoids are maximally produced in akinetes of *Nostoc punctiforme*: geobiological implications. *Geobiology*, **7**, 524–532.

Eickhoff M, Birgel D, Talbot HM, Peckmann J, Kappler A (2013) Oxidation of Fe(II) leads to increased C-2 methylation of pentacyclic triterpenoids in the anoxygenic phototrophic bacterium *Rhodopseudomonas palustris* strain TIE-1. *Geobiology*, **11**, 268–278.

Elvert M, Niemann H (2008) Occurrence of unusual steroids and hopanoids derived from aerobic methanotrophs at an active marine mud volcano. *Organic Geochemistry*, **39**, 167–177.

Farrimond P, Talbot HM, Watson DF, Schulz LK, Wilhelms a. (2004) Methylhopanoids: Molecular indicators of ancient bacteria and a petroleum correlation tool. *Geochimica et Cosmochimica Acta*, **68**, 3873–3882.

Fischer WW, Summons RE, Pearson A (2005) Targeted genomic detection of biosynthetic pathways: anaerobic production of hopanoid biomarkers by a common sedimentary microbe. *Geobiology*, **3**, 33–40.

Forster BHJ, Biemann K, Haigh G, NH T, JR C (1973) The Structure of Novel C35 Pentacycic Terpenes from *Acetobacter xylinum* C3H502 C9Hls C11H17. *Biochemical Journal*, **135**, 133–143.

Freeman KH, Hayes JM, Trendel JM, Albrecht P (1990) Evidence from carbon isotope measurements for diverse origins of sedimentary hydrocarbons. *Nature*, **343**, 254–256.

French KL, Hallman C, Hope JM, Schoon PL, Zumberge JA, Hoshino Y, Peters CA, George SC, Love GD, Brocks JJ, Buick R, Summons RE (2015) Reappraisal of hydrocarbon biomarkers in Archean rocks. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 5915-5920.

Frickey T, Kannenberg E (2009) Phylogenetic analysis of the triterpene cyclase protein family in prokaryotes and eukaryotes suggests bidirectional lateral gene transfer. *Environmental Microbiology*, **11**, 1224–1241.

Härtner T, Straub KL, Kannenberg E (2005) Occurrence of hopanoid lipids in anaerobic *Geobacter* species. *FEMS Microbiology Letters*, **243**, 59–64.

Hays LE, Grice K, Foster CB, Summons RE (2012) Biomarker and isotopic trends in a Permian-Triassic sedimentary section at Kap Stosch, Greenland. *Organic Geochemistry*, **43**, 67–82.

Knoll AH, Summons RE, Waldbauer JR, Zumberge JE (2007) The Geological Succession of Primary Producers in the Oceans. In: *The Evolution of Primary Producers in the Sea* (eds Falkowski P, Knoll AH), pp. 133–164. Academic Press, Boston.

Kuypers MMM, van Breugel Y, Schouten S, Erba E, Sinninghe Damsté JS (2004) N2-fixing cyanobacteria supplied nutrient N for Cretaceous oceanic anoxic events. *Geology*, **32**, 853–856.

Love GD, Grosjean E, Stalvies C, Fike DA, Grotzinger JP, Bradley AS, Kelly AE, Bhatia M, Meredith W, Snape CE, Bowring SA, Condon DJ, Summons RE (2009) Fossil steroids record the appearance of Demospongiae during the Cryogenian period. *Nature*, **457**, 718–721.

Moldowan JM, Seifert WK, Gallegos EJ (1985) Relationship between petroleum composition and depositional enviornment of petroleum source rocks. *The American Association of Petroleum Geologists Bulletin*, **69**, 1255–1268.

Nalin R, Putra SR, Domenach A, Rohmer M, Berry AM (2000) High hopanoid/total lipids ratio in Frankia mycelia is not related to the nitrogen status. *Microbiology*, **146**, 3013–3019.

Noble R, Alexander R, Kagi RI (1985) The occurrence of bisnorhopane, trisnorhopane and 25-norhopanes as free hydrocarbons in some Australian shales. *Organic Geochemistry*, **8**, 171–176.

Ourisson G, Albrecht P (1992) Hopanoids. 1. Geohopanoids: the most abundant natural products on Earth? *Accounts of Chemical Research*, **25**, 398–402.

Pancost RD, Aquilina A, Talbot HM, Lim K, Evershed R, Bull I, Gill F, Weijers J, Collinson M, Taylor K (2010) Biomarkers for methane cycling: From marine to terrestrial settings. *Geochimica et Cosmochimica Acta*, **74**, A787.

Pearson A, Flood Page SR, Jorgenson TL, Fischer WW, Higgins MB (2007) Novel hopanoid cyclases from the environment. *Environmental Microbiology*, **9**, 2175–2188.

Pearson A, Leavitt WD, Sáenz JP, Summons RE, Tam MC-M, Close HG (2009) Diversity of hopanoids and squalene-hopene cyclases across a tropical land-sea gradient. *Environmental Microbiology*, **11**, 1208–1223.

Pedersen MW, Overballe-petersen S, Ermini L, Sarkissian C Der, Haile J, Hellstrom M, Spens J, Thomsen PF, Bohmann K, Cappellini E, Schnell IB, Wales NA, Carøe C, Campos F, Schmidt AMZ, Gilbert MTP, Hansen AJ, Orlando L, Willerslev E (2015) Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society B*, **370**, 20130383.

Peters KE, Moldowan JM (1991) Effects of source, thermal maturity, and biodegradation on the distribution and isomerization of homohopanes in petroleum. *Organic Geochemistry*, **17**, 47–61.

Peters K, Walter C, Moldowan JM (2005) *The Biomarker Guide*. Cambridge University Press, United Kingdom.

Rashby SE, Sessions AL, Summons RE, Newman DK (2007) Biosynthesis of 2-methylbacteriohopanepolyols by an anoxygenic phototroph. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 15099–15104.

Rasmussen B, Fletcher IR, Brocks JJ, Kilburn MR (2008) Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature*, **455**, 1101–1104.

Rohmer M, Bouvier P, Ourisson G (1979) Molecular Evolution of Biomembranes: Structural Equivalents and Phylogenetic Precursors of Sterols. *Proceedings of the National Academy of Sciences of the United States of America*, **76**, 847–851.

Rohmer M, Bouvier-Nave P, Ourisson G (1984) Distribution of Hopanoid Triterpenes in Prokaryotes. *Journal of General Microbiology*, **130**, 1137–1150.

Rohmer M, Ourisson G (1976) Structure of bacteriohopanoetetrols from *Acetobacter xylinum*. *Tetrahedron Letters*, 3633–3636.

Rosing M (1999) 13 C-depleted carbon microparticles in >3700-Ma sea-floor sedimentary rocks from West Greenland. *Science*, **283**, 674–676.

Sáenz JP, Eglinton TI, Summons RE (2011a) Abundance and structural diversity of bacteriohopanepolyols in suspended particulate matter along a river to ocean transect. *Organic Geochemistry*, **42**, 774–780.

Sáenz JP, Wakeham SG, Eglinton TI, Summons RE (2011b) New constraints on the provenance of hopanoids in the marine geologic record: Bacteriohopanepolyols in marine suboxic and anoxic environments. *Organic Geochemistry*, **42**, 1351–1362.

Schweitzer MH, Zheng W, Organ CL, Avci R, Suo Z, Freimark LM, Lebleu VS, Duncan MB, Vander Heiden MG, Neveu JM, Lane WS, Cottrell JS, Horner JR, Cantley LC, Kalluri R, Asara JM (2009) Biomolecular characterization and protein sequences of the Campanian hadrosaur *B. canadensis*. *Science*, **324**, 626–631.

Silipo A, Vitiello G, Gully D, Sturiale L, Chaintreuil C, Fardoux J, Gargani D, Lee H-I, Kulkarni G, Busset N, Marchetti R, Palmigiano A, Moll H, Engel R, Lanzetta R, Paduano L, Parrilli M, Chang W-S, Holst O, Newman DK, Garozzo D, D’Errico G, Giraud E, Molinaro A (2014) Covalently linked hopanoid-lipid A improves outer-membrane resistance of a *Bradyrhizobium* symbiont of legumes. *Nature communications*, **5**, 5106.

Sinninghe Damsté JS, Schouten S (1997) Is there evidence for a substantial contribution of prokaryotic biomass to organic carbon in Phanerozoic carbonaceous sediments? *Organic Geochemistry*, **26**, 517–530.

Summons RE, Jahnke LL, Hope JM, Logan GA (1999) 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature*, **400**, 554–557.

Talbot HM, Farrimond P (2007) Bacterial populations recorded in diverse sedimentary biohopanoid distributions. *Organic Geochemistry*, **38**, 1212–1225.

Talbot HM, Farrimond P, Schaeffer P, Pancost RD (2005) Bacteriohopanepolyols in hydrothermal vent biogenic silicates. *Organic Geochemistry*, **36**, 663–672.

Talbot HM, Watson DF, Pearson EJ, Farrimond P (2003) Diverse biohopanoid compositions of non-marine sediments. *Organic Geochemistry*, **34**, 1353–1371.

Thiel V, Blumenberg M, Pape T, Seifert R, Michaelis W (2003) Unexpected occurrence of hopanoids at gas seeps in the Black Sea. *Organic Geochemistry*, **34**, 81–87.

Welander P V, Coleman ML, Sessions AL, Summons RE, Newman DK (2010) Identification of a methylase required for 2-methylhopanoid production and implications for the interpretation of sedimentary hopanes. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 8537–8542.

Welander P V, Doughty DM, Wu C-H, Mehay S, Summons RE, Newman DK (2012) Identification and characterization of *Rhodopseudomonas palustris* TIE-1 hopanoid biosynthesis mutants. *Geobiology*, **10**, 163–177.

Welander P V, Summons RE (2012) Discovery, taxonomic distribution, and phenotypic characterization of a gene required for 3-methylhopanoid production. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 12905–12910.

Willerslev E, Hansen AJ, Binladen J, Brand TB, Gilbert MTP, Shapiro B, Bunce M, Wiuf C, Gilichinsky DA, Cooper A (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, **300**, 791–795.

Van Winden JF, Talbot HM, Kip N, Reichart GJ, Pol A, McNamara NP, Jetten MSM, Op den Camp HJM, Sinninghe Damsté JS (2012) Bacteriohopanepolyol signatures as markers for methanotrophic bacteria in peat moss. *Geochimica et Cosmochimica Acta*, **77**, 52–61.

Wu C-H, Bialecka-Fornal M, Newman DK (2015) Methylation at the C-2 position of hopanoids increases rigidity in native bacterial membranes. *eLife*, 10.7554/eL.

Xie S, Pancost RD, Yin H, Wang H, Evershed RP (2005) Two episodes of microbial change coupled with Permo/Triassic faunal mass extinction. *Nature*, **434**, 494–497.

Xu Y, Cooke M, Talbot HM, Simpson M (2009) Bacteriohopanepolyol signatures of bacterial populations in Western Canadian soils. *Organic Geochemistry*, **40**, 79–86.

Zhu C, Talbot HM, Wagner T, Pan JM, Pancost RD (2010) Intense aerobic methane oxidation in the Yangtze Estuary: A record from 35-aminobacteriohopanepolyols in surface sediments. *Organic Geochemistry*, **41**, 1056–1059.

Zhu C, Talbot HM, Wagner T, Pan J-M, Pancost RD (2011) Distribution of hopanoids along a land to sea transect: Implications for microbial ecology and the use of hopanoids in environmental studies. *Limnology and Oceanography*, **56**, 1850–1865.

Zundel M, Rohmer M (1985) Prokaryotic triterpenoids: 3. The biosynthesis of 2B-methylhoanoids and 3B-methylhopanoids of *Methylobacterium organophilum* and *Scetobacteri pasteurianus* ssp. *pasteurianus*. *European Journal of Biochemistry*, **150**, 35–39.

Chapter 2

DIVERSE CAPACITY FOR 2-METHYLHOPANOID PRODUCTION CORRELATES WITH A SPECIFIC ECOLOGICAL NICHE

This chapter was first published as, Ricci JN, Coleman ML, Welander PV, Sessions AL, Summons RE, Spear JR, and Newman D (2014) Diverse capacity for 2-methylhopanoid production correlates with a specific ecological niche. *The ISME Journal*, **8**, 675-684.

Abstract

Molecular fossils of 2-methylhopanoids are prominent biomarkers in modern and ancient sediments that have been used as proxies for cyanobacteria and their main metabolism, oxygenic photosynthesis. However, substantial culture and genomic-based evidence now indicates that organisms other than cyanobacteria can make 2-methylhopanoids. Because little data directly addresses which organisms produce 2-methylhopanoids in the environment, we used metagenomic and clone library methods to determine the environmental diversity of *hpnP*, the gene encoding the C-2 hopanoid methylase. Here we show that *hpnP* copies from alphaproteobacteria and as yet uncultured organisms are found in diverse modern environments, including some modern habitats representative of those preserved in the rock record. In contrast, cyanobacterial *hpnP* genes are rarer and tend to be localized to specific habitats. To move beyond understanding the taxonomic distribution of environmental 2-methylhopanoid producers, we asked whether *hpnP* presence might track with particular variables. We found *hpnP* to be significantly correlated with organisms, metabolisms, and environments known to support plant-microbe interactions (p-value $< 10^{-6}$); in addition, we observed diverse *hpnP* types in closely-packed microbial communities from other environments, including stromatolites, hot springs, and hypersaline microbial mats. The common features of these niches indicate that 2-methylhopanoids are enriched in sessile microbial

communities inhabiting environments low in oxygen and fixed nitrogen with high osmolarity. Our results support the earlier conclusion that 2-methylhopanoids are not reliable biomarkers for cyanobacteria or any other taxonomic group, and raise the new hypothesis that, instead, they are indicators of a specific environmental niche.

Introduction

Morphological and molecular fossils left by microorganisms in ancient sedimentary rocks can provide a valuable window into the early history of life on Earth. Yet due to challenges inherent in working with billion-year old samples, the interpretation of these fossils has often been contentious (Schopf & Packer, 1987; Walter et al., 1992; Brasier et al., 2002, 2004). In this context, organic biomarkers have received attention due to their potential to provide more specific information about the composition of ancient microbial communities (Brocks & Pearson, 2005). Hopanes and steranes are among the more prominent classes of these biomarkers (Rohmer, 2010). These molecules can unambiguously be interpreted as the diagenetic remains of hopanoids and steroids, polycyclic triterpenoids found in the membranes of numerous organisms today (Rohmer et al., 1984; Ourisson et al., 1987). However, ambiguity regarding their distribution and function in modern bacteria clouds our ability to interpret their fossils in the rock record.

Hopanoids structurally resemble steroids, but unlike steroids they are primarily made by bacteria and do not require oxygen for their biosynthesis (Ourisson et al., 1987; Fischer et al., 2005; Rashby et al., 2007). Although hopanoids exhibit structural diversity of their side chains, much of this diversity is lost through diagenesis, resulting mainly in the preservation of their hydrocarbon skeletons, hopanes. Important exceptions to this are methyl groups at C-2 or C-3, which can be preserved. In fact, 2-methylhopanes have a rich history in the fossil record, found at discrete times and locations as far back as 2.7 billion years (Brocks et al., 1999), although this latter finding is under scrutiny (Rasmussen et al., 2008). The varied distribution of 2-methylhopanes in the more "recent" rock record (i.e., million year time scales), showing peaks in abundance correlated with ocean anoxic events, suggests their production may be linked to particular environmental triggers (Knoll et al., 2007).

Until recently, 2-methylhopanes were viewed as biomarkers for cyanobacteria and their main energy-generating metabolism, oxygenic photosynthesis (Summons et al., 1999). However, the finding of conditional 2-methylhopanoid production by the anoxygenic phototroph *Rhodopseudomonas palustris* called this interpretation into question (Rashby et al., 2007). From genomic data and culture-based work it is now clear that only a minority of cyanobacteria (i.e., 13% of all sequenced cyanobacterial species and 19% of all sequenced cyanobacterial genera) have the gene, *hpnP*, that encodes the enzyme responsible for methylating hopanoids at C-2 (Figure 2.1) (Welander et al., 2010, Talbot et al., 2008). Moreover, other bacterial species possess *hpnP*, including many from a subclade of alphaproteobacteria and an acidobacterium (Welander et al., 2010, Talbot et al., 2007a).

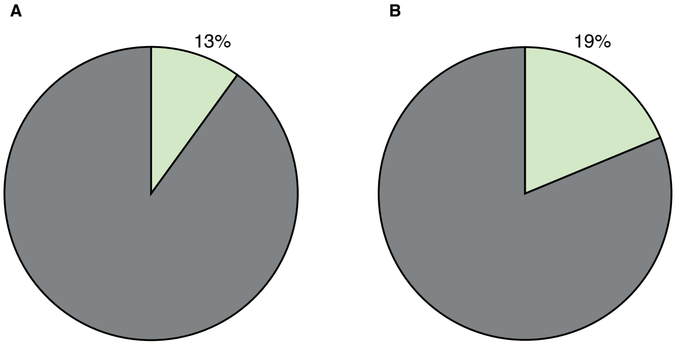


Figure 2.1 | Distribution of *hpnP* in cyanobacteria. Genomes with *hpnP* are represented in green and without *hpnP* in gray. (A) Distribution of *hpnP* among all finished cyanobacterial genomes on IMG. (B) Distribution of *hpnP* among cyanobacterial genera to account for biased sequencing among genomes. If one member of a genus had *hpnP* that genus was counted as having the gene.

Despite the known distribution of *hpnP* among cultured organisms, uncertainty remains about which producers of 2-methylhopanoids are environmentally relevant and whether there is a common ecology that correlates with the production of these molecules. Given that 2-methylhopanoids are produced by diverse bacteria, we asked two questions: 1) what is the potential for 2-methylhopanoid production by different organisms in modern environments, and 2) might 2-methylhopanoid producers inhabit common ecological niche(s)? Here we assess the distribution of *hpnP* in various environments, and find a statistically significant correlation with modern habitats that support plant-microbe interactions. We discuss the implications of these results for interpreting the ancient 2-methylhopane record.

Methods

Distribution of hpnP in cyanobacterial genomes

The abundance of *hpnP* genes among cyanobacteria was calculated using finished cyanobacterial genomes on the Joint Genomes Institute's Integrated Microbial Genomes (IMG) database (Markowitz et al., 2011). Additionally, we analyzed this dataset condensed to the genus level to reduce bias (Figure 2.1).

Distribution of hpnP in metagenomes

To retrieve HpnP sequences from public metagenomes, HpnP from *R. palustris* TIE-1 (NC_011004.1) was used as a query against the NCBI metagenomic proteins database, IMG/M, CAMERA, and myMGDB. All hits with an e-value equal to or less than $1x10^{-50}$ were subjected to phylogenetic analysis (Markowitz et al., 2011; Sun et al., 2011). Approximately 20% of hits retrieved clustered phylogenetically with known HpnP sequences, while all others clustered with sequences known to not be *hpnP*. Only sequences that cluster with known *hpnP* genes were identified as *hpnP* (Dataset S1). All searches were completed in December 2012.

Sequences were identified as *gyrB*, *psbC*, and *shc* if they had an e-value equal to or less than 1x10^{-20}. *Escherichia coli gyrB* (NC_000913.2), *N. punctiforme psbC* (YP_001866969.1) and *R. palustris* TIE-1 *shc* (NC_011004.1) were used as query sequences. This e-value was used because it captured known diversity of the genes without retrieving related sequences with other functions. Sequences of shc were determined to be cyanobacterial or alphaproteobacterial by top BLASTP hit (Table 2.1).

Clone library and sample preparation

DNA samples from Yellowstone National Park hot springs were collected and prepared as previously reported (Osburn et al., 2011). All other samples were extracted with the UltraClean Soil DNA Isolation Kit (MoBio, Carlsbad, CA, USA). DNA samples were stored at -20ºC. Information on samples can be found in Appendix A.

Nested PCR primers for *hpnP* were designed based on the conserved amino acid motifs (A/G)FMPPQ and (S/T)GII(L/M)G for the first pair, and (A/V)(L/I)GGPS and GIETP(E/D) for the second. The resulting primers were hpnP_1F 5'-GSB TTY ATG CCD CCB CAR GG, hpnP_1R 5'-TCN ARK CCV AKR ATR ATN CC, hpnP_2F 5'-GYB VTB GGH GGN CCN TCN GT, and hpnP_2R 5'-TCN GGN GTY TCD ATN CC, respectively (Figure 2.2A). To amplify *hpnP*, Promega PCR master mix (Promega, Madison, WI, USA) and cycles of 95°C for 2 min, 95°C denaturation for 30 s, 50°C for 30 s, and 72°C for 1 min for 35 cycles, followed by 72°C for 3 min were used. An aliquot the first PCR (1 µl) was used as a template for the second. We validated this method by testing the primer sets on diverse 2-methylhopanoid-producing organisms as well as some that do not make 2-methylhopanoids (Figure 2.2C).

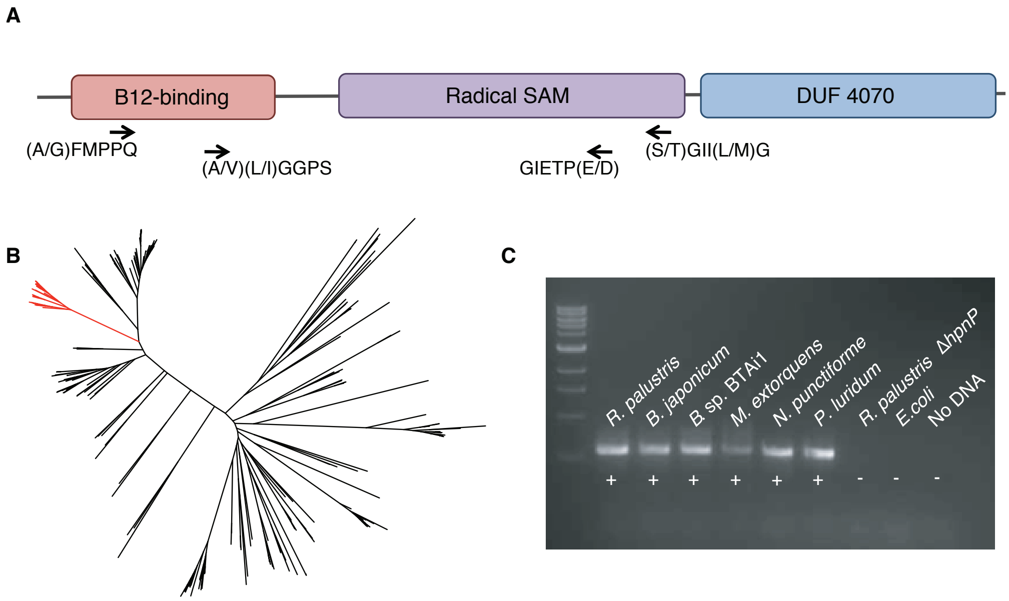


Figure 2.2 | Design of *hpnP* degenerate PCR primers. (A) Protein domain structure of HpnP determined by InterProScan and location of degenerate PCR primers with corresponding protein consensus sequences. HpnP is 527 amino acids in length. The domains B12 binding, radical SAM, and DUF4070 are 105, 162, and 140 amino acids in length, respectively, in *R. palustris* TIE-1. (B) Phylogeny of HpnP, red, and related homologous proteins. (C) Control PCR reactions for degenerate *hpnP* primers with genomic DNA from diverse cultured 2-methylhopanoid producers and negative controls.

PCR products from the second reaction were extracted with the Montage gel extraction kit (Millipore, Billerica, MA) and cloned with the TOPO TA Cloning® Kit for sequencing using One Shot Top10 electrocompetent cells (Invitrogen, Carlsbad, CA, USA). 24 to 100 clones per sample were amplified by PCR and restriction digested with AluI (New England Biolabs, Ipswich, MA, USA). Amplicons with unique digestion patterns were sequenced (Retrogen, San Diego, CA, USA). Sequences were trimmed to remove contaminating vector and poor quality regions, and then translated in Geneious 5.6.5. Representative sequences of 95% identity clusters were picked using CD-HIT and used to make phylogenetic trees (Huang et al., 2010). Sequences have been deposited in GenBank under the accession numbers KC603770 thru KC603846.

Rarefaction curves

The number of unique restriction digestion patterns was used as a proxy for *hpnP* diversity. To generate rarefaction curves, the species observed metric was used in the alpha diversity package of QIIME 1.5.0 (Caporaso et al., 2010). Datasets were rarefied 100 times in 2-step increments (Figure 2.3). To compare richness of *hpnP* between clone libraries, datasets were rarefied to 25 clones to avoid depth of sampling bias and averaged (Figure 2.3). Guerrero Negro sample 4 was not included in this analysis because of low sampling coverage (Appendix A).

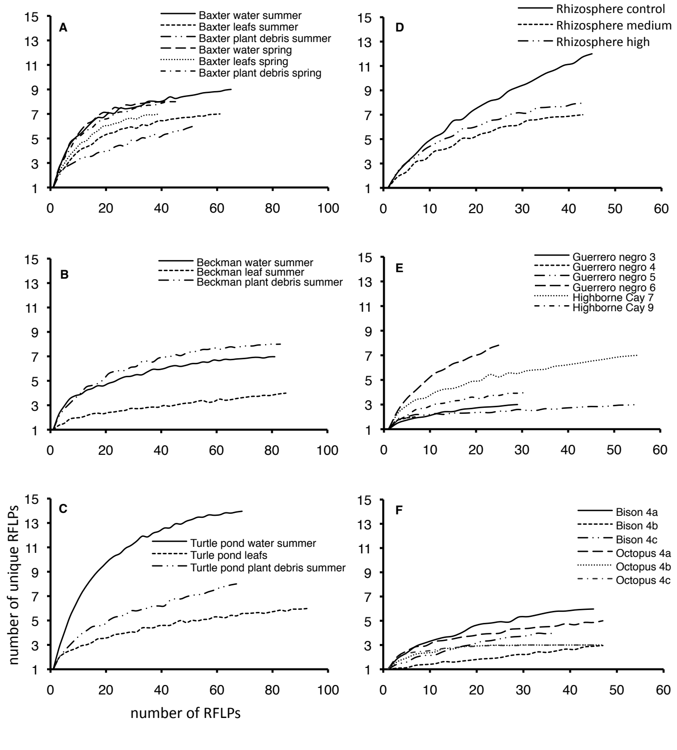


Figure 2.3 | Rarefaction curves for select *hpnP* clone libraries. RFLP patterns confirmed to be *hpnP* by sequencing were used to generate rarefaction curves in QIIME using the observed species metric and rarefied 100 times in increments of 2. Rarefaction curves are grouped by environment and average richnesses (± standard deviation) were calculated when data was rarefied to 25 RFLP patterns (A) Baxter pond 6.4 ± 1.7 (B) Beckman pond 4.5 ± 1.2 (C) Turtle pond 6.6 ± 3.5 (D) Rhizosphere 7 ± 1.2 (E) Guerrero negro 4.4 ± 3 and Highborne Cay 4.5 ± 1 (F) selected hot spring microbial mats 3.4 ± 1.

Alignment and phylogeny construction

All alignments were made using the MAFFT v6.859b l-ins-i algorithm (Katoh & Toh, 2008). Reference HpnP alignments comprised all HpnP sequences retrieved from NCBI as of December 2012 and *hpnP* from *Phormidium luridum* UTEX 426. Full length *P. luridum* UTEX 426 *hpnP* was obtained by inverse PCR using the degenerate *hpnP* PCR primers as a probe. Outgroup sequences were picked from the sister clade of HpnP (Welander et al., 2010). The reference alignment was trimmed in Gblocks 0.91b with relaxed parameters (Talavera & Castresana, 2007). Environmental HpnP sequences were then added to the reference HpnP alignments using the seed option in MAFFT. Phylogenetic trees were made by PhyML v3.0 using the LG model with aLRT supports and modified in iTOL (Guindon et al., 2010; Letunic & Bork, 2007).

HpnP types (cyanobacterial, alphaproteobacterial, or unknown) from metagenomic and clone library sequences were classified as such when they grouped phylogenetically with reference HpnP sequences of the same type (Figure 2.4, Appendix A). Unknown *hpnP* groups were defined by a lack of reference sequences. Metagenomic *hpnP* hits with long branches were examined for recombination events using Recombination Analysis Tool, but none were found (Etherington et al., 2005).

hpnP correlation with plants

The initial observation that many *hpnP*-containing organisms and metagenomes were plant-associated used the following criteria for a positive plant-association: the organism was isolated from a plant-associated environment, had a known plant interaction established in the literature, or the metagenome was from a plant derived environment (that is, soil and rhizosphere, wood compost, and insect fungal gardens, as these are maintained by leaf-cutting ants) based on available metadata (Figure 2.4 outer ring).

To assess the significance of the relationship between *hpnP* or *shc* and plant-associated organisms or environments, we used the hypergeometric test to evaluate if there was non-random overlap between organisms or environments that have *hpnP* or *shc* and those that are plant associated (Table 2.2). Plant-association for metagenomes was determined as described above. When addressing this analysis among organisms, we included in our analysis all finished bacterial genomes condensed to the species level on IMG. Organisms were counted as plant-associated if a plant species was listed under "host name" in the description of the genome. Since filling out this data field is voluntary, we would expect more false negatives than false positives. In an attempt to reduce the number of false negatives, we mined Pubmed for abstracts describing plant-associations among our list of organisms. To conduct an unbiased search, we used the following Boolean expression: two word name of the species AND host plant OR plant host OR plant-microbe OR root-coloniz* OR plant-associat* NOT pathogen, where * allows for multiple endings. Abstract hits were manually annotated and positive hits were combined with the plant-associated list from IMG.

The hypergeometric test was also used to assess an *hpnP* correlation with *nifD* and *moxF* in finished bacterial genomes from IMG (Table 2.3). Genomes were found to have *nifD* or *moxF* if they returned an e-value less than or equal to $1x10^{-50}$ when using *Nostoc* sp. PCC 7120 *nifD* (gi 4376092) and *Methylobacterium extorquens* AM1 *moxF* (YP_002965446) as queries. Similar results for *nifD* were also obtained for *nifH*.

Hopanoid analysis

Select samples were targeted for analysis by LC-MS for a limited number of hopanoids. These samples appear in Figure 2.5 with either an identifier if hopanoids were analyzed, or blank if hopanoids were not analyzed. We also attempted to quantify hopanoids, specifically unextended hopanoids that cannot be identified by LC-MS, using standard GC-MS techniques (Sessions et al., 2013), but we were unable to unambiguously detect hopanoids due to a high background. Samples were extracted as previously reported (Sessions et al., 2013).

Methylated and non-methylated bacteriohopanepolyols were identified by liquid chromatography-mass spectrometry (LC-MS) as previously described (Welander et al., 2012). Lipids were acetylated by incubating total lipid extract (TLE, 1 mg) from each sample in a 1:1 (v:v, 250 ml) mixture of acetic anhydride (Sigma-Aldrich, St. Louis, MO, USA) and pyridine (Sigma-Aldrich) for 1 hour at 70°C. Acetylated TLEs were dried down under a stream of N_2 and resuspended in methanol (1 ml) for a final TLE concentration of 1 mg/ml. Subsequently, each sample (5 ml) were loaded onto the LC-MS for analysis, a 1200 Series HPLC (Agilent Technologies, Santa Clara, CA) equipped with an autosampler and a binary pump linked to a Q-TOF 6520 mass spectrometer (Agilent Technologies) via an atmospheric pressure chemical ionization (APCI) interface (Agilent Technologies) operated in positive ion mode. The analytical procedure was adapted from (Talbot et al., 2001). A Poroshell 120 EC-C18 column (2.1 x 150 mm, 2.7 µm; Agilent Technologies), set at 30°C, was eluted isocratically first with MeOH/water (95:5, v:v) for 2 min at a flow rate of 0.15 ml/min, then using a linear gradient up to 20% (v) of isopropyl alcohol (IPA) over 18 min at a flow rate of 0.19 ml/min, and isocratic for 10 min. The linear gradient was then set to 30% (v) of IPA at 0.19 ml/min over 10 min, and maintained for 5 min. The column was subsequently eluted using a linear gradient up to 80% IPA (v) over 1 min at a flow rate of 0.15 ml/min and isocratic for 14 min. Finally the column was eluted with MeOH/water (95:5, v:v) at 0.15 ml/min for 5 min. The APCI parameters were as follows: gas temperature 325°C, vaporizer temperature 350°C, drying gas (N_2) flow 6 l/min, nebulizer (N_2) flow 30 l/min, capillary voltage 1200 V, corona needle 4 µA, fragmentor 150 V. Data were recorded by scanning from m/z 100 to 1600. Bacteriohopanepolyols were identified on the basis of accurate mass measurements of their protonated molecular ions, fragmentation patterns in MS-MS mode and by comparison of relative retention time and the mass spectra with published data (Talbot et al., 2007b; Talbot et al., 2003a, 2003b).

Results and Discussion

Environmental distribution of hpnP

To identify potential biological sources of environmental 2-methylhopanoids, we followed a two-pronged approach: 1) we searched all available metagenomes for the presence of *hpnP* (Figure 2.4, Table 2.1, Appendix A) and 2) we generated clone libraries of *hpnP* sequences from diverse environments (Figure 2.5, Appendix A). In the first survey, 59 metagenomes were found to have *hpnP*, which resulted in the identification of 139 partial *hpnP* sequences (Figure S2, Appendix A). We also searched the same metagenomes for *shc*, which encodes squalene hopene cyclase, the enzyme that catalyzes the first step in hopanoid biosynthesis. For the second approach, we designed degenerate PCR primers that amplify all known diversity of *hpnP* (Figure 2.2) and were used to retrieve 76 unique *hpnP* sequences from 62 samples (Appendix A). Due to the potential for PCR bias, we did not infer the abundance of any particular *hpnP* type within a given library. However, we did estimate the abundance of a particular *hpnP* type in an environment by counting the number of samples from that environment which contained that specific *hpnP* type (Figure 2.5). While we cannot exclude the possibility that horizontal gene transfer confounds our taxonomic assignment of *hpnP* sequences, there is no evidence of recent transfer events in the phylogeny of *hpnP* (Welander et al., 2010).

Based on both metagenomic and clone library data, we found that cyanobacterial *hpnP* copies are not ubiquitous in most modern habitats, constituting only 4% of metagenomic *hpnP* sequences (Figure 2.5, Table 2.1). Consistent with this finding, we found low abundances of metagenomic cyanobacterial *shc* sequences in all environments as seen previously (Table 2.1) (Pearson et al., 2007, 2009; Pearson & Rusch, 2009). These data suggest that not only are cyanobacteria minor producers of 2-methylhopanoids, but that they do not contribute substantially to hopanoid production in general. However, we cannot rule out the possibility that rare members of a community may be disproportionately active hopanoid producers.

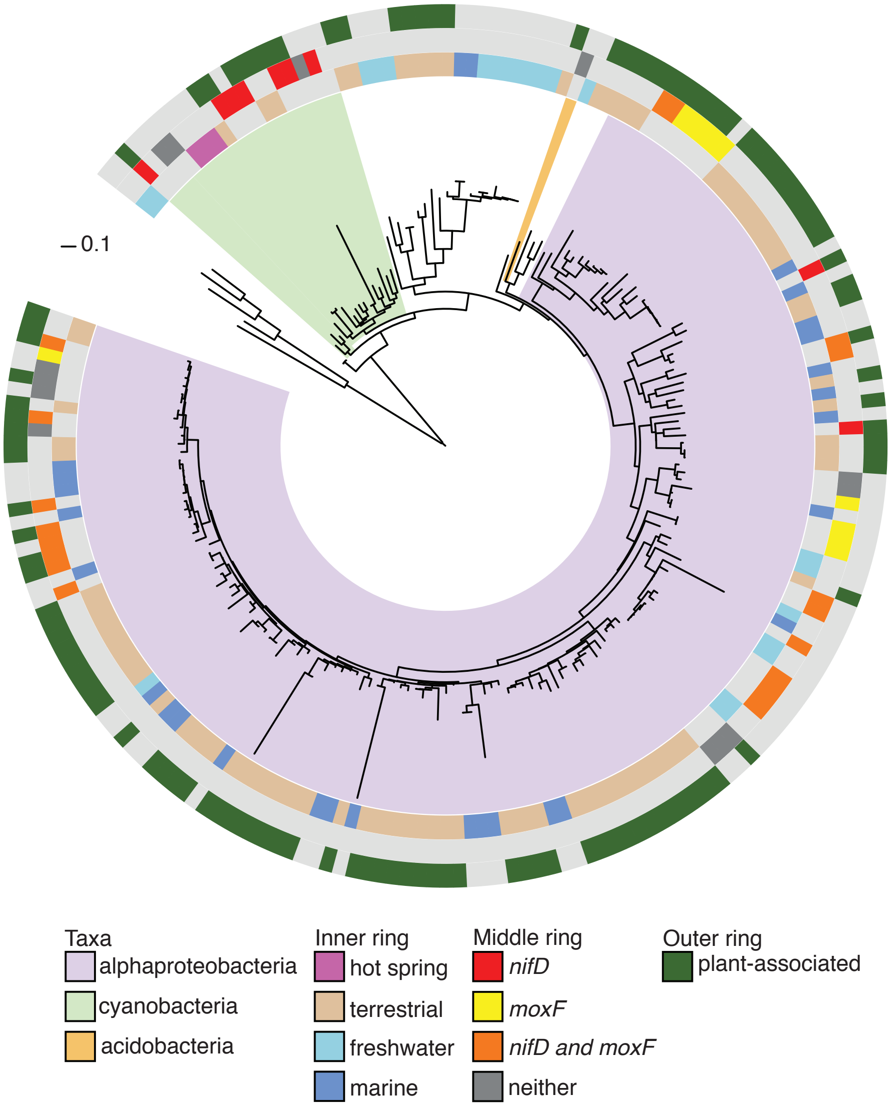


Figure 2.4 | HpnP diversity from metagenomes and its correlation with plant-microbe interactions. Metagenomic databases were searched for *hpnP*-like sequences. Sequences that could be phylogenetically classified as members of the HpnP family appear in this maximum likelihood HpnP phylogeny with HpnP sequences from genomes. The colored ranges on the tree's branches indicate alphaproteobacterial, cyanobacterial, or acidobacterial clades of HpnP determined by HpnP sequences from reference genomes (Appendix A). Metagenomic sequences that fall within one of these clades are classified as belonging to that taxon. In the inner ring surrounding the tree, HpnP sequences from metagenomes are colored by environment of origin, either hot spring, terrestrial, freshwater, or marine. In the middle ring, HpnP sequences from genomes are colored to indicate the presence of *nifD*, *moxF*, or both genes in the same genome. The outer ring indicates that the organism the sequence derives from was isolated from a plant-associated environment, has an established plant interaction, or that a metagenomic sequence is from a plant-associated environment (soil or rhizosphere, insect waste dumps, wood compost). The light grey background corresponds to no data in the inner two rings (e.g., genomes do not have an environment of origin color) or no plant-association found in the outer ring; the outgroup was not included in the analysis. aLRT support values and leaf names are shown in Appendix A. The scale bar is a measure of evolutionary distance equaling 0.1 substitutions per site.

Table 2.1 | Abundances of *hpnP* and *shc* in metagenomes

					hpnP					*shc*		
Environments[1]	Mbp	Reads	*gyrB*[2]	*psbC*[3]	Total	Cn	Al	Ac	Un	Total	Cn	Al
Hot springs	2543	6852165	1015	73	3	3	-	-	-	48	2	3
Terrestrial	44645	272985945	8562	126	87	3	74	-	10	1304	27	610
Soil and rhizosphere	41856	266263412	6037	125	80	3	68	-	9	1119	26	514
Insect fungal gardens	2376	5735261	2197	1	6	-	5	-	1	146	-	74
Wood compost	413	987272	328	-	1	-	1	-	-	39	1	22
Freshwater	58136	305595686	8810	200	21	-	10	-	11	554	16	49
Lentic and lotic	6852	17241435	5674	190	17	-	8	-	9	423	14	32
Groundwater	51172	288125962	2544	8	2	-	-	-	2	115	2	13
Wastewater	111	228289	592	2	2	-	2	-	-	16	-	4
Marine	203016	4793195841	13388	1759	28	-	26	-	2	488	4	192
Open ocean	4509	4359663	1142	534	-	-	-	-	-	48[4]	-[4]	24[4]
Coastal, upwelling, harbor	188571	4767074572	7805	491	2	-	-	-	2	137[4]	1[4]	37[4]
Estuary	2971	6629924	2405	10	9	-	9	-	-	183[4]	-[4]	87[4]
High latitude	1888	4405566	526	7	-	-	-	-	-	17	-	4
Ace Lake	1119	2465421	20	506	2	-	2	-	-	9	-	8
Deep sea hydrothermal vent	1221	2197778	424	3	15	-	15	-	-	14	-	4
Reef	627	597699	258	92	-	-	-	-	-	1[4]	-[4]	-[4]
Mangrove	153	148018	44	1	-	-	-	-	-	5[4]	-[4]	-[4]
Hypersaline	1411	3260701	490	50	-	-	-	-	-	71[4]	1[4]	27[4]
Equatorial upwelling	163	712274	102	30	-	-	-	-	-	-	-	-
Spring bloom	223	1064091	102	13	-	-	-	-	-	1	-	1
Trichodesmium bloom	159	280134	70	22	-	-	-	-	-	2	2	-

[1] Identifying information and descriptions of metagenomes included here appear in Dataset S2

[2] *gyrB*, estimates the number of bacterial genomes

[3] *psbC*, estimates the number of cyanobacterial genomes

[4] Some data from (Pearson & Rusch, 2009)

Abbreviations: Cn Cyanobacterial, Al Alphaproteobacterial, Ac Acidobacterial, Un Unknown

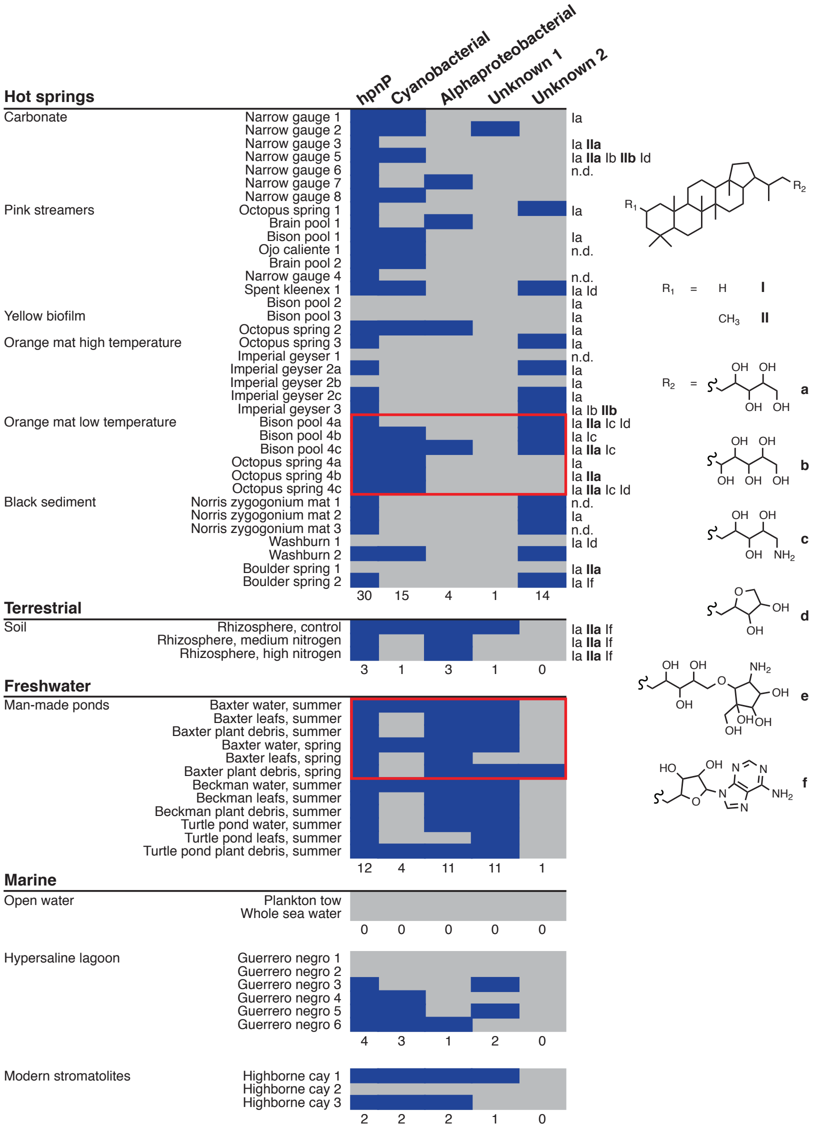


Figure 2.5 | Distribution of HpnP types in clone libraries and hopanoids from various environments. HpnP sequences generated from each sample were classified by HpnP type in a phylogenetic tree (Appendix A). The presence of HpnP or an HpnP type is indicated in blue. Total abundances of *hpnP* types are reported below each environment. A red outline highlights spatial and temporal heterogeneity. Sample descriptions appear in Appendix A. Presence of hopanoids was determined for select samples by LC-MS. Abbreviations include Ia bacteriohopanetetrol, IIa 2-methylbacteriohopanetetrol, Ib bacteriohopanepentol, IIb 2-methylbacteriohopanepentol, Ic aminobacteriohopnetriol, Id anhydrobacteriohopanetetrol, Ie adenosylhopane, and n.d. not detected. Other functionalized hopanoids, including aminohopanoids, were not detected. Methylated hopanoids (IIa and IIb) are highlighted in bold type.

Furthermore, some environmental *hpnP* sequences could not be identified as being from one of the known clades of 2-methylhopanoid producers. These sequences, which can be confidently identified as *hpnP* by homology, may represent new taxa of *hpnP*-containing organisms;

alternatively, as the database of *hpnP* sequences from cultured organisms grows, we may be able to assign these sequences to previously identified clades (Figure 2.4, Appendix A). Among clone library samples, the two unknown groups of *hpnP* segregate by environment of origin, with unknown group 1 found in most environments and unknown group 2 primarily in hot springs. 14 out of 36 hot spring clone libraries contained unknown *hpnP* sequences from group 2, nearly identical to the abundance of cyanobacterial *hpnP* sequences, 15 out of 36. Additionally, *hpnP* sequences belonging to unknown group 1 were found in 11 out of 12 pond samples (Figure 2.5). These observations indicate that potential novel groups of 2-methylhopanoid producing bacteria may play a major role in 2-methylhopanoid production in hot springs and freshwater environments.

Acknowledging that comparative metagenomic analysis can be biased by the samples that have been sequenced and the methods used to sequence them, we nevertheless found a robust pattern in the available data: the majority (63%) of metagenomic *hpnP* sequences belong to terrestrial habitats, such as soil, the rhizosphere, insect fungal gardens, and wood compost (Figure 2.4). This finding is not due to deeper sequencing of terrestrial environments. Using the abundance of *gyrB* to approximate the number of bacteria, because it is found in single copy in all bacterial genomes, we estimate that approximately 1% of terrestrial bacteria have *hpnP* whereas less than 0.4% of bacteria in other environments surveyed have *hpnP* (Table 2.1) (Biers et al., 2009). In comparing *hpnP* richness between clone library datasets, rhizosphere and pond samples were found to contain more unique *hpnP* genes than the other environments tested (Figure 2.3). Interestingly, 15% of terrestrial bacteria contain *shc*, while other habitats contain less than 7% of bacteria with *shc* (Table 2.1), implying that the same argument may be true for hopanoids in general. Taken together, these data suggest that terrestrial and freshwater environments harbor the majority of *hpnP* and *shc* diversity in modern ecosystems and should be considered likely sources of modern and possibly ancient (2-methyl)hopanoids. This finding is congruent with the presence of (2-methyl)hopanoids in terrestrial and freshwater settings (Pearson et al., 2009, Talbot & Farrimond, 2007, Cooke et al., 2008, Xu et al., 2009) and observations across a land-sea transect where (2-methyl)hopanoids appeared to be partially terrestrial in origin (Sáenz et al., 2011).

Paradoxically, the sedimentary context of the 2-methylhopane fossil record suggests ancient 2-methylhopanoid producers inhabited shallow, tropical marine environments (Knoll et al., 2007). It is noteworthy that our culture-independent search for *hpnP* sequences in all marine metagenomes, highly biased towards open water, coastal, and estuarine samples, did not identify any cyanobacterial *hpnP* sequences. This was not due to a lack of cyanobacteria, as 13% of bacteria in marine metagenomes were estimated to be cyanobacteria based on the abundance of *psbC,* which encodes a component of the photosynthetic machinery present at one copy per cyanobacterial genome (Table 2.1) (Mulkidjanian et al., 2006). Of those *hpnP* sequences identified, 93% were alphaproteobacterial and 7% were of unknown origin. These sequences derived from coastal waters, estuary sediments, Ace Lake (Antarctica), and deep-sea hydrothermal vents (Table 2.1). These data suggest that 2-methylhopanoids in some depositionally relevant modern marine environments (i.e., coastal water and estuaries) most likely derive from alphaproteobacteria.

To assess the capacity for 2-methylhopanoid production in other marine habitats that are preserved in the rock record, we analyzed *hpnP* clone libraries from Guerrero Negro hypersaline mats and Highborne Cay stromatolites (Hoehler et al., 2001; Dupraz & Visscher, 2005). Among Guerrero Negro samples that contained *hpnP*, 3 out of 4 had cyanobacterial *hpnP* genes while an equal number also contained alphaproteobacterial or unknown *hpnP* copies. Similarly in Highborne Cay, within the two samples found to have *hpnP*, both cyanobacterial and alphaproteobacterial copies of the gene were present (Figure 2.5). Although it is interesting that these habitats are the only marine environments where we find cyanobacterial *hpnP* genes, 2-methylhopanoid production cannot definitively be attributed to cyanobacteria in these environments. This result contrasts with a study of *hpnP* diversity in Hamelin Pool Shark Bay, Australia, another locality for stromatolite deposition, where mainly cyanobacterial *hpnP* was recovered (Garby et al., 2012). It remains to be determined if this inconsistency is due to an inherent difference between these locales or due to differences in sampling and methodology; for example, the *hpnP* PCR primers used in this study were on average more degenerate than those used in Garby et al. (2012).

In microbial ecology, it is well known that microheterogeneity can exist over small spatial scales (Hunt et al., 2008). In assessing our data, we observed spatial and temporal differences in *hpnP* diversity (Figure 2.5). These differences were evident even among samples with similar geochemistry, but the causes of variation are unknown. Related to this but at the level of lipid production, the presence of 2-methylhopanoids did not always correlate with the presence of *hpnP* (Figure 2.5). While the lack of 2-methylhopanoids may have been due to the detection limit of our analysis, two additional factors may explain this observation: first, 2-methylhopanoid production can depend on the growth condition (Rashby et al., 2007); second, the presence of 2-methylhopanoids, but not *hpnP*, may have resulted from bacteria no longer present assuming the degradation rate of hopanoids is considerably slower than the disappearance of DNA. Therefore, the presence of *hpnP* does not require that a community is making 2-methylhopanoids but indicates that it has the capacity to do so when the environment calls for it.

hpnP is correlated with plants

Using two independent methods we have shown that diverse microbes have the genetic potential to produce 2-methylhopanoids in a multitude of modern environments tested, making it difficult to use 2-methylhopanoids as unambiguous biomarkers for any particular taxonomic group. This leaves open the question of whether 2-methylhopanoids instead reflect a deeper underlying physiological function for the organisms that produce them. Notably, 43% of organisms with *hpnP* and 63% of *hpnP* sequences from metagenomes are plant-associated, defined as bacteria that form commensal or mutualistic symbiotic interactions with land plants. Notable examples include *Bradyrhizobium* spp., *Methylobacterium* spp., and *Nostoc* spp. (Figure 2.4) (Bravo et al., 2001; Knani et al., 1994; Meeks, 2009). To assess if this observation was non-random, we used the hypergeometric test calculated with the number of plant-associated bacteria estimated at the species level to alleviate bias in the dataset. We also performed this test with *shc*-containing bacteria and metagenomes. In all cases, we found the enrichment between (meta)genomes containing *hpnP* or *shc* and those that are plant-associated to be significant (p values < 10^{-6}): 46% and 51% of *hpnP*-containing bacteria and metagenomes are plant-associated, as well as 24% and 30% of those containing *shc*, in contrast to only 9% of bacteria (Table 2.2). There is thus a preferential association of 2-methylhopanoid production, and to a lesser extent hopanoid production, with plant symbiosis.

Table 2.2 | Hypergeometric probability of *hpnP* or *shc* correlating with plant-associated organisms or environments

	hpnP		*shc*	
	genomes	metagenomes	genomes	metagenomes
# genomes or metagenomes	1200	474	1200	474
# plant-associated	107	93	107	93
# contain *hpnP* or *shc*	26	59	183	221
# overlap	12	30	44	66
p value	$4.6\text{x}10^{-7*}$	$5.7\text{x}10^{-9*}$	$6.6\text{x}10^{-12*}$	$1.2\text{x}10^{-7*}$

*p values < 0.001 are significant.

To expand on this, we investigated the possibility that metabolisms used to establish plant-microbe interaction might be correlated to *hpnP*. It is common for plant symbionts to provide fixed nitrogen to plants or utilize methanol, a byproduct of plant metabolism, as a carbon source (Bravo et al., 2001; Gourion et al., 2006; Meeks, 2009). We found that 74% of *hpnP*-containing bacteria had either *nifD* and *moxF*, which encode proteins necessary for nitrogen fixation and methanol utilization, respectively (Figure 2.4 middle ring). Using the hypergeometric test, we found *hpnP* to be significantly overrepresented among bacteria containing *nifD* or *moxF* (p values < 10^{-6}; Table

2.3). Since the presence of *nifD* or *moxF* is not exclusive to plant symbionts, these numbers are likely overestimates, whereas the percentage of *hpnP* containing bacteria that are plant-associated is an underestimate of the number of *hpnP*-containing bacteria that are plant associated because not all organisms have been tested for the ability to form symbioses.

Consistent with this finding, the *hpnP* gene in *R. palustris* TIE-1 is regulated by an extracytoplasmic function (ECF) transcription factor conserved in alphaproteobacteria that plays a role in establishing plant symbioses in *Bradyrhizobium japonicum* and *Methylobacterium extorquens* (Gourion et al., 2009, 2006). This factor induces *hpnP* expression in response to osmotic stress in *R. palustris* (Kulkarni et al., 2013); osmolyte production by plants in the rhizosphere is well documented, and in some cases has been shown to promote plant-microbe symbioses (Miller & Wood, 1996; Khamar et al., 2010).

Table 2.3 | Hypergeometric probability of *hpnP* correlating with *nifD* or *moxF*

	nifD	*moxF*
# genomes	1200	1200
# with *nifD* or *moxF*	199	124
# with *hpnP*	26	26
# overlap	16	13
p value	$2.1\text{x}10^{-7*}$	$2.7\text{x}10^{-7*}$

*p values < 0.001 are significant.

Geobiological implications

While the presence of *hpnP* significantly correlates with plant-associated bacteria in modern environments, the 2-methylhopane record predate the rise of land plants (Clarke et al., 2011). Thus ancient symbioses clearly cannot explain the presence of 2-methylhopanes in the remote past. A possible explanation for why today we find the capacity for 2-methylhopanoid production enriched in habitats containing microbe-plant associations is that the capacity to make 2-methylhopanoids is selected by particular environmental conditions present in these habitats that are similar to those in the ancient depositional context. Specifically, we note that many modern environments containing 2-methylhopanoid producers comprise sessile microbial communities that have suboxia or anoxia, high osmolarity, and limited fixed nitrogen; these same parameters have also been used to describe the depositional context of 2-methylhopanes. For example, increased 2-methylhopane indices have been measured in ancient sedimentary rocks recording ocean anoxic events, which may have favored nitrogen fixing organisms (Knoll et al., 2007) and in sessile microbial mat and stromatolites, which contain high osmolytes in the form of extracellular polysaccharides and excreted small molecules (Summons et al., 1999). While none of the described niche parameters are solely responsible for the presence of *hpnP* or 2-methylhopanoids, and 2-methylhopanoids are not required for occupancy of this niche, their combination appears to be correlated with the capacity for 2-methylhopanoid production. Determining the underlying cellular role for 2-methylhopanoids given the described niche is necessary to provide a more definitive interpretation for ancient 2-methylhopanes. In conclusion, our ecological data demonstrate that 2-methylhopanoids cannot be used as taxonomic biomarkers for any particular group but suggest 2-methylhopanoids may be diagnostic for the confluence of particular environmental parameters.

Acknowledgements

We acknowledge members of the Newman lab for constructive comments on the manuscript. This work was supported by grants from the Howard Hughes Medical Institute to DKN and a NASA award (NNX12AD93G) to DKN, ALS, and RES. Research access to Yellowstone hot springs was granted to JRS from the Yellowstone Center for Resources. We thank V Orphan, D Des Marais, the NASA Ames Research Center, and ESSA Exportadora del Sal, SA, de CV, Guerrero Negro, Baja, California Sur, Mexico for samples of Guerrero Negro microbial mats. We are grateful to E Allen, J Valliere, D Caron, and A Lie for help with sample collection. JNR was supported by an NSF graduate fellowship, MLC by an Agouron Institute postdoctoral fellowship, and PVW by a NASA Astrobiology Institute postdoctoral fellowship. DKN is an HHMI Investigator.

References

Biers EJ, Sun S, Howard EC (2009) Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Applied Environmental Microbiology*, **75**, 2221–2229.

Brasier M, Green O, Lindsay J, Steele A (2004) Earth's oldest (approximately 3.5 Ga) fossils and the "Early Eden hypothesis": questioning the evidence. *Origins of Life and Evolution of Biospheres*, **34**:257–269.

Brasier M, Green OR, Jephcoat AP, Kleppe AK, Van Kranendonk MJ, Lindsay JF, Steele A, Grassineau NV (2002) Questioning the evidence for Earth's oldest fossils. *Nature*, **416**, 76–81.

Bravo J, Perzl M, Härtner T, Kannenberg EL, Rohmer M (2001) Novel methylated triterpenoids of the gammacerane series from the nitrogen-fixing bacterium *Bradyrhizobium japonicum* USDA 110. *European Journal of Biochemistry*, **268**, 1323-1331.

Brocks JJ, Logan GA, Buick R, Summons RE (1999) Archean molecular fossils and the early rise of eukaryotes. *Science*, **285**, 1033–1036.

Brocks JJ, Pearson A (2005) Building the Biomarker Tree of Life. *Review in Mineralogy and Geochemistry*, **59**, 233–258.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335–336.

Clarke JT, Warnock RCM, Donoghue PCJ (2011) Establishing a time-scale for plant evolution. *New Phytologist*, **192**, 266–301.

Cooke MP, Talbot HM, Farrimond P (2008) Bacterial populations recorded in bacteriohopanepolyol distributions in soils from Northern England. *Organic Geochemistry*, **39**, 1347-1358.

Dupraz C, Visscher PT (2005) Microbial lithification in marine stromatolites and hypersaline mats. *Trends in Microbiology*, **13**, 429–438.

Etherington GJ, Dicks J, Roberts IN (2005) Recombination Analysis Tool (RAT): a program for the high-throughput detection of recombination. *Bioinformatics*, **21**, 278–281.

Fischer WW, Summons RE, Pearson A (2005) Targeted genomic detection of biosynthetic pathways: anaerobic production of hopanoid biomarkers by a common sedimentary microbe. *Geobiology*, **3**, 33–40.

Garby TJ, Walter MR, Larkum AW, Neilan BA (2012) Diversity of cyanobacterial biomarker genes from the stromatolites of Shark Bay, Western Australia. *Environmental Microbiology*, **15**, 1464–1475.

Gourion B, Rossignol M, Vorholt JA (2006) A proteomic study of *Methylobacterium extorquens* reveals a response regulator essential for epiphytic growth. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 13186–13191.

Gourion B, Sulser S, Frunzke J, Francez-Charlot A, Stiefel P, Pessi G, Vorholt JA, Fischer H-M (2009) The PhyR-sigma(EcfG) signalling cascade is involved in stress response and symbiotic efficiency in *Bradyrhizobium japonicum*. *Molecular Microbiology*, **73**, 291–305.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, **59**, 307–321.

Hoehler TM, Bebout BM, Des Marais DJ (2001) The role of microbial mats in the production of reduced gases on the early Earth. *Nature*, **412**, 324–327.

Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science*, **320**, 1081–1085.

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, **9**, 286–298.

Khamar HJ, Breathwaite EK, Prasse CE, Fraley ER, Secor CR, Chibane FL, Elhai J, Chui W-L (2010) Multiple roles of soluble sugars in the establishment of *Gunnera-Nostoc* endosymbiosis. *Plant Physiology*, **154**, 1381–1389.

Knani M, Corpe WA, Rohmer M (1994) Bacterial hopanoids from pink-pigmented facultative methylotrophs (PPFMs) and from green plant surfaces. *Microbiology*, **140**, 2755–2759.

Knoll AH, Summons RE, Waldbauer JR, Zumberge JE (2007) The Geological Succession of Primary Producers in the Oceans. In: T*he Evolution of Primary Producers in the Sea*, (eds Falkowski P, Knoll AH), pp. 133–164. Academic Press, Boston.

Kulkarni G, Wu C-H, Newman DK (2013) The general stress response factor EcfG regulates expression of the C-2 hopanoid methylase HpnP in *Rhodopseudomonas palustris* TIE-1. *Journal of Bacteriology*, **195**, 2490–2498.

Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127–128.

Markowitz VM, Chen I-MA, Chu K, Szeto E, Palaniappan K, Grechkin Y, Ratner A, Jacob B, Pati A, Huntemann M, Liolios K, Pagani I, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC (2011) IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Research*, **40**, D123–D129.

Meeks JC (2009) Physiological Adaptations in Nitrogen-fixing *Nostoc*–Plant Symbiotic Associations. In: *Microbiology Monographs: Prokaryotic Symbionts in Plants* (ed Pawlowski K), pp. 181–205. Springer-Verlag, Munster Germany.

Miller KJ, Wood JM (1996) Osmoadaptation by rhizosphere bacteria. *Annual Reviews of Microbiology*, **50**, 101–136.

Mulkidjanian AY, Koonin EV, Makarova KS, Mekhedov SL, Sorokin A, Wolf YI, Dufresne A, Partensky F, Burd H, Kaznadzey D, Haselkorn R, Galperin MY (2006) The cyanobacterial genome core and the origin of photosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 13126–13131.

Osburn MR, Sessions AL, Pepe-Ranney C, Spear JR (2011) Hydrogen-isotopic variability in fatty acids from Yellowstone National Park hot spring microbial communities. *Geochimica et Cosmochimica Acta*, **75**, 4830–4845.

Ourisson G, Rohmer M, Poralla K (1987) Prokaryotic hopanoids and other polyterpenoid sterol surrogates. *Annual Reviews of Microbiology*, **41**, 301–333.

Pearson A, Flood Page SR, Jorgenson TL, Fischer WW, Higgins MB (2007) Novel hopanoid cyclases from the environment. *Environmental Microbiology*, **9**, 2175–2188.

Pearson A, Leavitt WD, Sáenz JP, Summons RE, Tam MC-M, Close HG (2009) Diversity of hopanoids and squalene-hopene cyclases across a tropical land-sea gradient. *Environmental Microbiology*, **11**, 1208–1223.

Pearson A, Rusch DB (2009) Distribution of microbial terpenoid lipid cyclases in the global ocean metagenome. *The ISME Journal*, **3**, 352–363.

Rashby SE, Sessions AL, Summons RE, Newman DK (2007) Biosynthesis of 2-methylbacteriohopanepolyols by an anoxygenic phototroph. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 15099–15104.

Rasmussen B, Fletcher IR, Brocks JJ, Kilburn MR (2008) Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature*, **455**, 1101–1104.

Rohmer M (2010) Handbook of Hydrocarbon and Lipid Microbiology. In: *Chemistry and Physics of Lipids* (ed Timmis KN), Springer, Berlin Heidelberg.

Rohmer M, Bouvier-Nave P, Ourisson G (1984) Distribution of Hopanoid Triterpenes in Prokaryotes. *Journal of General Microbiology*, **130**, 1137–1150.

Sáenz JP, Eglinton TI, Summons RE (2011) Abundance and structural diversity of bacteriohopanepolyols in suspended particulate matter along a river to ocean transect. *Organic Geochemistry*, **42**, 774–780.

Schopf JW, Packer B (1987) Early Archean (3.3-billion to 3.5-billion-year-old) microfossils from Warrawoona Group, Australia. *Science*, **237**, 70–73.

Sessions AL, Zhang L, Welander PV, Doughty D, Summons RE, Newman DK (2013) Identification and quantification of polyfunctionalized hopanoids by high temperature gas chromatography–mass spectrometry. *Organic Geochemistry*, **56**, 120–130.

Summons RE, Jahnke LL, Hope JM, Logan GA (1999) 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature*, **400**, 554–557.

Sun S, Chen J, Li W, Altintas I, Lin A, Peltier S, Stocks K, Allen EE, Ellisman M, Grethe J, Wooley J (2011) Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Research*, **39**, D546–D551.

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, **56**, 564–577.

Talbot HM, Farrimond P (2007) Bacterial populations recorded in diverse sedimentary biohopanoid distributions. *Organic Geochemistry*, **38**, 1212-1225.

Talbot HM, Rohmer M, Farrimond P (2007a) Rapid structural elucidation of composite bacterial hopanoids by atmospheric pressure chemical ionization liquid chromatography/ion trap mass spectrometry. *Rapid Communications in Mass Spectrometry*, **21**, 880–892.

Talbot HM, Rohmer M, Farrimond P (2007b) Structural characterization of unsaturated bacterial hopanoids by atmospheric pressure chemical ionization liquid chromatography/ion trap mass spectrometry. *Rapid Communications in Mass Spectrometry*, **21**, 1613–1622.

Talbot HM, Squier AH, Keely BJ, Farrimond P (2003a) Atmospheric pressure chemical ionization reversed-phase liquid chromatography/ion trap mass spectrometry of intact bacteriohopanepolyols. *Rapid Communications in Mass Spectrometry*, **17**, 728–737.

Talbot HM, Summons RE, Jahnke L, Cockell CS, Rohmer M, Farrimond P (2008) Cyanobacterial bacteriohopanepolyol signatures from cultures and natural environmental settings. *Organic Geochemistry*, **39**, 232-263.

Talbot HM, Summons RE, Jahnke L, Farrimond P (2003b) Characteristic fragmentation of bacteriohopanepolyols during atmospheric pressure chemical ionization liquid chromatography/ion trap mass spectrometry. *Rapid Communications in Mass Spectrometry,* **17**, 2788–2796.

Talbot HM, Watson DF, Murrell JC, Carter JF, Farrimond P (2001) Analysis of intact bacteriohopanepolyols from methanotrophic bacteria by reversed-phase high-performance liquid chromatography-atmospheric pressure chemical ionization mass spectrometry. *Journal of Chromatography A*, **921**, 175–185.

Walter MR, Grotzinger JP, Schopf JW (1992) Proterozoic stromatolites. In: *The Proterozoic Biosphere: A Multidisciplinary Study* (eds Schopf JW, Klien C), pp. 253–260. Cambridge University Press, Cambridge.

Welander PV, Coleman ML, Sessions AL, Summons RE, Newman DK (2010) Identification of a methylase required for 2-methylhopanoid production and implications for the interpretation of sedimentary hopanes. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 8537–8542.

Welander PV, Doughty DM, Wu C-H, Mehay S, Summons RE, Newman DK (2012) Identification and characterization of *Rhodopseudomonas palustris* TIE-1 hopanoid biosynthesis mutants. *Geobiology*, **10**, 163–177.

Xu Y, Cooke MP, Talbot HM, Simpson MJ (2009) Bacteriohopanepolyol signatures of bacterial populations in Western Canadian soils. *Organic Geochemistry*, **40**, 79-86.

Chapter 3

PHYLOGENETIC ANALYSIS OF HPNP REVEALS THE ORIGIN OF 2-METHYLHOPANOID PRODUCTION IN ALPHAPROTEOBACTERIA

This chapter was first published as, Ricci JN, Michel AJ, Newman DK (2015) Phylogenetic analysis of HpnP reveals the origin of 2-methylhopanoid production in Alphaproteobacteria. *Geobiology*, **13**, 267-277.

Abstract

Hopanoids are bacterial steroid-like lipids that can be preserved in the rock record on billion-year timescales. 2-Methylhopanoids are of particular interest to geobiologists because methylation is one of the few chemical modifications that remain after diagenesis and catagenesis. 2-methylhopanes, the molecular fossils of 2-methylhopanoids, are episodically enriched in the rock record, but we do not have a robust interpretation for their abundance patterns. Here, we exploit the evolutionary record found in molecular sequences from extant organisms to reconstruct the biosynthetic history of 2-methylhopanoids using the C-2 hopanoid methylase, HpnP. Based on HpnP phylogenetic analysis, we find that 2-methylhopanoids originated in a subset of the Alphaproteobacteria. This conclusion is statistically robust and reproducible in multiple trials varying the outgroup, trimming stringency, and ingroup dataset used to infer the evolution of this protein family. The capacity for 2-methylhopanoid production was likely horizontally transferred from the Alphaproteobacteria into the Cyanobacteria after the Cyanobacteria's major divergences. Together, these results suggest that the ancestral function of 2-methylhopanoids was not related to oxygenic photosynthesis but instead to a trait already present in the Alphaproteobacteria. Moreover, given that early 2-methylhopane deposits could have been made solely by alphaproteobacteria before the acquisition of *hpnP* by cyanobacteria, and that the Alphaproteobacteria are thought to be ancestrally aerobic, we infer that

2-methylhopanoids likely arose after the oxygenation of the atmosphere. This finding is consistent with the geologic record; the oldest known syngenetic 2-methylhopanes occur after the rise of oxygen, in middle Proterozoic strata of the Barney Creek Formation.

Introduction

Hopanoids are a diverse class of pentacyclic triterpenoid lipids produced by some bacteria. Due to their structural similarity to steroids, they have been proposed to play a role in membrane integrity and resistance to various stresses, such as high temperature and extreme pH (Poralla et al., 1984; Kannenberg & Poralla, 1999; Welander et al., 2009, 2012; Doughty et al., 2011). The hydrocarbon backbones of hopanoids can be preserved in the rock record as biomarkers or molecular fossils. During the transformation from sediment to rock, hopanoids lose many functional groups that differentiate them in cells today. One of the few chemical modifications maintained throughout these fossilization processes is methylation at the C-2 position. 2-methylhopanes, the molecular fossil derivatives of 2-methylhopanoids, date to 1.64 billion years ago and exhibit an increase in abundance during ocean anoxic events, making them important biomarkers with episodic enrichment in the fossil record (Rasmussen et al., 2008; Knoll et al., 2007).

2-methylhopanoids were once considered biomarkers of Cyanobacteria (Summons et al., 1999), but multiple lines of evidence suggest this interpretation is no longer valid (Rashby et al., 2007; Welander et al., 2010; Ricci et al., 2014). In an attempt to understand what their molecular fossils may be telling us about ancient environments and biological communities, a variety of approaches have been used to gain insight into the taxonomic sources and cellular functions of 2-methylhopanoids. Upon the identification of the enzyme responsible for C-2 hopanoid methylation, HpnP, in *Rhodopseudomonas palustris* TIE-1, HpnP homologs were found in modern cyanobacteria, alphaproteobacteria, and an acidobacterium (Welander et al., 2010). Environmental and metagenomic surveys further established that diverse HpnP sequences are present in many locales, but are disproportionately associated with plants and terrestrial environments (Ricci et al., 2014). Consistent with this finding, bacteriohopanepolyols, including 2-methylhopanoids, isolated from particulate matter along a river to ocean transect were shown to derive from terrigenous sources (Sáenz et al., 2011). Based on studies of *R. palustris* TIE-1, *hpnP* appears to be regulated by a variety of stressors and 2-methylhopanoids are enriched in the outer membrane compared to the inner membrane (Doughty et al., 2011; Kulkarni et al., 2013). Similarly, 2-methylhopanoids are concentrated in the outer membrane of akinetes—survival structures—of the cyanobacterium *Nostoc punctiforme* (Doughty et al., 2009, 2014). These results suggest that 2-methylhopanoids may play a specific role in conferring stress resistance, consistent with recent in vitro experiments showing that 2-methylhopanoids rigidify membranes more than their desmethyl equivalents (Wu et al., 2015).

Despite the emerging picture of the modern biological sources and function of 2-methylhopanoids, better understanding of the evolution of 2-methylhopane synthesis is needed to interpret the significance of episodic 2-methylhopane occurrences in the rock record. While 2-methylhopanoids originate from certain bacterial sources on Earth today, this does not necessarily mean related organisms produced them in the past. One way to constrain this uncertainty is to determine the likely ancestral source of 2-methylhopanoids by decoding the evolutionary history of the HpnP methylase. Phylogenetic analysis may reveal the history of a gene or protein, the type of organism(s) in which it evolved, the order of inheritance between organisms, and the mode of inheritance between different species (Omland et al., 2008; Yang & Rannala, 2012; Vogl & Bryant, 2012). Whereas the presence of biomarkers and other fossils in the geologic record can be used to link these signals to discrete intervals in time, comparative DNA, RNA, and protein sequence analysis from living organisms can reveal their relative temporal histories. Molecular phylogenies can be also be calibrated against the geologic record to constrain evolutionary processes in absolute time (Shih & Matzke, 2013). Molecular analyses are often accompanied by statistical probabilities, lending rigor to inferences derived from these techniques.

Previously, we attempted to reconstruct the phylogeny of HpnP to gain insight into which organisms first produced 2-methylhopanoids (Welander et al., 2010). This prior attempt to resolve the HpnP phylogeny was inconclusive due to the paucity of HpnP sequences available at the time. Since then, we made a targeted effort to clone more diverse environmental *hpnP* sequences from various habitats (Ricci et al., 2014), as diverse sequences are needed to root the HpnP family. In addition, many more *hpnP*-containing genomes have been deposited in publically-available databases in the past few years. This increase in HpnP sequence diversity prompted us to revisit the challenge of inferring the evolutionary history of HpnP. Here we show that it is now possible to derive a conserved and robust HpnP evolutionary history. We report the phylum of HpnP origin, the mode of *hpnP* inheritance between and within phyla, and the last *hpnP*-containing ancestor of each phylum. These data provide a picture of how HpnP and therefore 2-methylhopanoid production has evolved, which has important implications for the relative timing of its origin with respect to the evolution of oxygenic photosynthesis.

Methods

Environmental data

We used a combination of genomic and environmental HpnP sequences to increase the diversity in our phylogeny. We recovered full-length environmental *hpnP* sequences from Imperial Geyser in Yellowstone National Park. The target *hpnP* was identified based on its unique phylogenetic location in a previously published study of environmental *hpnP* diversity (Ricci et al., 2014). We initially employed nested inverse PCR to retrieve the flanking sequences of the gene with primers designed within an approximately 550 bp region identified in previous work (Table 3.1) (Uchiyama & Watanabe, 2006; Ricci et al., 2014). This approach retrieved 83% of the gene sequence leaving 273 bp at its 5' region unknown. We then used the DNA Walking SpeedUp^{TM} kit (Seegene, Seoul, South Korea), a commercially available arbitrary PCR kit, to acquire the remainder of the gene. Two variants of the target *hpnP* were sequenced, due to the presence of a single amino acid discrepancy. We included both alternatives in our phylogeny as both are likely found in the environment. The environmental *hpnP* genes were sequenced at Retrogen (San Diego, CA, USA) and were assembled in Geneious 6.0.5 created by Biomatters. Available at http://www.geneious.com/. Sequences are available under Genbank accession numbers KP204881 and KP204882.

Table 3.1 | Environmental genome walking primers

Primer Name	Target	Method	Primer Sequence
IGU1iPCR1F	U1 IG1	Inverse PCR	AATGATCTGCTGTGGGGGTC
IGU1iPCR1R	U1 IG1	Inverse PCR	GGCAGAAAGAGCGGGGTTAT
IGU1iPCR2F	U1 IG1	Inverse PCR	CGTTCCGTGTGGCTATCCAA
IGU1iPCR2R	U1 IG1	Inverse PCR	CGGTGCAATTTGCCTGTGAA
IGU1SP1F	U1 IG1	SpeedUp Forward	AACAAAGTAGATGGCTCCCG
IGU1SP2F	U1 IG1	SpeedUp Forward	AATGATCTGCTGTGGGGGTC
IGU1SP3F	U1 IG1	SpeedUp Forward	CGTTCCGTGTGGCTATCCAA

HpnP phylogenetic diversity comparison

To quantify the increase in HpnP sequence diversity, we calculated the phylogenetic diversity metric (total branch length) and Colless's Imbalance using Mesquite 2.75 (Table 3.2) (Maddison & Maddison, 2011; Maddison et al., 2011). The percentile of imbalance is based on an equiprobable distribution of 1000 random trees. Ingroup only phylogenies were generated by aligning sequence in MAFFT 7.158 using L-INS-i (Katoh & Toh, 2008). The alignments were trimmed in Gblocks 0.91b with relaxed parameters; see below (Talavera & Castresana, 2007). Trees were constructed in

PhyML 3.1 with the same settings used for the ingroup plus outgroup trials, see below (Guindon et al., 2010).

Table 3.2 | Improved phylogenetic diversity

Phylogeny	Number of unique sequences	PD score*	Colless's Imbalance†
Welander et al., 2010	28	8.95	0.387(48)
This study, genomic sequences only	115	14.02	0.111(2)
This study, genomic and environmental sequences	117	14.07	0.132(4)

*Phylogenetic diversity (PD) score is the sum of all branch lengths.
†Colless's Imbalance ranges from 0(even) to 1(lateralized). The percentile of the target tree's imbalance is found in parentheses and is based on an equiprobable distribution of random trees (n=1000).

HpnP phylogeny

The top 500 homologs of *R. palustris* TIE-1 HpnP (GI 192292635) from the NCBI non-redundant (nr) database (excluding uncultured and environmental sample sequences) were retrieved in July 2014. These sequences were aligned using the L-INS-i algorithm of MAFFT 7.158 (Katoh & Toh, 2008), and a tree was constructed using the default parameters in PhyML 3.1 (Guindon et al., 2010). From this tree of HpnP (greater than 55% identity) and its homologs (greater than 30% identity), outgroup sequences were selected from a closely related clade of class B Radical SAM enzymes of unknown function using iTOL (Figure 3.1A, Appendix B) (Letunic & Bork, 2007; Zhang et al., 2012). The regions of amino acid conservation between different HnpP sequences occurred throughout the length of the protein.

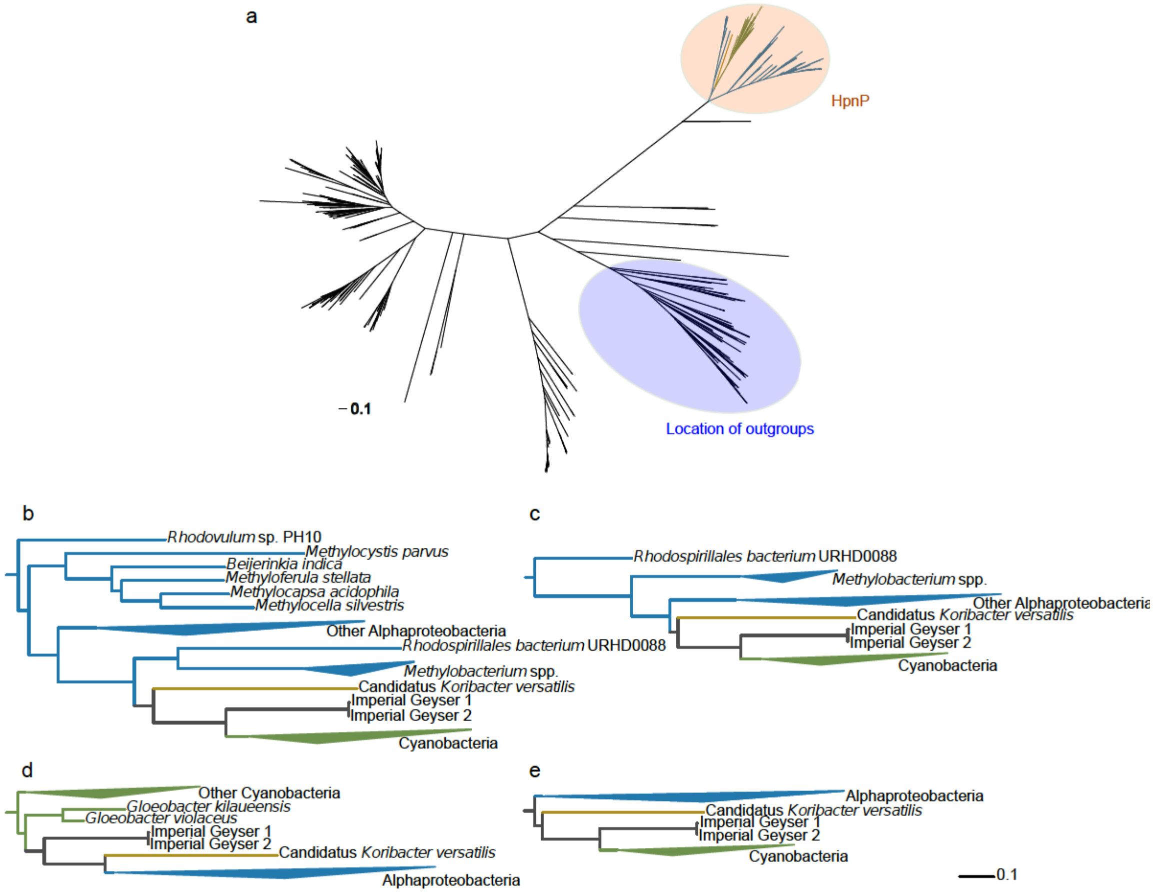


Figure 3.1 | The HpnP family and hypothesized rooted topologies. (A) Unrooted maximum likelihood phylogeny of HpnP and homologs. The location of the HpnP sequences and outgroup sequences used in this study are highlighted. (B, C) HpnP phylogenies rooted in the Alphaproteobacteria with *Rhodovulum* sp. PH10 and Rhodospirillales bacterium URHD0088 as the most basal branches, respectively. (D) HpnP phylogeny rooted in the Cyanobacteria. (E) HpnP phylogeny with monophyletic alphaproteobacterial and cyanobacterial clades. Branch colors: Alphaproteobacteria – blue, Cyanobacteria – green, Acidobacteria – yellow, and unknown – gray. The scale bars represents 0.1 substitutions per site.

Ingroup and outgroup sequences were aligned with and without environmental HpnP sequences using the L-INS-i algorithm of MAFFT 7.158 (Katoh & Toh, 2008). The alignments were trimmed with three different thresholds to tune the signal-to-noise ratio of homologous and non-homologous sites using Gblocks 0.91b: none, relaxed, and stringent (default) (Talavera & Castresana, 2007). For relaxed trimming the number of sequences for a conserved and flanking position were set to the minimum, the number of contiguous non-conserved positions was set to 20, the block length was 5, and gaps at all positions were allowed. Phylogenies were created from the resulting alignments in

PhyML 3.1 using the LG model, selected by ProtTest 3.2 using Akaike Information Criterion, 5 random starting trees, SPR and NNI branch swapping, and gamma rate categories and substitution parameters estimated from the data (Guindon et al., 2010; Darriba et al., 2011). Best trees from each trial, varying outgroups, datasets, and trimming, are reported in Table 3.3.

Table 3.3 | Comparison of phylogeny trials and alternate topologies

						log-likelihood of topology*				p-value of topology†			
Trial	Outgroup	Trimming	Dataset	No. sites	No. taxa	A	B	C	D	A	B	C	D
1	1	none	genomic	631	120	**-34,084**	-34,092	-34,102	-34,092	**0.663**	**0.405**	0.042	**0.397**
2	1	relaxed	genomic	553	120	**-31,319**	-31,321	-31,331	-31,322	**0.629**	**0.456**	0.043	**0.381**
3	1	stringent	genomic	416	120	**-22,762**	-22,788	-22,797	-22,788	**0.844**	**0.198**	0.014	**0.226**
4	1	none	environmental	631	122	**-34,634**	-34,637	-34,651	-34,641	**0.532**	**0.617**	0.034	**0.256**
5	1	relaxed	environmental	553	122	**-31,830**	-31,827	-31,842	-31,832	**0.434**	**0.686**	0.020	**0.205**
6	1	stringent	environmental	414	122	**-22,942**	-22,954	-22,967	-22,958	**0.688**	**0.426**	0.018	**0.143**
7	2	none	genomic	619	125	**-37,236**	-37,248	-37,256	-37,247	**0.698**	**0.318**	**0.052**	**0.515**
8	2	relaxed	genomic	541	125	**-34,468**	-34,474	-34,482	-34,473	**0.737**	**0.326**	**0.051**	**0.399**
9	2	stringent	genomic	419	125	**-25,642**	-25,672	-25,676	-25,669	**0.809**	**0.160**	**0.064**	**0.310**
10	2	none	environmental	619	127	**-37,786**	-37,795	-37,805	-37,795	**0.635**	**0.178**	0.047	**0.470**
11	2	relaxed	environmental	541	127	**-34,975**	-34,980	-34,990	-34,980	**0.662**	**0.405**	0.044	**0.423**
12	2	stringent	environmental	419	127	**-26,049**	-26,068	-26,074	-26,066	**0.737**	**0.245**	**0.065**	**0.402**

Topologies: A, root in the Alphaproteobacteria on branch to *Rhodovulum* sp. PH10; B, root in the Alphaproteobacteria on branch to *Rhodospirillales bacterium* URHD0088; C, root in the Cyanobacteria; D root unknown with monophyletic Alphaproteobacteria and Cyanobacteria

* log likelihood of best unconstrained trees are bold.

†Possible trees as determined by the approximately unbiased test are bold.

For each trial, alternate tree topologies were generated (Figure 3.1B-E). The log-likelihoods of these alternative trees were calculated in PhyML 3.1 by estimating substitution parameters and determining branch lengths without optimizing the topology. The approximately unbiased (AU) test found in CONSEL 0.20 was used to compare the site likelihood between the best tree and the alternate trees (Table 3.3) (Shimodaira & Hasegawa, 2001; Shimodaira, 2002). Trees with a p-value < 0.05 can be eliminated as possible trees, while any tree with a p-value ≥ 0.05 could be the true tree. The phylogeny generated from trial 11 was selected for the remaining analyses because it has a larger and presumably more balanced and resolved outgroup, includes environmental HpnP sequences, and employs a median trimming setting.

Ancestral state reconstruction

The reconstruction of ancestral phyla containing *hpnP* was calculated in Mesquite 2.75 using the phylogeny from trial 11 (Table 3.3). The parsimony reconstruction assumed unordered states. The likelihood reconstruction used the Mk1 model (Markov k-state 1 parameter model) (Maddison & Maddison, 2006, 2011).

Species and gene tree reconciliation

Reference species trees for Alphaproteobacteria and Cyanobacteria were constructed using mainly full-length 16S rDNA gene sequences from genomes with *hpnP* and from finished genomes without *hpnP* but from organisms in para- or polyphyletic relationships with organisms that contain *hpnP*. The phylogenies were generated as described above for HpnP. Reconciliations between undated species trees and their respective rooted HpnP tree lacking branch support were carried out using Jane version 4 and Ranger-DTL version 1.0 (Conow et al., 2010; Bansal et al., 2012). Event costs were assigned as 1 for loss, 2 for duplication, and 3 for transfer.

Our species tree and HpnP tree reconciliations were completed with 16S rDNA gene phylogenies, even though multigene concatenated phylogenies are thought to better represent the species tree (Williams et al., 2007). We used 16S rDNA gene sequences because many of the genomes containing *hpnP* are incomplete and missing genes that would be used to generate a multigene species tree. The 16S rDNA gene phylogenies for both the Alphaproteobacteria and Cyanobacteria have good congruence with multigene trees (Williams et al., 2007; Shih et al., 2013). While improved species trees may alter some conclusions about the transfer of *hpnP* genes within the major phyla, our conclusions about interphylum transfers should be insensitive to this methodological aspect.

Results

Improved HpnP diversity

To quantify how the known HpnP diversity has improved, we compared our current dataset with the one previously published (Table 3.2) (Welander et al., 2010). The number of full-length HpnP sequences increased approximately four-fold between the two datasets. This increase in the number of sequences is partially due to targeted genome sequencing efforts of particular genera (e.g., Bradyrhizobium spp.); because these sequences are very similar to ones that are already known, they do not significantly increase the phylogenetic breadth. To correct for this we used the phylogenetic diversity metric, which is the sum of the branch lengths of the tree. The phylogenetic diversity score is 1.6 times higher in the current phylogeny than in Welander et al. (2010). Additionally, we estimated the phylogenies' imbalance using Colless's Imbalance, which assesses how evenly distributed branches are on a tree (Colless DH, 1982). A more even tree implies that the tree is better sampled than a lateralized tree. The phylogenies reported in this study fall into a lower percentile of imbalance and have a Colless's Imbalance score 2.9–3.5 times lower than those in Welander et al. (2010), signifying that our current tree is more balanced. Taken together, these metrics indicate that the phylogeny reported here better samples HpnP diversity than our previous attempt.

HpnP phylogenies have a conserved topology

In our analysis we varied the outgroup, trimming stringency, and ingroup dataset with and without environmental sequences; each of these trials selected the same outgroup location (Figure 3.1B, Table 3.3 bolded log likelihood of topology values). The best tree comprised a monophyletic clade of the Cyanobacteria, Acidobacteria, and environmental HpnP sequences nested within the Alphaproteobacteria, as well as another monophyletic cyanobacterial only clade within this larger group (Figure 3.1B, 2). The most basal branch of the best topology was Rhodovulum sp. PH10. The ingroup topology of HpnP was generally well supported (Figure 3.2), but we observed small changes in the ingroup topology between trials. Generally these differences were among poorly supported nodes with low bootstrap and aLRT values in the Cyanobacteria, but they did not affect the overall topology between phyla or outgroup placement.

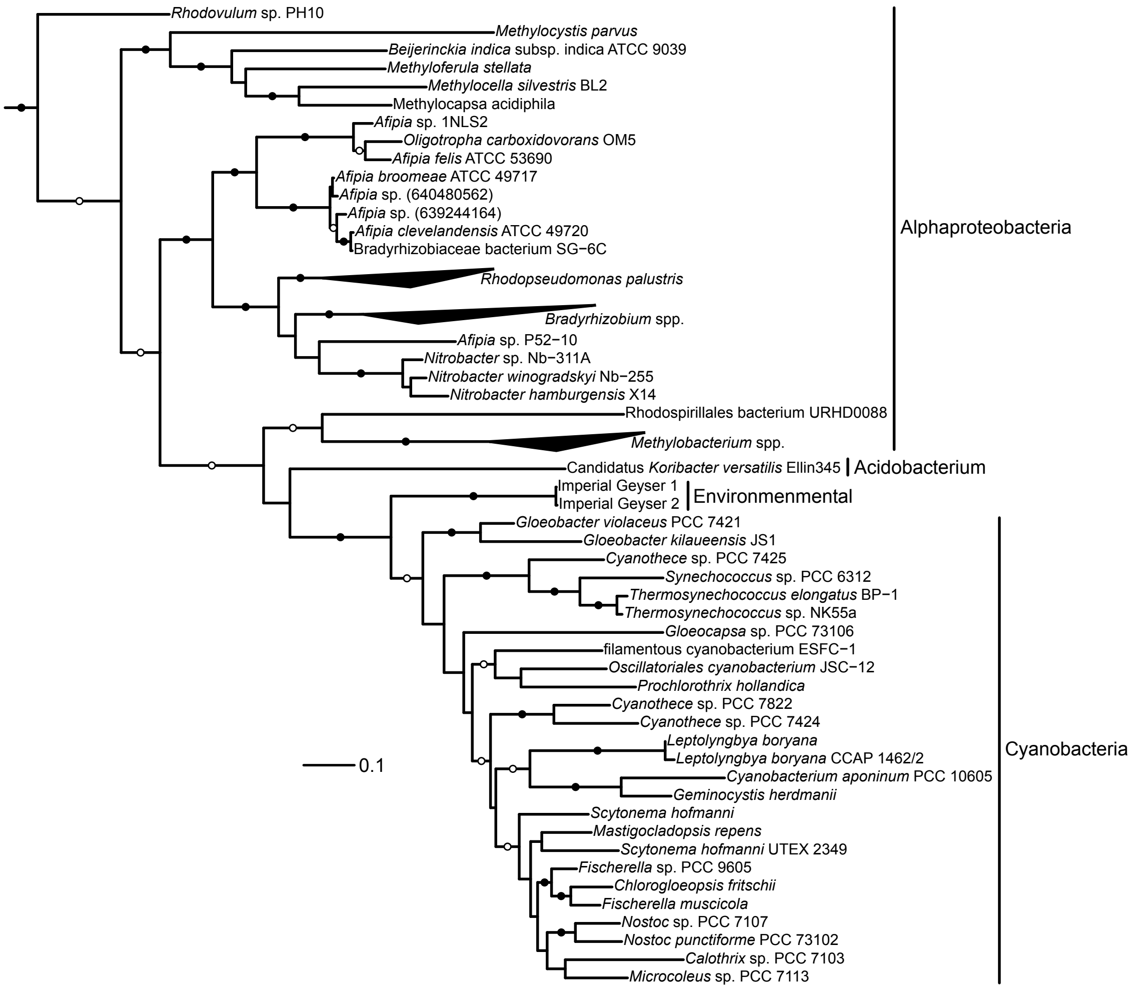

Rhodovulum sp. PH10
Methylocystis parvus
Beijerinckia indica subsp. indica ATCC 9039
Methyloferula stellata
Methylocella silvestris BL2
Methylocapsa acidiphila
Afipia sp. 1NLS2
Oligotropha carboxidovorans OM5
Afipia felis ATCC 53690
Afipia broomeae ATCC 49717
Afipia sp. (640480562)
Afipia sp. (639244164)
Afipia clevelandensis ATCC 49720
Bradyrhizobiaceae bacterium SG−6C
Rhodopseudomonas palustris
Bradyrhizobium spp.
Afipia sp. P52−10
Nitrobacter sp. Nb−311A
Nitrobacter winogradskyi Nb−255
Nitrobacter hamburgensis X14
Rhodospirillales bacterium URHD0088
Methylobacterium spp.
Alphaproteobacteria
Candidatus *Koribacter versatilis* Ellin345
Acidobacterium
Imperial Geyser 1
Imperial Geyser 2
Environmenmental
Gloeobacter violaceus PCC 7421
Gloeobacter kilaueensis JS1
Cyanothece sp. PCC 7425
Synechococcus sp. PCC 6312
Thermosynechococcus elongatus BP−1
Thermosynechococcus sp. NK55a
Gloeocapsa sp. PCC 73106
filamentous cyanobacterium ESFC−1
Oscillatoriales cyanobacterium JSC−12
Prochlorothrix hollandica
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Leptolyngbya boryana
Leptolyngbya boryana CCAP 1462/2
Cyanobacterium aponinum PCC 10605
Geminocystis herdmanii
Scytonema hofmanni
Mastigocladopsis repens
Scytonema hofmanni UTEX 2349
Fischerella sp. PCC 9605
Chlorogloeopsis fritschii
Fischerella muscicola
Nostoc sp. PCC 7107
Nostoc punctiforme PCC 73102
Calothrix sp. PCC 7103
Microcoleus sp. PCC 7113
Cyanobacteria
0.1

Figure 3.2 | Representative maximum likelihood phylogeny of HpnP. aLRT values and 1000 bootstrap replicates were calculated for branch support. Closed circles denote branches with both aLRT SH-like (cutoff ≥ 0.85) and bootstrap (cutoff ≥ 85%) values equal to or greater than the cutoffs, while open circles indicate branches with only aLRT values equal to or greater than the cutoff. The phylogeny shown is the best tree from trial 11, but the topology is representative of the best trees from all trials. The scale bar represents 0.1 substitutions per site.

We then sought to understand if our best topology was significantly better than competing topologies. We generated three alternate topologies similar to those seen in Welander et al. (2010) (Figure 3.1C-E) and compared the four phylogenies using the AU test (Table 3.3 p-value of topology). In eight of twelve trials the topology with the outgroup placed in the Cyanobacteria was eliminated from the pool of possible trees (Table 3.3, Figure 3.1D). The four trials (7-9 and 12) that consider the cyanobacterial root possible used outgroup 2, whereas none of the trials using outgroup 1 considered this root possible. Additionally, trials 7–9 did not include the environmental HpnP sequences; once these additional sequences were included, two of three trials (10 and 11 out of 10-12) eliminated the cyanobacterial root. The stringent trimming used in trial 12 may have removed sites needed to distinguish between the alternate topologies. These findings imply that the AU test is sensitive to the outgroup used but is able to eliminate the cyanobacterial root with added sequences and conservative trimming. Considering that the AU test eliminates the cyanobacterial root in the majority of trials and the trials that do not remove this topology improve with reasonable modifications, the cyanobacterial root of HpnP is improbable.

The AU test was unable to eliminate the alternate alphaproteobacterial and unknown root topologies in any trial (Figure 3.1C, E). While these topologies are considered possible HpnP phylogenies, they were never found as the most statistically well-supported tree in any trial where the topology was unconstrained. Accordingly, we infer the topology of Figure 3.1B (see also Figure 3.2) to be the most likely given our current data.

Ancestral phyla reconstruction

We used two distinct statistical methods of ancestral state reconstruction, parsimony and maximum likelihood, to infer the phyla of ancestral *hpnP*. It is most parsimonious for the last common ancestor of *hpnP* to reside in the Alphaproteobacteria (Figure 3.3). The maximum likelihood approach was congruent with the parsimony reconstruction, finding a 0.99 normalized likelihood for an ancestral alphaproteobacterial *hpnP*. The maximum likelihood reconstruction allowed further resolution of a key internal node, highlighted in Figure 3.3, over the parsimony approach. This internal node was likely an alphaproteobacterium or a cyanobacterium rather than an acidobacterium. Here, we assume the transitions between phyla were horizontal gene transfers because the alternative hypothesis would require multiple loss events, which is less parsimonious. Thus based on the current data, *hpnP* likely arose in the Alphaproteobacteria and was transferred from this phylum into the Cyanobacteria. Horizontal gene transfer events have been observed between these phyla (Beiko et al., 2005), and their 2-methylhopanoid producing members are known to co-occur in multiple habitats (Ricci et al., 2014). At this time, we cannot resolve which of these groups transferred *hpnP* into the Acidobacteria.

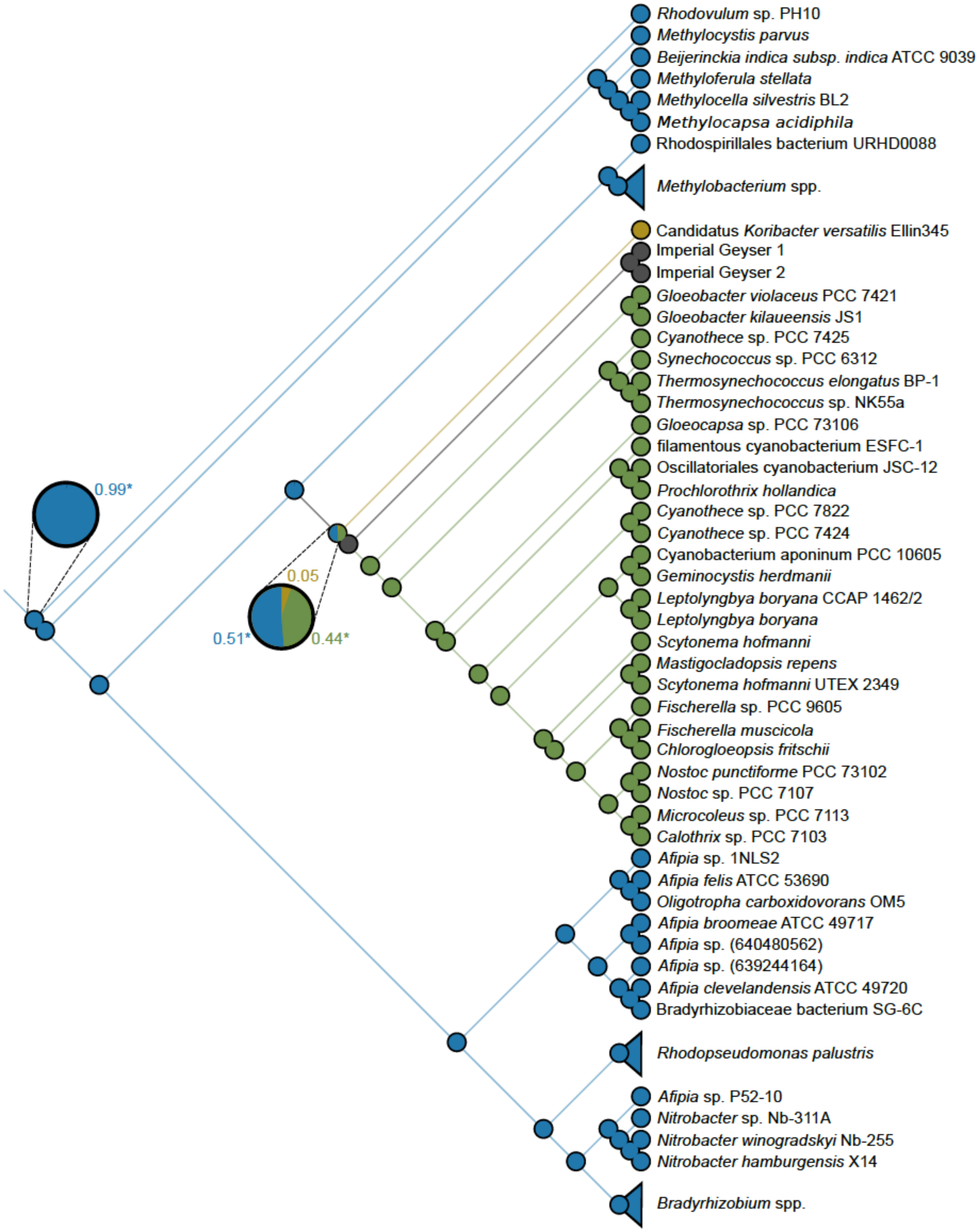

Rhodovulum sp. PH10
Methylocystis parvus
Beijerinckia indica subsp. indica ATCC 9039
Methyloferula stellata
Methylocella silvestris BL2
Methylocapsa acidiphila
Rhodospirillales bacterium URHD0088
Methylobacterium spp.
Candidatus Koribacter versatilis Ellin345
Imperial Geyser 1
Imperial Geyser 2
Gloeobacter violaceus PCC 7421
Gloeobacter kilaueensis JS1
Cyanothece sp. PCC 7425
Synechococcus sp. PCC 6312
Thermosynechococcus elongatus BP-1
Thermosynechococcus sp. NK55a
Gloeocapsa sp. PCC 73106
filamentous cyanobacterium ESFC-1
Oscillatoriales cyanobacterium JSC-12
Prochlorothrix hollandica
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Cyanobacterium aponinum PCC 10605
Geminocystis herdmanii
Leptolyngbya boryana CCAP 1462/2
Leptolyngbya boryana
Scytonema hofmanni
Mastigocladopsis repens
Scytonema hofmanni UTEX 2349
Fischerella sp. PCC 9605
Fischerella muscicola
Chlorogloeopsis fritschii
Nostoc punctiforme PCC 73102
Nostoc sp. PCC 7107
Microcoleus sp. PCC 7113
Calothrix sp. PCC 7103
Afipia sp. 1NLS2
Afipia felis ATCC 53690
Oligotropha carboxidovorans OM5
Afipia broomeae ATCC 49717
Afipia sp. (640480562)
Afipia sp. (639244164)
Afipia clevelandensis ATCC 49720
Bradyrhizobiaceae bacterium SG-6C
Rhodopseudomonas palustris
Afipia sp. P52-10
Nitrobacter sp. Nb-311A
Nitrobacter winogradskyi Nb-255
Nitrobacter hamburgensis X14
Bradyrhizobium spp.
0.99*
0.05
0.51*
0.44*

Figure 3.3 | Maximum likelihood and parsimony ancestral state reconstructions of HpnP phyla. Ancestral state reconstructions are based on the phylogeny in Figure 2. Branch colors specify the most parsimonious ancestral state. The pie charts at each internal node contain the normalized likelihoods of the ancestor being from a particular phylum, while the circles present on leafs indicate the phylum assigned to extant organisms. At all internal nodes, the dominant state is the only significant ancestral phylum except for one node where two significant values are denoted with an asterisk. Colors: Alphaproteobacteria – blue, Cyanobacteria – green, Acidobacteria – yellow, and unknown – gray (treated as missing data).

hpnP and species tree reconciliations within phyla

Because our earlier study had suggested that *hpnP* was subject to more gene transfer and gene loss within the Cyanobacteria than the Alphaproteobacteria (Welander et al., 2010), we sought to reassess this observation. Using gene and species reconciliation methods that consider speciation (vertical), duplication, transfer (horizontal), and loss events, we examined the transfer of *hpnP* within individual phyla. The results of two separate reconciliation programs, Jane and Ranger-DTL, were congruent with our earlier observation (Table 3.4). In the Alphaproteobacteria, 42% or 43% of gene acquisition events were vertical, while only 34% or 37% of events were horizontal. On the other hand, in the Cyanobacteria, 33% of gene transfer events were vertical, while 39% or 47% of events were horizontal. This pattern of inheritance suggests that 2-methylhopanoids serve a more useful function in the Alphaproteobacteria who make them than in the Cyanobacteria as a whole. Consistent with widespread interphylum horizontal gene transfer, *hpnP* was transferred into the Cyanobacteria after the group's major divergences and was not present in the last common ancestor of the phylum (Figure 3.4). We note that the diversity between the species trees under consideration is different; the cyanobacterial species tree includes 16S rDNA sequences from the entire phylum, whereas the alphaproteobacterial species tree focuses on a subset of an order, based on which organisms contain *hpnP*. This difference in diversity may have contributed to the dissimilarity in the inheritance mode we found. Nevertheless, the fact that *hpnP* is found sporadically throughout the Cyanobacteria yet concentrated within a monophyletic clade of alphaproteobacterial families itself may reflect different strategies in gene inheritance and transfer between organisms depending on the degree of their relatedness.

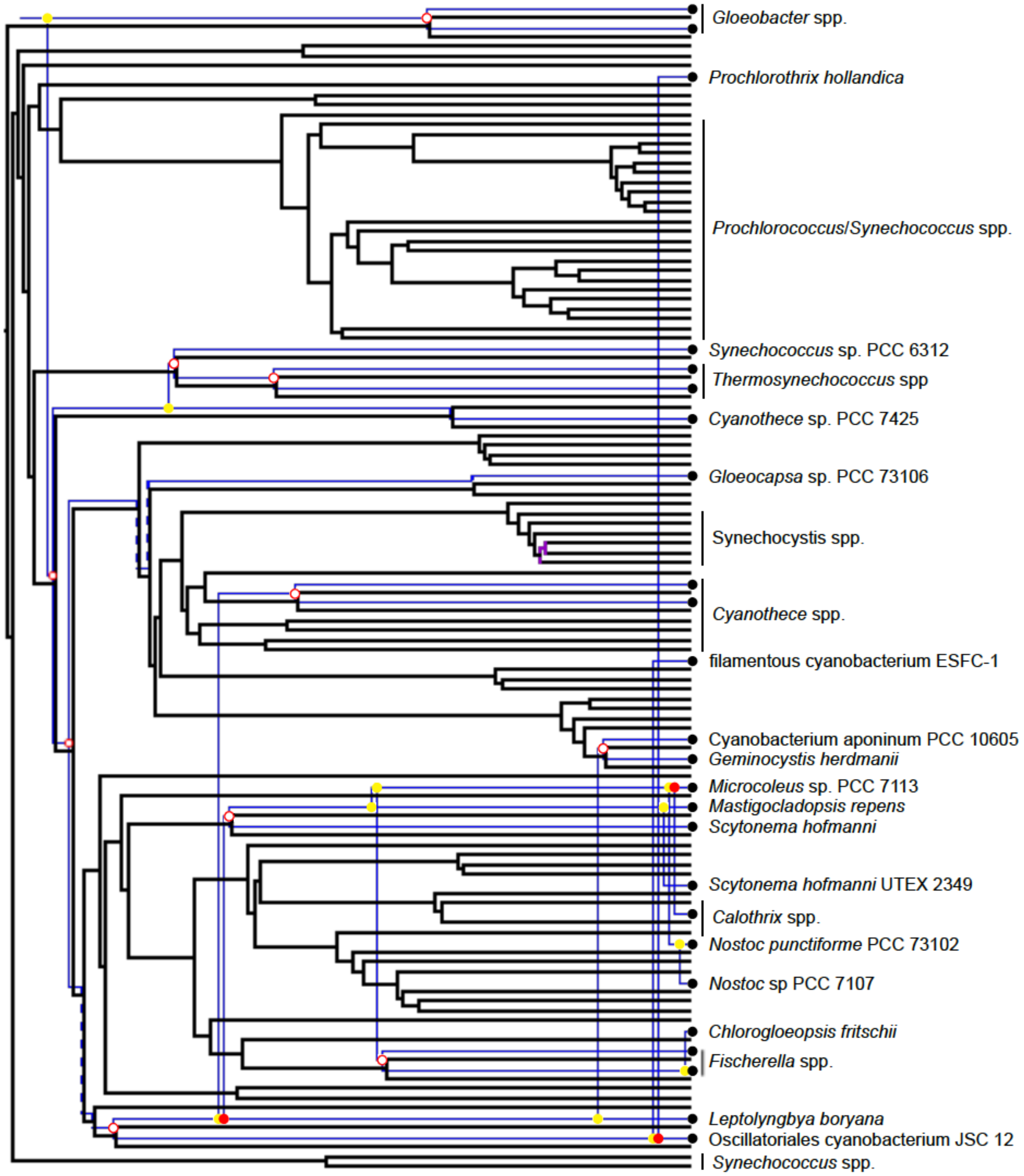


Figure 3.4 | Cyanobacterial HpnP and species tree reconciliation. The 16S rRNA species tree (black) is overlaid with the HpnP tree (blue). The reconciliation was completed in Jane using 1, 2, and 3 as the event costs for loss, duplication, and horizontal transfer, respectively. Vertical transfer – open circle, horizontal transfer – closed circle, loss – dashed line, red – best reconciliation, yellow – other equally good reconciliations, purple – polytomy due to identical sequences.

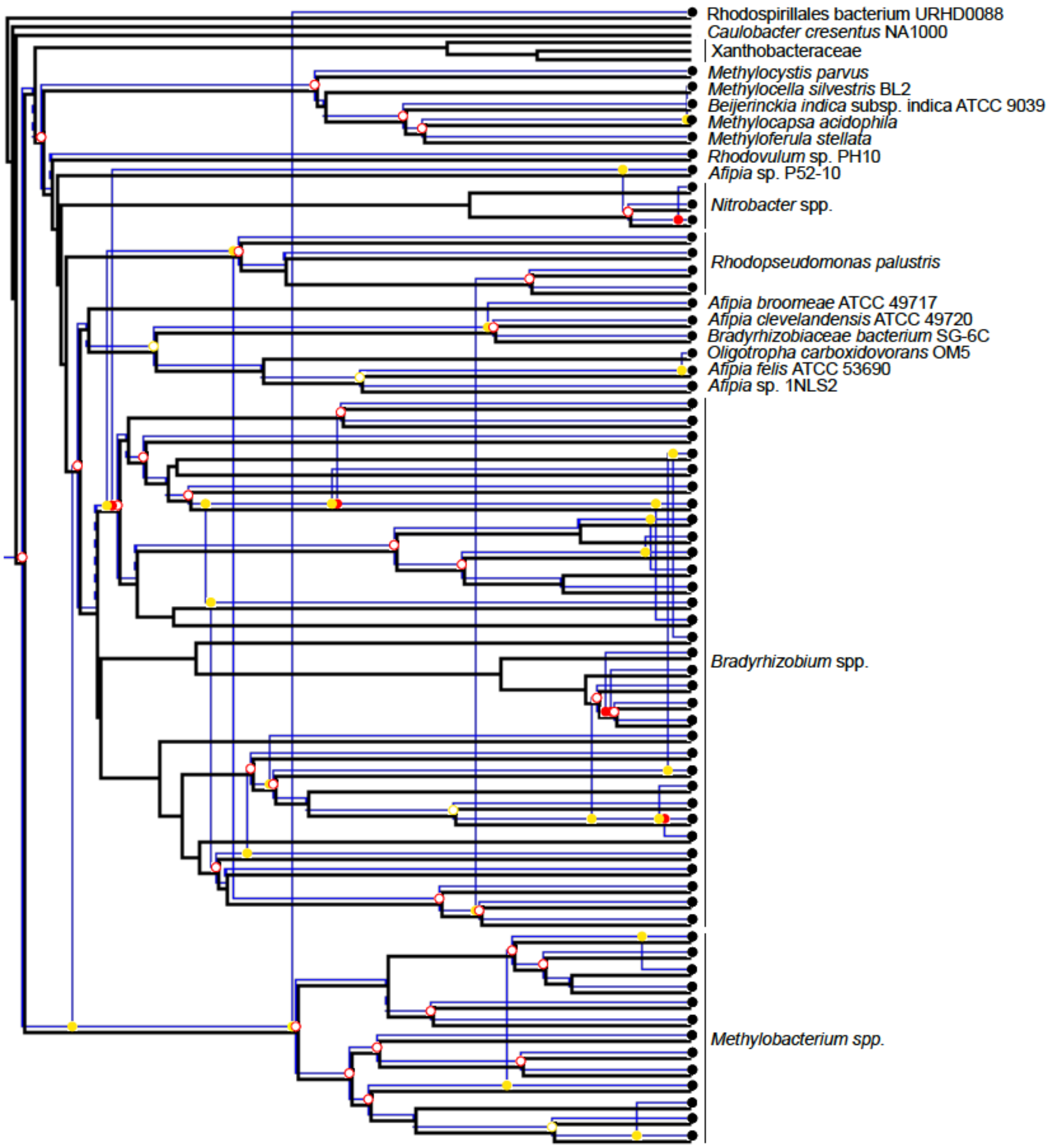

Rhodospirillales bacterium URHD0088
Caulobacter cresentus NA1000
Xanthobacteraceae
Methylocystis parvus
Methylocella silvestris BL2
Beijerinckia indica subsp. indica ATCC 9039
Methylocapsa acidophila
Methyloferula stellata
Rhodovulum sp. PH10
Afipia sp. P52-10
Nitrobacter spp.
Rhodopseudomonas palustris
Afipia broomeae ATCC 49717
Afipia clevelandensis ATCC 49720
Bradyrhizobiaceae bacterium SG-6C
Oligotropha carboxidovorans OM5
Afipia felis ATCC 53690
Afipia sp. 1NLS2
Bradyrhizobium spp.
Methylobacterium spp.

Figure 3.5 | Alphaproteobacterial HpnP and species tree reconciliation. The 16S rRNA species tree (black) is overlaid with the HpnP tree (blue). The reconciliation was completed in Jane using 1, 2, and 3 as the event costs for loss, duplication, and horizontal transfer, respectively. Vertical transfer – open circle, horizontal transfer – closed circle, loss – dashed line, red – best reconciliation, yellow – other equally good reconciliations.

Table 3.4. Summary of HpnP and species tree reconciliations

	Alphaproteobacteria		Cyanobacteria	
	Jane	Ranger-DTL	Jane	Ranger-DTL
Total Cost*	107	107	48	48
Total Events	82	85	30	33
Speciation	35	36	10	11
Transfer	30	29	14	13
Loss	17	20	6	9
Duplication	0	0	0	0

*Event costs: loss – 1, duplication – 2, transfer – 3

Gene and species tree reconciliations can also predict when a gene arose or was acquired. Based on our reconciliation, HpnP most likely originated in the last common ancestor of the Bradyrhizobiaceae, Beijerinckiaceae, Methylobacteriaceae, and Methylocystaceae families of the Alphaproteobacteria (Figure 3.5). *hpnP* was mainly inherited vertically throughout this clade (Table 3.4). From the Alphaproteobacteria, *hpnP* was horizontally transferred to the Cyanobacteria along the *Gloeobacter* spp. lineage (Figure 3.4). The C-2 methylase gene was then horizontally transferred to a deeply branching node that was the last common ancestor of *Nostoc* spp., *Cyanothece* spp. as well as many others, but it was not present in the last common ancestor of the phylum. Subsequently, *hpnP* was predominantly horizontally rather than vertically transferred within the cyanobacterial phylum. We note that in both the Alphaproteobacteria and Cyanobacteria HpnP and species tree reconciliations, *hpnP* was lost and then later re-acquired by some clades. At a relatively similar time as the horizontal transfer between Alphaproteobacteria and Cyanobacteria, *hpnP* appears to have also been laterally transferred to the Acidobacteria (Figure 3.3). Due to a paucity of information about *hpnP* in the Acidobacteria, it is unclear which phylum, Alphaproteobacteria or Cyanobacteria, donated *hpnP* to the Acidobacteria, and we do not know whether and how *hpnP* has been inherited between members of the Acidobacteria subsequently.

Discussion

Comparison to previous work

The significant expansion in available HpnP sequences over the past four years enabled us to identify an unambiguous and robust phylogenetic history for HpnP. The current dataset yields consistent outgroup placement regardless of the outgroup composition, trimming stringency, and addition of environmental sequences (Figure 3.2, Table 3.3). The elimination of the tree rooted in the Cyanobacteria under the majority of trials based on the AU test also lends support to our evolutionary model. These results contrast with our previous study (Welander et al., 2010), where three different outgroup placements were found in 18 trials and none of these alternate roots could be eliminated with statistical confidence.

Between 2010 and present, approximately four times the number of HpnP sequences have become available from genome sequencing efforts (Table 3.2). While many sequenced genomes had very similar HpnP, some targeted genomic and environmental sequencing efforts recovered new HpnP sequences (Shih et al., 2013; Ricci et al., 2014). These unique sequences were critical in enabling resolution of the root of the HpnP tree by increasing both diversity and balance. The increase in the number of sequences was not limited to HpnP, but also extended to neighboring class B Radical SAM enzymes with unknown function, allowing us to select an outgroup that is phylogenetically closer to HpnP than previously used (Figure 3.1A) (Welander et al., 2010; Zhang et al., 2012). Outgroup selection is critical to identifying the true ingroup topology because an outgroup that is too distant from the ingroup can cause long branch attraction (Smith, 1994). The combined effects of increased diversity in the ingroup and a more closely related outgroup led to a more robust HpnP topology.

Implications for the rock record

Our phylogenetic analysis suggests that 2-methylhopanoids are younger than previously thought. Based on the genomic composition of extant members of the alphaproteobacterial clade, the last common alphaproteobacterial ancestor is predicted to have been capable of aerobic respiration and adapted to an oxic environment (Boussau et al., 2004; Schoepp-Cothenet et al., 2009). Given our confidence in the origin of 2-methyhopanoid biosynthesis in the Alphaproteobacteria, this implies that 2-methylhopanoids appeared after the rise of oxygen. This interpretation supports the conclusion that 2-methylhopane deposits at 2.7 Ga are contaminants (Rasmussen et al., 2008; French et al., 2015) and is consistent with the earliest unambiguous 2-methylhopanes dating to 1.64 Ga (Brocks et al., 2005). Additionally, 2-methylhopanoids evolving after the rise of oxygen fits with our conclusion that Cyanobacteria horizontally acquired *hpnP* after the phylum diverged because estimates suggest that the ancestor of Cyanobacteria arose prior to the great oxidation event (Schirrmeister et al., 2013).

Because *hpnP* may have existed for some time in the Alphaproteobacteria prior to its transfer to the Cyanobacteria, it is possible that the earliest 2-methylhopanes have a solely alphaproteobacterial origin. Without distinct fossils to calibrate HpnP molecular evolution, we are unable to infer how long this time interval may have been. Yet the earliest unambiguous 2-methylhopane deposits in the Barney Creek Formation derive from environments that likely supported anoxygenic phototrophs based on the presence aromatic carotenoid biomarkers, such as okenane, chlorobactane, and isorenieratane, which are commonly used biomarkers for phototrophic Gammaproteobacteria and Chlorobi (Brocks & Schaeffer, 2008). Consistent with the presence of more general carotenoid biomarkers, such lycopene and the spirilloxanthin series, it is possible that this habitat harbored other anoxygenic phototrophs, including alphaproteobacteria. We note that gammacerane, made by many 2-methylhopanoid producing alphaproteobacteria, was not found in the Barney Creek

Formation, but its absence does not preclude the presence of alphaproteobacteria because tetrahymanol, the precursor of gammacerane, is synthesized in inverse abundance relative to 2-methylhopanoids (Neubauer et al., 2015) and may have a different preservation potential than other biomarkers.

Because HpnP originated in the Alphaproteobacteria, the first function of 2-methylhopanoids must have been useful to ancient alphaproteobacteria. We can therefore assume that 2-methylhopanoids are functionally decoupled from oxygenic photosynthesis or any other metabolism that was not present in ancient *hpnP*-containing alphaproteobacteria. Although it is possible that the ancestral function of 2-methylhopanoids or the physiology of its producers may no longer be the same in modern organisms, for the sake of argument, we will assume conservation in function. Given this assumption, what can we infer from patterns of 2-methylhopanoid occurrence in modern organisms? Today we know that many HpnP-containing alphaproteobacteria are anoxygenic phototrophs, diazotrophs, and can utilize methanol as a carbon source (Ricci et al., 2014). It is tempting to speculate that the ancestral function of 2-methylhopanoids might have provided an advantage to these strains, indirectly enabling them to perform these functions.

Recent biophysical experiments have revealed that methylation can enhance the ability of hopanoids to rigidify membranes under physiologically relevant conditions (Wu et al., 2015). Consistent with this, 2-methylhopanoids are enriched in the outer membrane of akinetes, tough survival cell types made by the cyanobacterium *N. punctiforme* (Doughty et al, 2009). In *R. palustris* TIE-1, 2-methylhopanoids are regulated via a pathway that responds to a variety of stressors (Kulkarni et al., 2013). Collectively, these observations suggest that 2-methylhopanoids play a role in stress resistance. Although there is no reason to think 2-methylhopanoids are required for any particular metabolism (Welander et al, 2009, Kulkarni et al, 2013), the significant correlation of *hpnP* presence with organisms, metabolisms, and environments that support plant-microbe interactions (Ricci et al, 2014) indicates that 2-methylhopanoids promote fitness in these contexts. For example, bacteria experience a variety of stresses (e.g., osmotic, low pH, etc.) in the process of establishing a functional symbiosis with plants, where they provide the plant fixed nitrogen (Gibson et al., 2009). Intriguingly, peaks in 2-methylhopane abundance in the Phanerozoic Eon have been correlated with episodes of ocean anoxic events (Knoll et al., 2007), and it has been suggested that the capacity for nitrogen fixation may have been helpful during such times (Kuypers et al., 2004). Because multiple types of stressors were likely operative during these episodes, 2-methylhopanoids may have provided a selective advantage. Despite recent progress, much remains to be learned regarding how 2-methylhopanoids, as well as other hopanoid types, contribute to fitness in modern organisms (Wu et al., 2015).

Uncertainties underpinning the HpnP phylogeny

While new sequence data may increase confidence in our conclusions, it is also possible that they will beget revisions. For example, additional HpnP sequences could support a different tree topology. If a different topology were chosen, it would be from the pool of possible HpnP topologies (Figure 3.1C, E). It is improbable that new data will find the cyanobacterial-rooted topology, as the most-supported because current data already eliminates this possibility (Figure 3.1D). Additionally, a special scenario could occur if new non-alphaproteobacterial HpnP sequences were found to form a paraphyletic clade surrounding the currently known HpnP sequences. This topology would suggest that HpnP originated in this new group rather than the Alphaproteobacteria. Regardless, it is unlikely that the Cyanobacteria could have composed this hypothetical paraphyletic clade for two reasons. First, a hypothetical cyanobacterial-rooted topology would imply multiple horizontal transfers of *hpnP* between cyanobacteria and alphaproteobacteria. We know that such an event happened once, but having multiple lateral transfers between distantly related organisms is less probable. Second, to fall in a basal position, additional cyanobacterial

HpnP sequences would need to be very different from ones currently known and would therefore likely be found in poorly studied regions of the phylum. A recent genome sequencing effort focused on gathering genomes from diverse cyanobacteria did not find any *hpnP* sequences that meet this criterion (Shih et al., 2013). We also note that no hopanoid biosynthesis genes have been identified in the non-photosynthetic proposed sister clade of Cyanobacteria (Di Rienzi et al., 2013). Given the improved coverage of the Cyanobacteria, it is unlikely that new HpnP sequences will be found that yield a topology incongruent with the Cyanobacteria acquiring the gene secondarily.

Conclusion

Current data provide strong support that HpnP and 2-methylhopanoids arose in the Alphaproteobacteria and were subsequently horizontally transferred into the Cyanobacteria. These conclusions require that, for some interval of time, alphaproteobacteria constituted the only flux of these molecules to sedimentary environments, which is consistent with a strict reading of the molecular fossil record. Additionally, because Alphaproteobacteria are thought to be ancestrally aerobic, 2-methylhopanoids did not likely appear until after the rise of oxygen. Furthermore, our findings suggest that the original function of 2-methylhopanoids did not involve oxygenic photosynthesis but instead was related to a feature present in the first 2-methylhopanoid-producing alphaproteobacterium. Some questions remain about the evolutionary history of HpnP. They include: from what enzyme did the HpnP protein evolve and what was its original function, how and under which conditions did *hpnP* transfer between phyla, and what is the evolutionary relationship between HpnP and HpnR, the C-3 hopanoid methylase? More information and sequence data are needed to address these questions. Regardless, our study demonstrates the power of using phylogenetic analyses to gain insight into problems of geobiological importance.

Acknowledgements

We would like to thank Woodward Fischer, James Hemp, Jena Johnson, and members of the Newman Lab for their helpful suggestions. We are grateful to John Spear for providing samples from Yellowstone National Park through the Yellowstone Center for Resources (permit #5664). This work was supported by grants from NASA (NNX12AD93G), the National Science Foundation (1224158), and the Howard Hughes Medical Institute (HHMI) to DKN. JNR is supported by an NSF graduate fellowship. DKN is an Investigator of the Howard Hughes Medical Institute.

References

Bansal MS, Alm EJ, Kellis M (2012) Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, **28**, i283–i291.

Beiko RG, Harlow TJ, Ragan MA (2005) Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 14332–14337.

Boussau B, Karlberg EO, Frank AC, Legault B, Andersson SGE (2004) Computational inference of scenarios for alphaproteobacterial genome evolution. *Proceedings of the National Academy of Sciences*, **101**, 9722–9727.

Brocks JJ, Love GD, Summons RE, Knoll AH, Logan GA, Bowden SA (2005) Biomarker evidence for green and purple sulphur bacteria in a stratified Palaeoproterozoic sea. *Nature*, **437**, 866–870.

Brocks JJ, Schaeffer P (2008) Okenane, a biomarker for purple sulfur bacteria (Chromatiaceae), and other new carotenoid derivatives from the 1640Ma Barney Creek Formation. *Geochimica et Cosmochimica Acta*, **72**, 1396–1414.

Colless DH (1982) Phylogenetics: The theory and practice of phylogenetic systematics. *Systematic Zoology*, **31**, 100–104.

Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R (2010) Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology*, **5**, 1–10.

Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, **27**, 1164–1165.

Di Rienzi SC, Sharon I, Wrighton KC, Koren O, Hug LA, Thomas BC, Goodrich JK, Bell JT, Spector TD, Banfield JF, Ley RE (2013) The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *eLife*, **2**, 1–25.

Di Rienzi SC, Sharon I, Wrighton KC, Koren O, Hug LA, Thomas BC, Goodrich JK, Bell JT, Spector TD, Banfield JF, Ley RE (2013) The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *eLife*, **2**, 1–25.

Doughty DM, Coleman ML, Hunter RC, Sessions AL, Summons RE, Newman DK (2011) The RND-family transporter, HpnN, is required for hopanoid localization to the outer membrane of *Rhodopseudomonas palustris* TIE-1. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, E1045–E1051.

Doughty DM, Dieterle M, Sessions AL, Fischer WW, Newman DK (2014) Probing the Subcellular Localization of Hopanoid Lipids in Bacteria Using NanoSIMS. *PloS one*, **9**, e84455.

Doughty DM, Hunter R, Summons RE, Newman DK (2009) 2-Methylhopanoids are maximally produced in akinetes of *Nostoc punctiforme*: geobiological implications. *Geobiology*, **7**, 524–532.

French KL, Hallman C, Hope JM, Schoon PL, Zumberge JA, Hoshino Y, Peters CA, George SC, Love GD, Brocks JJ, Buick R, Summons RE (2015) Reappraisal of hydrocarbon biomarkers in Archean rocks. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 5915-5920.

Gibson KE, Kobayashi H, Walker GC (2008) Molecular Determinants of a Symbiotic Chronic Infection. *Annual Review of Genetics*, 42, 413-441.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, **59**, 307–321.

Kannenberg EL, Poralla K (1999) Hopanoid Biosynthesis and Function in Bacteria. *Naturwissenschaften*, **86**, 168–176.

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, **9**, 286–298.

Knoll AH, Summons RE, Waldbauer JR, Zumberge JE (2007) The Geological Succession of Primary Producers in the Oceans. In: *The Evolution of Primary Producers in the Sea* (eds Falkowski P, Knoll AH), pp. 133–164. Academic Press, Boston.

Kulkarni G, Wu C-H, Newman DK (2013) The general stress response factor EcfG regulates expression of the C-2 hopanoid methylase HpnP in *Rhodopseudomonas palustris* TIE-1. *Journal of Bacteriology*, **195**, 2490–2498.

Kuypers MMM, van Breugel Y, Schouten S, Erba E, Sinninghe Damsté JS (2004) N2-fixing cyanobacteria supplied nutrient N for Cretaceous oceanic anoxic events. *Geology*, **32**, 853–856.

Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, **23**, 127–128.

Maddison W, Maddison D (2006) StochChar: A package of Mesquite modules for stochastic models of character evolution. Version 1.1.

Maddison W, Maddison D (2011) Mesquite: A modular system for evolutionary analysis. Version 2.75. http://mesquiteproject.org.

Maddison W, Maddison D, Midford P (2011) Tree Farm package for Mesquite, version 2.75.

Neubauer C, Dalleska NF, Cowley ES, Shikuma NJ, Wu C-H, Sessions AL, Newman DK (2015) Loss of hopanoid methylation leads to lipid remodeling and differential subcellular localization in *Rhodopseudomonas palustris* TIE-1. *Geobiology*, accepted.

Omland KE, Cook LG, Crisp MD (2008) Tree thinking for all biology: the problem with reading phylogenies as ladders of progress. *BioEssays*, **30**, 854–867.

Poralla K, Hartner T, Kannenberg E (1984) Effect of Temperature and pH on the Hopanoid Content of *Bacillus acidocaldarius*. *FEMS Microbiology Letters*, **23**, 253–256.

Rashby SE, Sessions AL, Summons RE, Newman DK (2007) Biosynthesis of 2-methylbacteriohopanepolyols by an anoxygenic phototroph. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 15099–15104.

Rasmussen B, Fletcher IR, Brocks JJ, Kilburn MR (2008) Reassessing the first appearance of eukaryotes and cyanobacteria. *Nature*, **455**, 1101–1104.

Ricci JN, Coleman ML, Welander P V, Sessions AL, Summons RE, Spear JR, Newman DK (2014) Diverse capacity for 2-methylhopanoid production correlates with a specific ecological niche. *The ISME journal*, **8**, 675–684.

Sáenz JP, Eglinton TI, Summons RE (2011) Abundance and structural diversity of bacteriohopanepolyols in suspended particulate matter along a river to ocean transect. *Organic Geochemistry*, **42**, 774–780.

Schirrmeister BE, de Vos JM, Antonelli A, Bagheri HC (2013) Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 1791–1796.

Schoepp-Cothenet B, Lieutaud C, Baymann F, Verméglio A, Friedrich T, Kramer DM, Nitschke W (2009) Menaquinone as pool quinone in a purple bacterium. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 8549–8554.

Shih PM, Matzke NJ (2013) Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 12355–60.

Shih PM, Wu D, Latifi A, Axen SD, Fewer DP, Talla E, Calteau A, Cai F, Tandeau de Marsac N, Rippka R, Herdman M, Sivonen K, Coursin T, Laurent T, Goodwin L, Nolan M, Davenport KW, Han CS, Rubin EM, Eisen JA, Woyke T, Gugger M, Kerfeld CA (2013) Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 1053–1058.

Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, **51**, 492–508.

Shimodaira H, Hasegawa M (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, **17**, 1246–1247.

Smith AB (1994) Rooting molecular trees: problems and strategies. *Biological Journal of Linnean Society*, **51**, 279–292.

Summons RE, Jahnke LL, Hope JM, Logan GA (1999) 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature*, **400**, 554–557.

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, **56**, 564–577.

Uchiyama T, Watanabe K (2006) Improved inverse PCR scheme for metagenome walking. *BioTechniques*, **41**, 183–188.

Vogl K, Bryant D (2012) Biosynthesis of the biomarker okenone: χ-ring formation. *Geobiology*, **10**, 205–215.

Welander P V, Coleman ML, Sessions AL, Summons RE, Newman DK (2010) Identification of a methylase required for 2-methylhopanoid production and implications for the interpretation of sedimentary hopanes. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 8537–8542.

Welander P V, Doughty DM, Wu C-H, Mehay S, Summons RE, Newman DK (2012) Identification and characterization of *Rhodopseudomonas palustris* TIE-1 hopanoid biosynthesis mutants. *Geobiology*, **10**, 163–177.

Welander P V, Hunter RC, Zhang L, Sessions AL, Summons RE, Newman DK (2009) Hopanoids play a role in membrane integrity and pH homeostasis in *Rhodopseudomonas palustris* TIE-1. *Journal of Bacteriology*, **191**, 6145–6156.

Williams KP, Sobral BW, Dickerman AW (2007) A robust species tree for the Alphaproteobacteria. *Journal of Bacteriology*, **189**, 4578–8456.

Wu C-H, Bialecka-Fornal M, Newman DK (2015) Methylation at the C-2 position of hopanoids increases rigidity in native bacterial membranes. *eLife* 2015;10.7554/eLife.05663.

Yang Z, Rannala B (2012) Molecular phylogenetics: principles and practice. *Nature Reviews Genetics*, **13**, 303–314.

Zhang Q, van der Donk WA, Liu W (2012) Radical-Mediated Enzymatic Methylation: A Tale of Two SAMS. *Accounts of Chemical Research*, **45**, 555–564.

Chapter 4

THE ROLE OF HOPANOIDS AND 2-METHYLHOPANOIDS IN *NOSTOC PUNCTIFORME*

Introduction

Hopanoids are pentacyclic triterpenoids produced by select bacteria and have structural similarity to cholesterol. Their hydrocarbon skeletons can be preserved in the geologic record on a billion-year timescale (Brocks et al. 2005), and they have the potential to provide insight into ancient microbial life and the conditions faced by it in the past. A particular subclass of hopanoids, 2-methylhopanoids, were once thought to be biomarkers of Cyanobacteria and their main metabolism oxygenic photosynthesis (Summons et al. 1999). It has now been shown that organisms other than Cyanobacteria also produce 2-methylhopanoids and that not all Cyanobacteria produce these lipids (Rashby et al. 2007; Talbot et al. 2008; Welander et al. 2010). Although the hypothesis that 2-methylhopanoids are indicators of Cyanobacteria is incorrect, the question of what hopanoids and their C-2 methylated counterparts are unique biomarkers for remains open. Progress has been made on understanding that the environmental distribution of 2-methylhopanoids is taxonomically diverse and that their initial evolution occurred in the Alphaproteobacteria (Ricci et al. 2014; Ricci et al 2015). Studies exploring the biological function of hopanoids in modern organisms have occurred mainly in Proteobacteria and have found that hopanoids assist in rigidifying the cell membrane, aid in stress tolerance, and are important in plant-microbe symbioses (Welander et al. 2009; Schmerk et al. 2011; Kulkarni et al. 2013; Silipo et al. 2014; Ricci et al. 2014; Wu et al. 2015a). Limited studies of hopanoids in Cyanobacteria have focused on their location in the cell as well as on the production of squalene, the hopanoid precursor, for biotechnology applications (Doughty et al. 2009; Englund et al. 2014). Here, we sought to understand the physiological role of hopanoids and 2-

methylhopanoids in Cyanobacteria by following up on initial work in *Nostoc punctiforme* (Doughty et al. 2009).

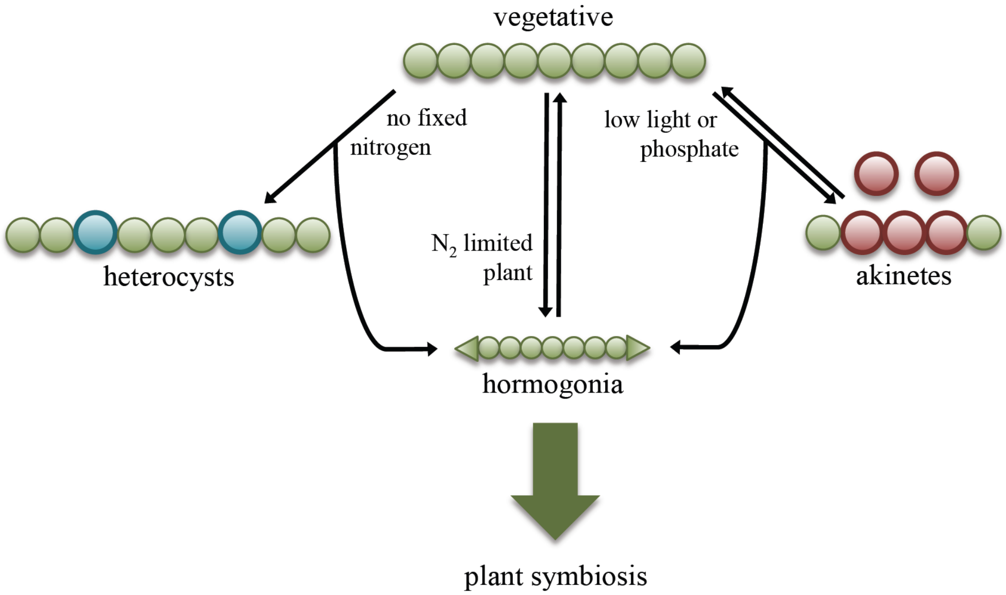


Figure 4.1 | Cell types and cycle of *N. punctiforme*. *N. punctiforme* filaments forms 4 cell types: photosynthetic vegetative, nitrogen fixing heterocysts, spore-like akinetes, and motile hormogonia.

N. punctiforme is a filamentous diazotrophic symbiotic cyanobacterium that lives in a variety of freshwater and terrestrial habitats. As part of its life cycle, *N. punctiforme* forms four distinct cell types (Figure 4.1) (Meeks et al. 2002). Under non-nutrient-limiting conditions, filaments are exclusively composed of photosynthetic vegetative cells. When nitrogen-limited, approximately 8% of vegetative cells terminally differentiate into heterocysts, which fix and distribute nitrogen to neighboring cells. The creation of a spatial-separated cell type dedicated to nitrogen fixation allows *N. punctiforme* to keep its oxygen-sensitive nitrogenase away from oxygen generated by photosynthesis (Flores & Herrero 2010). Spore-like akinetes form when *N. punctiforme* is exposed to other stresses, such as low light or phosphate, and can survive under harsher conditions than vegetative cells, including extreme cold and desiccation (Wong & Meeks 2002; Argueta &

Summers 2005). Finally, vegetative cells can become motile hormogonia when exposed to certain stresses, such as transfer to fresh growth medium or nitrogen limitation. It is unknown why filaments become hormogonia rather than forming another cell type, but it has been suggested that hormogonia formation may be due to a confluence of factors as opposed to a single stressor (Flores & Herrero 2010). Hormogonia are generated through multiple rounds of cell division without the addition of biomass and can be easily characterized by their small cell size, pointed terminal cells, and the presence of gas vesicles. After a 48–72 hour period hormogonia cease to be motile and differentiate back into vegetative filaments, which may contain heterocysts (Campbell & Meeks 1989).

N. punctiforme forms symbioses with multiple diverse plant species, including members of the hornworts, liverworts, gymnosperms, and angiosperms (Meeks & Elhai 2002). In each of these symbioses, the plant partner initiates the interaction when it is limited for fixed nitrogen by signaling to *N. punctiforme* to form hormogonia, glide toward it, and settle at the site of infection. Once there, *N. punctiforme* forms vegetative filaments with a greater number of heterocysts than found when free-living and begins transferring fixed nitrogen to its host, while the plant partner provides a carbon source to the bacterium. One particularly well-studied *N. punctiforme* symbiosis is its interaction with the hornwort *Anthoceros punctatus* (Wong & Meeks 2002; Meeks 2009). In this association, *N. punctiforme* infects the mucilaginous cavities at the growing edge of gametophyte tissue to create visible symbiotic colonies, which are composed of 25% heterocysts (Meeks & Elhai 2002).

Previous work on the role of hopanoids in *N. punctiforme* has focused on characterizing their abundance in the organism's non-symbiotic life cycle. This analysis concluded that hopanoids, as well as 2-methylhopanoids, were most abundant in the outer membranes of akinetes (Doughty et al. 2009). Yet, exclusively 2-methylhopanoids were found in the outer membrane of heterocysts and the thylakoid membrane of their associated vegetative cells (Figure 1.3), suggesting that they are important in this locale compared to unmethylated hopanoids. We hypothesized that hopanoids may be important in particular cell types with and without various stressors due to their location in the cell and their role in other species. We also speculate that hopanoids function in *N. punctiforme*'s symbiotic associations based on physiological studies in other organisms and plant-microbe correlations with hopanoids. In this study, we used a genetic approach to address these questions and understand the biological role of both hopanoids and 2-methylhopanoids in *N. punctiforme* more broadly.

Methods

Bacterial and plant growth conditions

N. punctiforme ATCC 29133 (PCC 73102) WT was obtained from John Meeks, and an S variant of the same species was acquired from Michael Summers, referred to as *N. punctiforme* ATCC 29133 and *N. punctiforme* S respectively (Table 1). While *N. punctiforme* ATCC 29133 has full symbiotic functionality, *N. punctiforme* S does not; yet *N. punctiforme* S grows more homogeneously in liquid medium than *N. punctiforme* ATCC 29133. *N. punctiforme* strains and mutants were routinely grown in quarter strength Allen and Arnon medium (AA/4) or in full strength Allen and Arnon medium with 1% noble agar (Allen & Arnon 1955). The medium was supplemented with 2.5 mM nitrate, 5 mM ammonium, and 5 mM morpholinepropanesulfonic acid (MOPS) pH 7.8 to achieve vegetative growth unless otherwise stated. Cultures were incubated at 25°C with 100 rpm shaking under cool white fluorescent lights at 15–19 μmol photon/m^2/s. When stated cultures were incubated at 13°C and 40°C with all other conditions the same.

For spot stress assays, 50 ml *N. punctiforme* S cultures grown to 10 μg chlorophyll a per ml were harvested by centrifugation at 2000 x g for 5 min. Cells were resuspended in fresh medium and separated into 10 ml aliquots and supplemented with, no addition, 0.45 mM EDTA, 300mM NaCl, or 300 mM mannitol and incubated under standard conditions for 3 days. Five µl of 2 fold serial dilutions were spotted on solid medium and incubated for 10 days. For desiccation assays, 5 µl of 2 fold serial dilutions of vegetatively grown cultures were spotted onto filters. Once dry, the control filter was transferred to AA solid medium and incubated under standard conditions for 10 days. The remaining filters were incubated in otherwise empty Petri dishes under standard conditions for 3 or 8 hours before being transferred to AA solid medium and incubated for 10 days. Doubling times for growth curves were calculated by monitoring µg chlorophyll a per ml of culture (Meeks et al. 1983).

A. punctatus was grown in Hutner's medium supplemented with 5 mM 2-(N-morpholino)ethanesulfonic acid (MES) and 0.5% glucose pH 6.4 and incubated at 20°C with 50 rpm shaking under cool white fluorescent lights at 16 µmol photon/m^2/s on 14:10 hour light:dark cycle (Enderlin & Meeks 1983). To reconstitute the *N. punctiforme* and *A. punctatus* symbiosis, 100 µl of 10 µg chlorophyll a per ml vegetatively grown *N. punctiforme* ATCC 29133 was added to 5 g of *A. punctatus* in 10 ml of Hutner's without ammonium nitrate. After incubating for 2 weeks under standard *A. punctatus* growth conditions, the plant tissue was examined for *N. punctiforme* colonies under the dissecting scope (Wong & Meeks 2002).

Escherichia coli DH5α-MCR strains were grown in lysogeny broth (LB) at 37°C with 250 rpm shaking. Cultures were supplemented with 25 µg/ml kanamycin or 30 µg/ml chloramphenicol when appropriate.

Construction of Nostoc punctiforme hopanoid mutants

Using the same plasmid constructs, *shc* and *hpnP* were deleted in *N. punctiforme* S and *shc* was deleted in *N. punctiforme* ATCC 29133. Mutants were constructed by interrupting genes of interest with the omega neomycin phosphotransferase gene cassette (Ω-*npt*) (Cohen & Meeks 1997). Plasmids pJR101 and pJR102, targeted against *shc* and *hpnP* respectively, were introduced by triparental conjugation and selected for integration into the chromosome with 10 μg/ml neomycin for 2 weeks on plates. After colonies grew up on plates, they were transferred to liquid medium with selection and sequentially transferred for 3 months to out grow *E. coli*. Elimination of the plasmid was achieved with neomycin selection and 5% (w/v) sucrose counterselection for 1 month (Cai & Wolk 1990).

N. punctiforme S Δ*shc*::Ω-*npt* (Δ*shc*) was complemented with pJR103 inserted into the strain through electroporation and selection with 5 μg/ml ampicillin, as described previously (Summers et al. 1995), to create *N. punctiforme* S Δ*shc*::Ω-*npt* pSCR202::*shc* (Δ*shc* complement). Control WT and mutant strains were created by introducing empty pSCR202 into *N. punctiforme* S WT and Δ*shc* as described above. These control strains were used in experiments that included the complemented *shc* mutant.

All plasmids and strains used in this study can be found in Table 4.1. DNA sequences for all plasmids and cloning intermediates were confirmed by sequencing at Retrogen (http://sequencing.retrogen.com/).

Hopanoid analysis and lipidomics

50 ml *N. punctiforme* S WT, Δ*shc*, and Δ*hpnP*::Ω-*npt* (Δ*hpnP*) cultures were grown vegetatively to 10 μg chlorophyll a per ml. Total lipid extracts were obtained by micro-scale extraction as described previously (Wu et al. 2015b). In brief, 1 ml was harvested by centrifugation at 14,000 x g for 2 min and resuspended in 50 μl of ddH_2O in a 400 μl glass insert within a gas chromatography (GC) vial. Cells were extracted with 125 μl methanol and 62.5 μl dichloromethane (DCM) and sonicated for 30 min at room temperature. An additional 200 μl DCM was added and samples were mixed. The

organic layer was collected and divided in half. The aliquot for GC-MS hopanoid analysis was dried under N_2 at room temperature, then acetylated with 100 µl 1:1 pyridine:acetic anhydride for 30 min at 60°C. The acetylated sample was run on a Restek Rxi-XLB column (30 m x 0.25 mm x 0.1 µm) in a Thermo Scientific TraceGC coupled to an ISQ mass spectrometer as described in Welander et al. (2009).

The total lipid extracts reserved for LC-MS were resuspended in 9:1 methanol:DCM and run on an Acquity I-Class UPLC coupled to a Xevo G2-S TOF mass spectrometer (Waters Corporation). Samples (injection volume 5 µl) were run in instrument triplicates and in randomized order. LC-TOF-MS^E data was collected in positive mode using electrospray ionization (ESI). Separation of intact polar lipids was achieved on an Acquity UPLC CSH C18 column (2.1 x 100 mm, 1.7 µm, Waters Corporation) following a protocol established by the manufacturer (Isaac et al. 2011). Mass features are defined as ions with a unique m/z and retention time and were detected and quantified in R version 3.1.3 (R Core Team 2014) using the xcms package (Smith et al. 2006). Principal component analysis (PCA) was performed using the function *prcomp* in R version 3.1.3 (R Core Team 2014). Only mass features of intensities greater than 1% relative to the maximum detected intensity were included in PCA analysis. LC-MS relative peak intensities were calculated by dividing the uncorrected peak intensities by the maximum peak intensity of that particular sample. Relative peak intensities of the mutants were normalized to the WT chromatographs by subtracting the WT trace from those of the mutants.

Results and Discussion

N. punctiforme hopanoid biosynthetic mutants

Mutants unable to synthesize all hopanoids and 2-methylhopanoids specifically were created to study the function of hopanoids in *N. punctiforme*. Based on homology to *Rhodopseudomonas* *palustris* TIE-1 *shc* and *hpnP*, we identified *shc* (41% amino acid identity), which encodes for the first step in hopanoid biosynthesis, and *hpnP* (58% amino acid identity), which encodes the C-2 methylase. We deleted *shc* and *hpnP* by insertion of the Ω-*npt* cassette (Table 4.1). These two strains were employed because *N. punctiforme* ATCC 29133 was symbiotically competent, while *N. punctiforme* S was not, but *N. punctiforme* S was easier to manipulate in free-living growth conditions. From these reasons, *N. punctiforme* ATCC 29133 was used for symbiotic assays and *N. punctiforme* S was used for all other experiments.

Table 4.1 | Strains and plasmids used in this study

Strain or plasmid	Genotype, description, or construction	Reference or source
Strains		
N. punctiforme ATCC 29133	Wild-type from J. Meeks (PCC 73102)	Rippka & Herdman, 1992
DKN 1650	*N. punctiforme* ATCC 29133 Δ*shc*::Ω-*npt*	This study
N. punctiforme S	Derivative of wild-type from M. Summers	M. Summers
DKN 1648	*N. punctiforme* S Δ*shc*::Ω-*npt* (**Δ*shc***)	This study
DKN 1649	*N. punctiforme* S Δ*hpnP*::Ω-*npt* (**Δ*hpnP***)	This study
DKN 1683	*N. punctiforme* S pSCR202	This study
DKN 1684	*N. punctiforme* S Δ*shc*::Ω-*npt* pSCR202	This study
DKN 1685	*N. punctiforme* S Δ*shc*::Ω-*npt* pSCR202::*shc* (**Δ*shc* complement)**	This study
E. coli DH5α-MCR	Methylation-dependent restriction-defective derivative of strain DH5α	Grant et al., 1990
Plasmids		
pSCR9	Ω-*npt* cassette source vector; Nm^r/Km^r	Cohen & Meeks, 1997
pRL271	Conjugatable non-replicating cyanobacterial vector; *sacB* Em^r Cm^r	Cai & Wolk, 1990
pSCR202	Complementation plasmid, Ap^r	Summers et al., 1995
pJR101	1.9kb SpeI-XhoI amplicon from genomic DNA containing *shc* with Ω-*npt* inserted at NdeI cloned into SpeI & XhoI of pRL271; *sacB* Em^r Cm^r Nm^r/Km^r	This study
pJR102	2.5kb SpeI-XhoI amplicon form genomic DNA containing *hpnP* with Ω-*npt* inserted at EcoRV cloned into SpeI & XhoI of pRL271; *sacB* Em^r Cm^r Nm^r/Km^r	This study
pJR103	1.9kb SalI-SmalI amplicon from genomic DNA containing *shc* inserted into the same sites of pSCR202, Ap^r	This study

Abbreviations: Ap ampicillin, Cm chloramphenicol, Em erythromycin, Km kanamycin, Nm neomycin, r resistance

To confirm that these hopanoid mutants did not make their respective downstream products, we analyzed their hopanoid composition by established GC-MS methods (Figure 4.2). In WT *N. punctiforme* strains we identified the following hopanoids: diploptene (II), 2-methylbacteriohopanetetrol (III), bacteriohopanetetrol (IV), and 2-methylbacteriohopanepentol (V). We also identified squalene (I) in both WT strains but only labeled its associated peak in *N. punctiforme* ATCC 29133 because its peak was near baseline in *N. punctiforme* S. No hopanoids were identified in both Δ*shc* strains, and an increase in the abundance of squalene was observed in *N. punctiforme* ATCC 29133. In *N. punctiforme* S Δ*hpnP*, no 2-methylhopanoids were identified. An increase in the abundance of bacteriohopanetetrol and appearance of bacteriohopanepentol (VI) was observed. The absence of hopanoids or 2-methylhopanoids and the increase of their respective biosynthetic precursors validated the *shc* and *hpnP* mutant constructs.

The major hopanoids in *N. punctiforme* were confirmed to be bacteriohopanetetrol, 2-methylbacteriohopanetetrol, and 2-methylbacteriohopanepentol (Figure 4.2) (Doughty et al. 2009). In *N. punctiforme* ATCC 29133, the C_{30} hopanoid diploptene was also observed but was not found in *N. punctiforme* S. Bacteriohopanepentol was below the limit of detection in both WT strains but appears to be an intermediate in the synthesis of 2-methylbacteriohopanepentol as it accumulated when *hpnP* was not present. Similar biosynthesis of other methylated hopanoids, where *hpnP* methylates otherwise complete hopanoid end products, has been observed in *R. palustris*. We did not observe the production of bacteriohopane cyclitol ether, which was reported in an earlier study (Doughty et al. 2009). More recent work has identified *hpnJ* as the gene responsible for bacteriohopane cyclitol ether (Schmerk et al. 2015); a homology search could not identify *hpnJ* in *N. punctiforme*. These two lines of evidence suggest that *N. punctiforme* does not make bacteriohopane cyclitol ether.

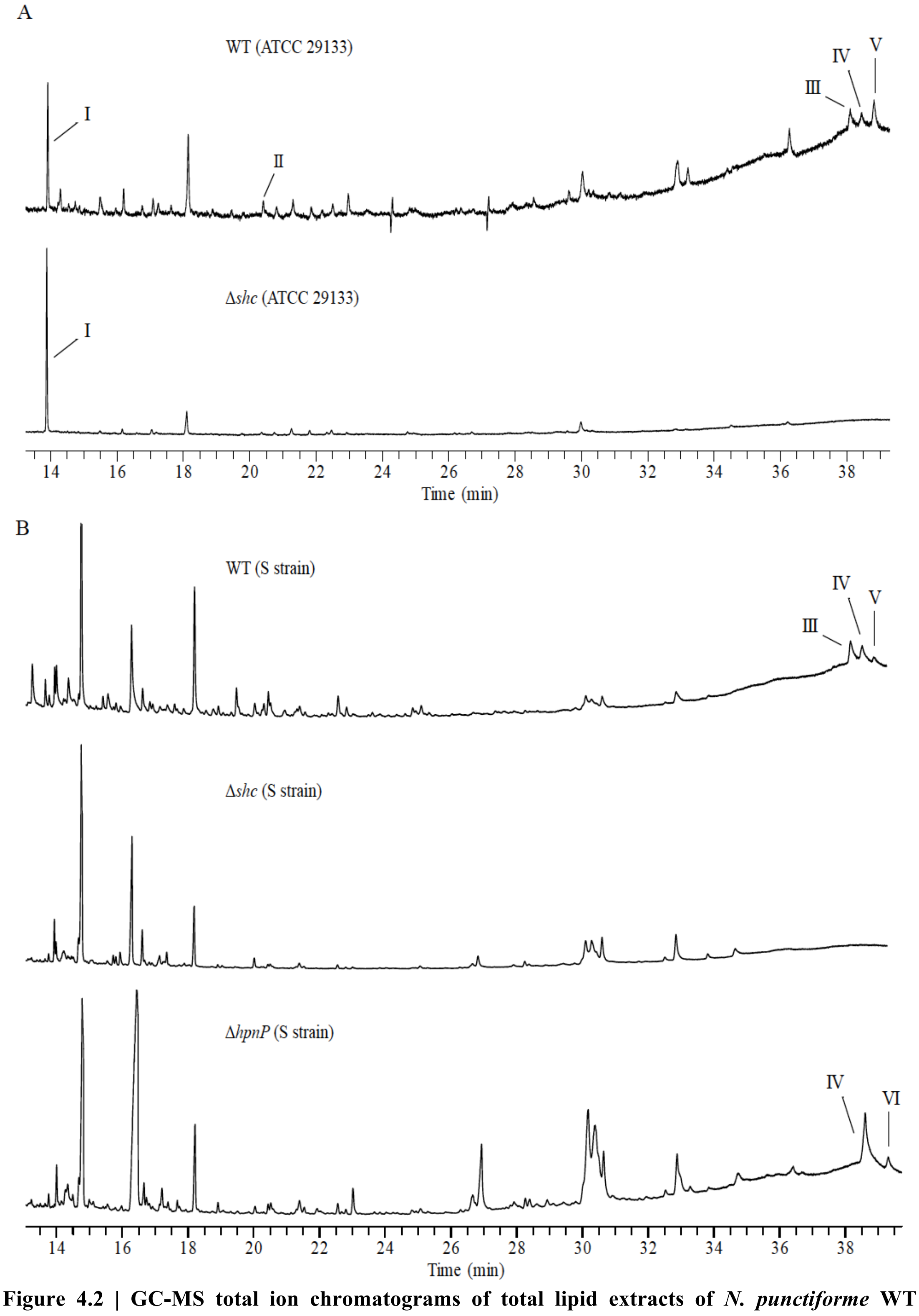


Figure 4.2 | GC-MS total ion chromatograms of total lipid extracts of *N. punctiforme* WT and hopanoid mutants by GC-MS. *N. punctiforme* Δ*shc* and Δ*hpnP* do not make hopanoids and 2-methylhopanoids respectively. (A) *N. punctiforme* ATCC 29133 (B) *N. punctiforme* S. Abbreviations include I squalene, II diplotene, III 2-methylbacteriohopanetetrol, IV bacteriohopanetetrol, V 2-methylbacteriohopanepentol, VI bacteriohopanopentol.

We compared the lipidomes of WT and hopanoid mutants to detect whether the absence of hopanoids or 2-methylhopanoids caused *N. punctiforme* S to change the composition of other lipids in its membrane. We found that the lipidomes of vegetative WT, Δ*shc*, and Δ*hpnP* were remarkably similar with less than 20% change in height between all detectable lipid peaks (Figure 4.3A). In contrast, the lipidomes of *R. palustris* TIE-1 WT, Δ*shc*, and Δ*hpnP* show many differences with some lipid abundances changing more than 80% (Figure 4.3B). These data suggest that *N. punctiforme* S does not compensate extensively for lack of hopanoids or 2-methylhopanoids under this condition.

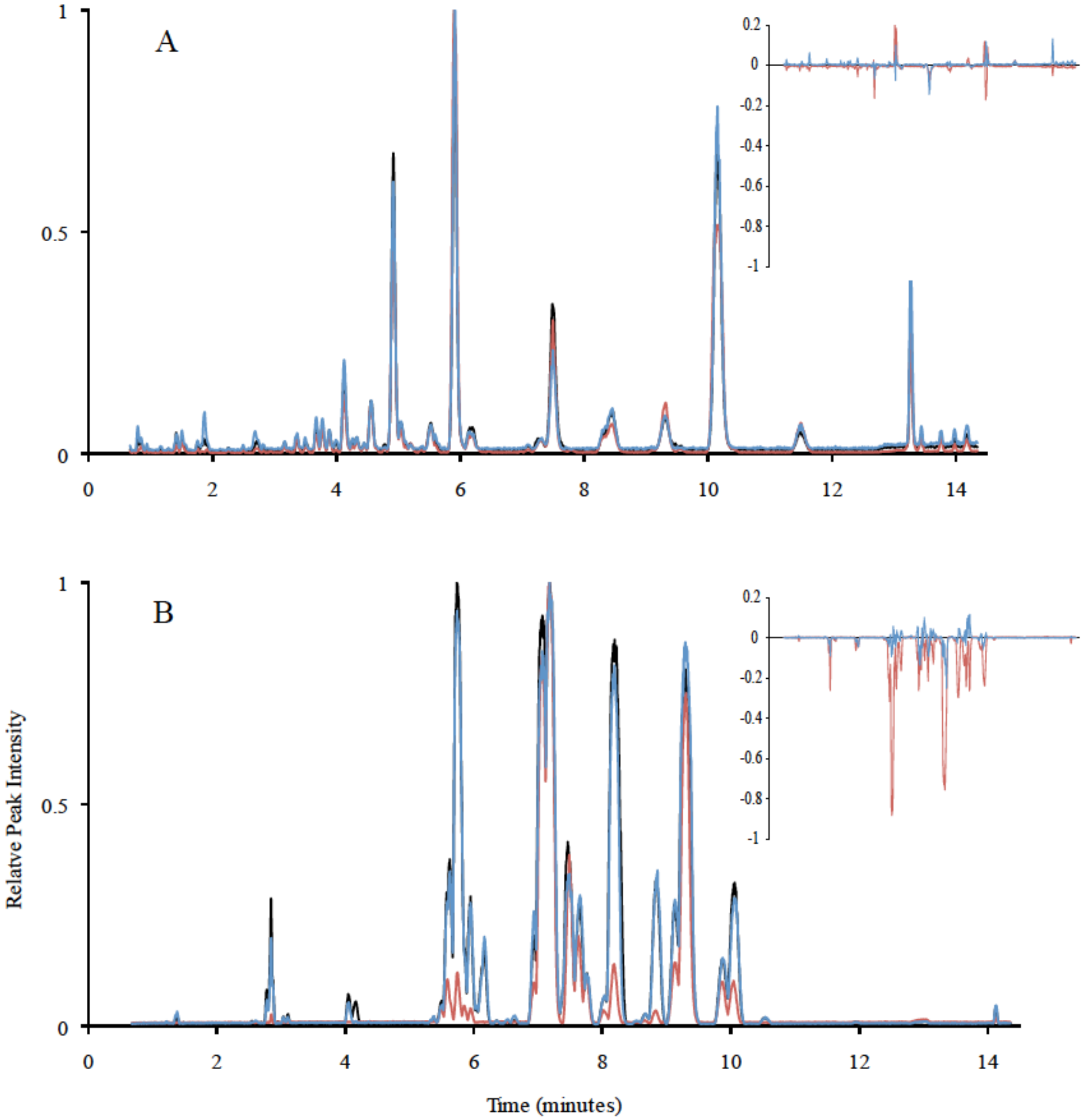


Figure 4.3 | Lipidomes of vegetative *N. punctiforme* hopanoid mutants show little variation compared to the hopanoid mutants of *R. palustris* TIE-1. (A) Overlay of *N. punctiforme* S WT, Δ*shc*, and Δ*hpnP* TLE LC-MS chromatographs. Inset contains *N. punctiforme* S Δ*shc* and Δ*hpnP* TLE LC-MS chromatographs normalized to WT (B) Overlay of *R. palustris* TIE-1 WT, Δ*shc*, and Δ*hpnP* TLE LC-MS chromatographs. Inset contains *R. palustris* TIE-1 Δ*shc* and Δ*hpnP* TLE LC-MS chromatographs normalized to WT.

Characterization of the hopanoid-lacking mutant grown under extreme temperatures

The lack of hopanoids affects the growth rate of *N. punctiforme* at extreme temperatures. Under standard temperature, 25°C, there was no difference in doubling time between *N. punctiforme* S WT and Δ*shc* (Figure 4.4). At elevated temperature, 40°C, *N. punctiforme* S Δ*shc* had a significantly larger doubling time compared to WT, while at 13°C *N. punctiforme* S Δ*shc* had a significantly smaller doubling time than WT. Similar to *R. palustris* Δ*shc*, *N. punctiforme* S Δ*shc* has a larger doubling time at elevated temperature (Kulkarni et al. 2013) (Figure 4.4). These phenotypes at extreme temperatures are consistent with hopanoids rigidifying the membrane. Some hopanoids, particularly 2-methylbacteriohopanetetrol, have been shown to rigidify cell membranes (Sáenz et al. 2012; Wu et al. 2015a). It is likely that the lack of hopanoids in *N. punctiforme* S Δ*shc* caused this strain to have a more fluid membrane than WT, leading to a growth defect at a high temperature when membranes are more fluid in general. Furthermore, the *N. punctiforme* S hopanoid-lacking mutant has a faster growth rate than WT at low temperature (Figure 4.4). At cold temperatures, when membranes are often more rigid, the absence of hopanoids may allow the membrane to be more fluid. We speculate that the partial return to standard temperature membrane fluidity allows the strain to achieve a higher growth rate.

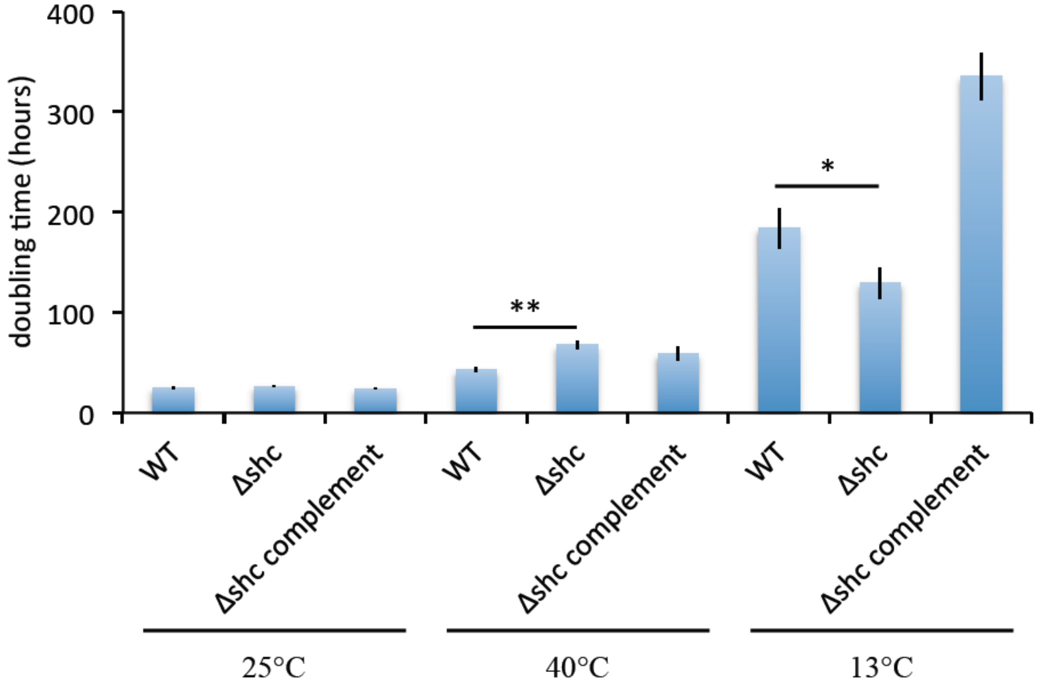


Figure 4.4 | Growth rate changes of *N. punctiforme* S hopanoid-lacking mutant under varying temperatures. *N. punctiforme* Δ*shc* has growth phenotypes at high and low temperature. Biological triplicate growth curves were measured for each condition. Error bars indicate standard deviation. * equals a p-value of less than 0.05 and ** represents a p-value of less than 0.01 by student's t test assuming a double-tailed distribution and equal variance between datasets.

To gain insight into the role of hopanoids at extreme temperatures, we quantified their abundances under these conditions. The total abundance of hopanoids in WT under different temperatures varied little, but the composition of hopanoids changed (Figure 4.5). At 40°C in WT, no 2-methylbacteriohopanepentol was observed, while bacteriohopanetetrol and 2-methylbacteriohopanetetrol were more abundant than at 25°C. This change in abundance of 2-methylbacteriohopanetetrol may play a role in the differential growth rate of *N. punctiforme* Δ*shc* at high temperature. Future work will be needed to confirm this observation, including identifying the gene responsible for bacteriohopanepentol biosynthesis and carrying out phenotypic analysis of the mutant.

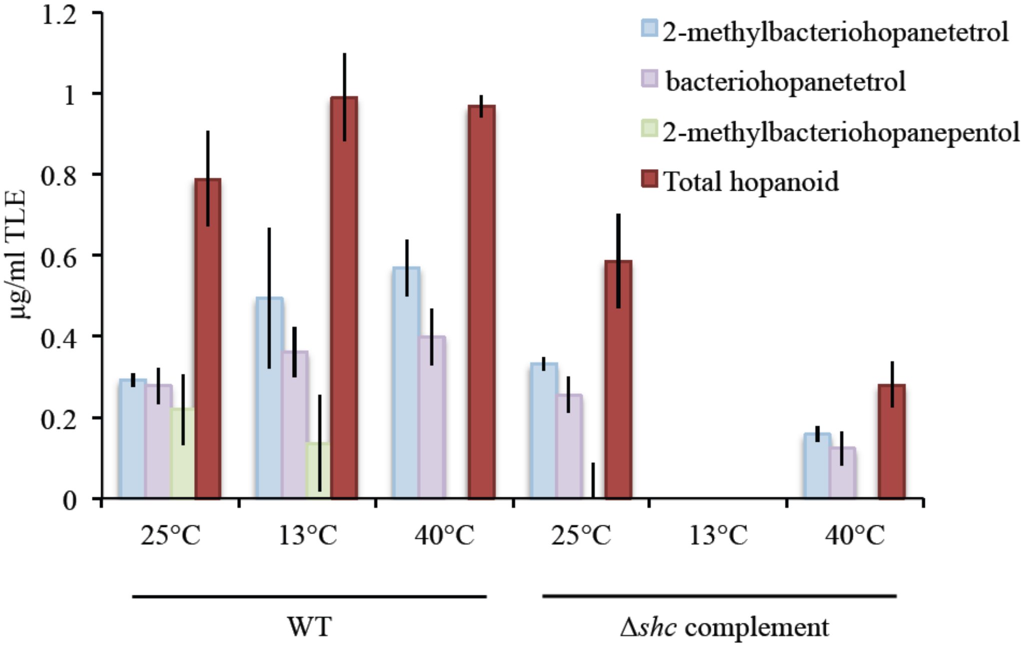


Figure 4.5 | Hopanoid abundance in *N. punctiforme* S WT and Δ*shc* complement across varying temperatures. Hopanoid concentrations vary little in WT over different temperatures. *N. punctiforme* Δ*shc* complement contains less hopanoids than WT. Measurements were made by GC-MS with biological triplicates.

We also measured changes in the greater lipid environment between *N. punctiforme* S WT and Δ*shc* at 13, 25, and 40°C. Principle component analysis generally separated lipidomes by temperature along PCA1 (29.3%), while PCA2 (21.3%) further separated the lipidomes by genetic background (Figure 4.6). While *N. punctiforme* S WT and Δ*shc* showed few changes at 25°C, as shown in Figure 4.3, the difference in the lipid composition of these strains at 13 and 40°C was greater (Figure 4.6). These data plus the unvarying abundances of hopanoids in WT at different temperatures imply that the greater lipid environment affects the function of hopanoids more than their total abundance under changing temperature. It is known that the composition of membranes changes with temperature to adjust fluidity (van Meer et al. 2008; Neubauer et al. 2015). Additionally, changes in the effect of hopanoids on different membrane compositions have been observed in vitro (Wu et al. 2015a).

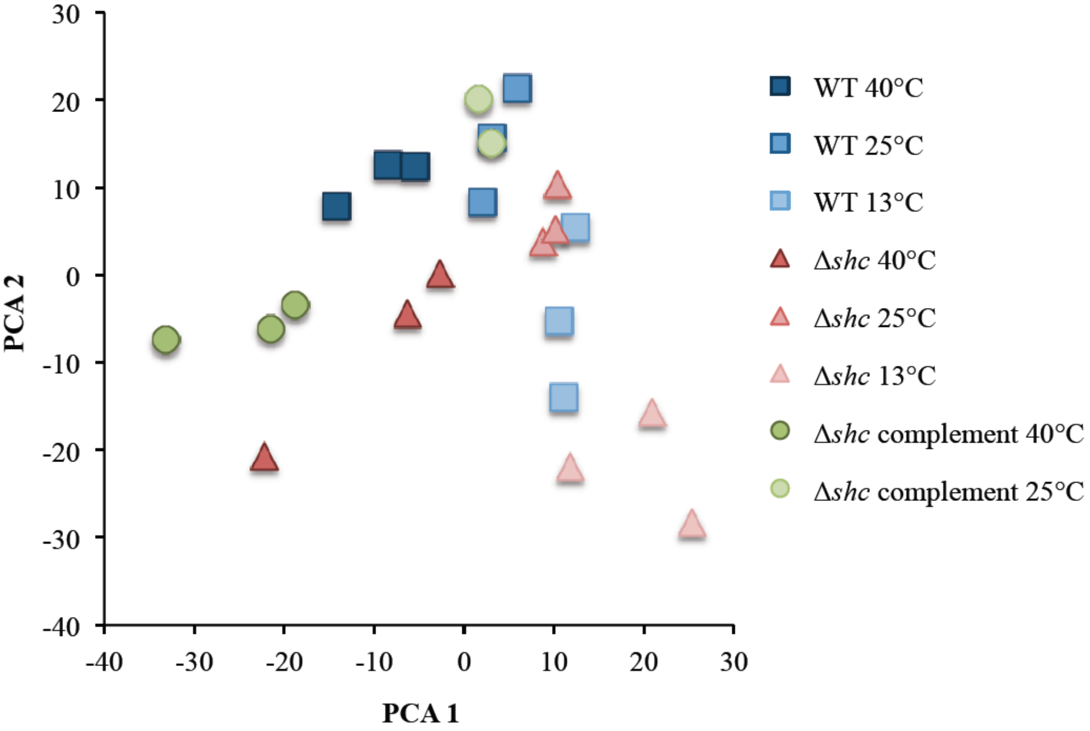


Figure 4.6 | Principle component analysis (PCA) of the lipidomes of *N. punctiforme* S WT, Δ*shc*, and Δ*shc* complement grown under varying temperatures. Measurements were made by LC-MS with biological duplicates or triplicates. Mass features of intensities greater than 1% relative to the maximum detected intensity were included in the PCA.

N. punctiforme S Δ*shc* complement was able to partially return the doubling time to WT levels at 40°C, but at 13°C the complementation strain attained a doubling time larger than WT. *N. punctiforme* S Δ*shc* complement produced less hopanoids compared to WT at all temperatures, including no hopanoids observed at 13°C. The composition of individual hopanoids at 40°C was similar between *N. punctiforme* S Δ*shc* complement and WT. On the other hand, at 25°C the reduction in total hopanoids abundance in *N. punctiforme* S Δ*shc* complement compared to WT appeared to be due mainly to decreased levels of 2-methylbacteriohopanepentol. Additionally, the lipidome of *N. punctiforme* S Δ*shc* complement at 25°C closely resembles WT at that temperature, and its lipidome at 40°C shares similarity with WT and Δ*shc*. These observations of *N. punctiforme* Δ*shc* complement's lipidomes and hopanoid composition are in line with the partial recovery of WT doubling time at 40°C. Yet, it is unclear why *N. punctiforme* S Δ*shc* complement was unable to restore WT function and hopanoid levels at cold temperature but was still able to recover moderate activity at high temperature. One plausible explanation is that overexpression of *shc* from a plasmid might lead to the build up of unknown intermediates or Shc itself, which might be toxic to the cell leading to a slower growth rate. Post-transcriptional regulation could result in the lack of production or degradation of hopanoids, which would explain their absence in *N. punctiforme* S Δ*shc* complement at low temperature. This could be confirmed by expressing *shc* from its native promoter on the chromosome to ensure native levels of expression.

N. punctiforme hopanoid mutants have no defect under various stresses

To understand other roles hopanoids and 2-methylhopanoids play in *N. punctiforme*. Hopanoid- and 2-methylhopanoid-lacking *N. punctiforme* S showed WT survival under ionic osmotic stress (NaCl), non-ionic osmotic stress (mannitol), envelope stress (EDTA), and desiccation (Figure 4.7). Under these stress conditions, WT displayed reduced growth in spot dilutions compared to ambient growth conditions, suggesting that physiologically relevant concentrations of stressors were used. These data imply that hopanoids and 2-methylhopanoids do not play a significant role protecting vegetative cells from ionic and non-ionic osmotic stress, outer destabilization, and desiccation.

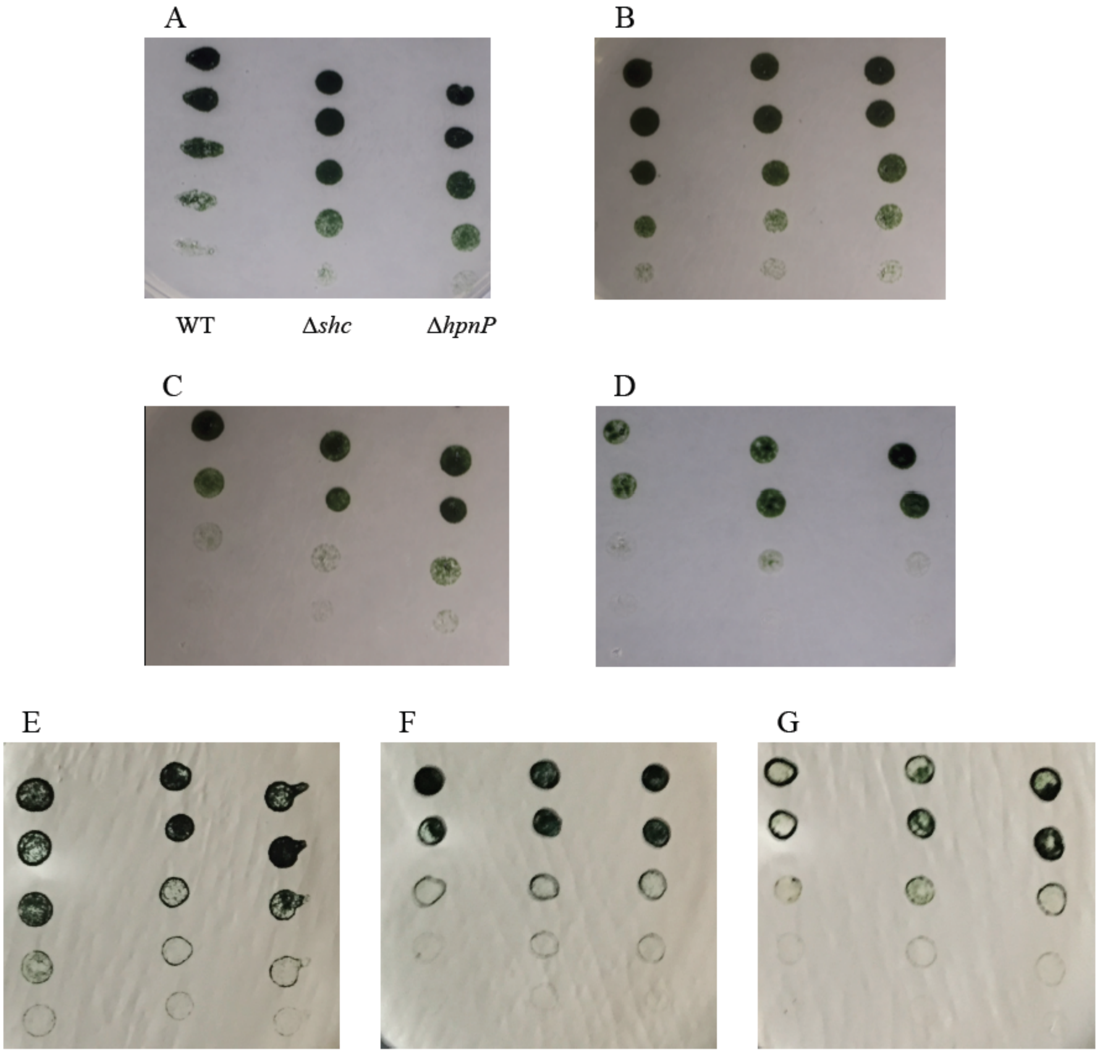


Figure 4.7 | Survival of *N. punctiforme* S hopanoid mutants under stress. (A) no stress (B) 0.45 mM EDTA (C) 300 mM NaCl (D) 300 mM mannitol (E) desiccation 0 hours (F) desiccation 3 hours (G) desiccation 8 hours. Spots were serially diluted 2 fold.

Deleterious effects of some of these stresses have been observed in hopanoid mutants of other organisms. For instance, *R. palustris* Δ*shc* displays growth defects in the presence of EDTA and non-ionic osmolytes (Welander et al. 2012; Kulkarni et al. 2013), and *Bradyrhizobium* sp. BTai1 Δ*shc* is sensitive to ionic osmolytes (Silipo et al. 2014). Additionally, *hpnP* is up-regulated in the presence of non-ionic osmolytes in *R. palustris* (Kulkarni et al. 2013). While these other organisms have growth defects under cell envelope and osmotic stress, *N. punctiforme* did not (Figure 4.7).

These results are contextualized by lipidomics data suggesting that *N. punctiforme* hopanoid mutants do not compensate for their missing hopanoids with other lipids to the extent that *R. palustris* does under standard growth conditions (Figure 4.3 and 4.6). Additionally, *N. punctiforme* membranes generally contain less hopanoids than *R. palustris*, except akinetes (Doughty et al. 2009; Welander et al. 2009). It is likely that hopanoids do not play an important role in outer membrane stabilization and osmotic stress in *N. punctiforme* under the vegetative conditions tested as opposed to other species, but it is possible that hopanoids are needed in other cell types, such as akinetes, under stressful conditions. Further work with other stressors and cell types is needed to confirm these hypotheses.

Hopanoid-lacking N. punctiforme can form symbioses with A. punctatus

We tested the symbiotic efficiency of hopanoid-lacking *N. punctiforme* ATCC 29133 in *A. punctatus*. *N. punctiforme* ATCC 29133 Δ*shc* was able to infect *A. punctatus* similar to WT (Figure 4.8). Two weeks post infection, *A. punctatus* with either symbiont was not chlorotic, and the number and size of *N. punctiforme* colonies were similar. The health of *A. punctatus* infected with either strain was tracked for a total of 8 weeks with no defects observed. Our qualitative assessment of the symbiosis showed no defect in plant health or the appearance of *N. punctiforme* symbiotic colonies (Figure 4.8), although a more subtle deficiency could be present. This result contrasts with *Bradyrhizobium* sp. BTai1 Δ*shc*, which forms ineffective symbioses and stunts the growth of the legume *Aeschynomene evenia* (Silipo et al. 2014). It is possible that hopanoids are important in *N. punctiforme*'s interactions with other plants; these interactions span a large phylogenetic diversity, involve the infection of divergent plant structures, and have different physiological effects on *N. punctiforme* as evidenced by the diversity in the percentage of heterocysts in each interaction (Meeks & Elhai 2002).

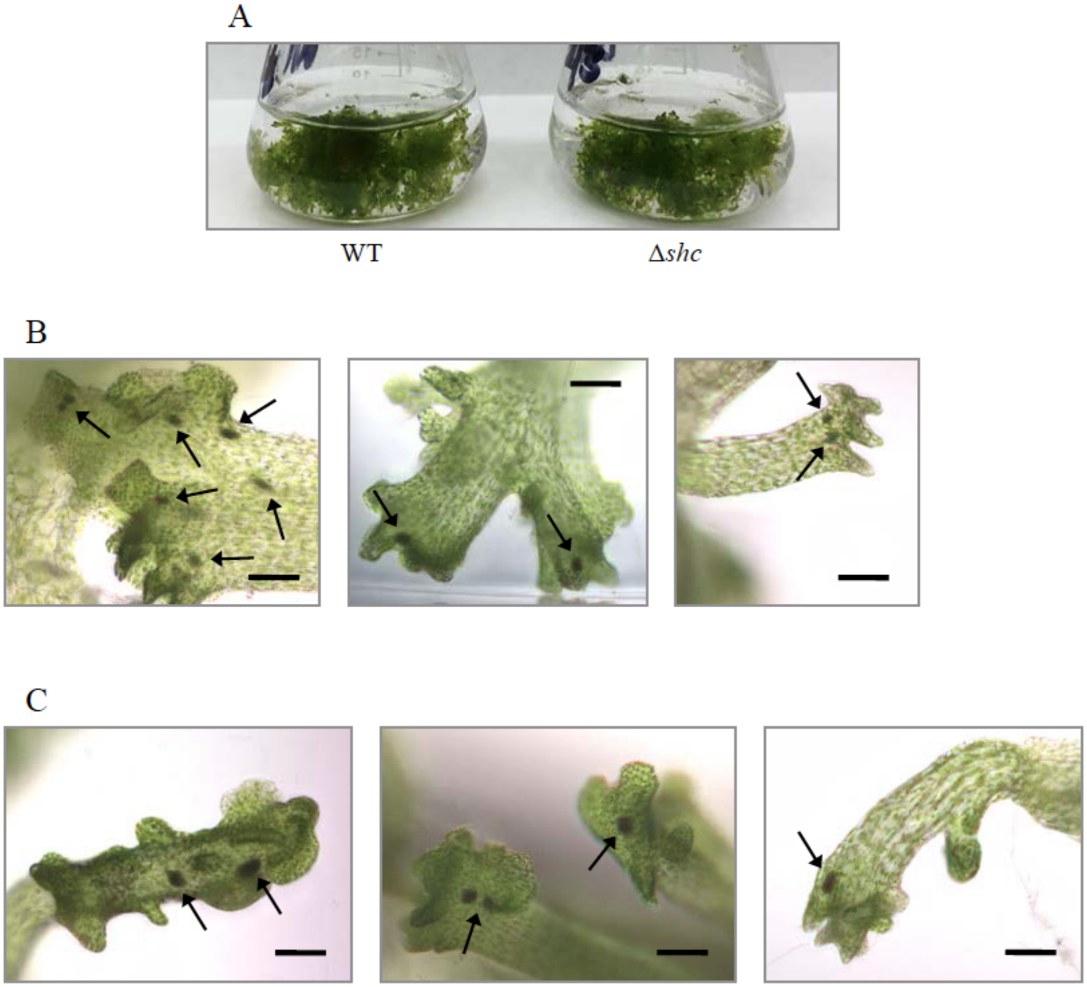


Figure 4.8 | Symbiosis effectiveness with a *N. punctiforme* ATCC 29133 mutant lacking hopanoids. (A) *A. punctatus* infected with *N. punctiforme* ATCC 29133 WT and Δ*shc*. *N. punctiforme* (B) WT and (C) Δ*shc* colonies in *A. punctatus* tissue. *N. punctiforme* symbiotic colonies are denoted with arrows (Meeks & Elhai, 2002). Scale bars are 200 μm.

Conclusion

In conclusion, we learned that hopanoids are important to *N. punctiforme* at extreme temperatures and likely assist in rigidifying the membrane under these conditions. Changes in the general lipid environment at high and low temperatures seem to affect the importance of hopanoids more so than changes in hopanoid abundance. If these observations about hopanoids being important under high temperatures are true for all Cyanobacteria, they may explain the presence of many hopanoid-producing Cyanobacteria in high temperature environments such as hot springs (Summons et al. 1999).

However, hopanoids do not appear to play a significant role in coping with cell envelope or osmotic stressors in vegetative cells, nor are they important for *N. punctiforme* to interact with one of its plant partners, *A. punctatus*. While our experiments presented in this study focus on vegetative cells, future work will concentrate on other cell types, particularly akinetes that have elevated levels of hopanoids. Nevertheless, this study used a genetic approach to provide insight into the biological function of hopanoids in Cyanobacteria and suggests that these molecules may play overlapping as well as distinct roles in different taxa. Additional work will be needed in Cyanobacteria and other organisms to further explore the diversity of hopanoid functions.

Acknowledgements

This chapter was a collaborative effort with Michael Summers (California State University, Northridge). We thank Caj Neubauer, Chia-Hung Wu, and Gargi Kulkarni for helpful discussions and assistance with LC-MS and GC-MS. We are grateful to Jack Meeks and Elsie Campbell for their help with *N. punctiforme* mutant construction and *A. punctatus*.

References

Allen M, Arnon D (1955) Studies on nitrogen-fixing blue-green algae. I. Growth and nitrogen fixation by *Anabaena cylindrica* Lemm. *Plant Physiology*, **30**, 366–372.

Argueta C, Summers ML (2005) Characterization of a model system for the study of *Nostoc punctiforme* akinetes. *Archives of Microbiology*, **183**, 338–46.

Brocks JJ, Love GD, Summons RE, Knoll AH, Logan GA, Bowden SA (2005) Biomarker evidence for green and purple sulphur bacteria in a stratified Palaeoproterozoic sea. *Nature*, **437**, 866–870.

Cai YP, Wolk CP (1990) Use of a conditionally lethal gene in *Anabaena* sp. strain PCC 7120 to select for double recombinants and to entrap insertion sequences. *Journal of Bacteriology*, **172**, 3138–45.

Campbell EL, Meeks JC (1989) Characteristics of hormogonia formation by symbiotic *Nostoc* spp. in response to the presence of *Anthoceros punctatus* or its extracellular products. *Applied and Environmental Microbiology*, **55**, 125–131.

Cohen MF, Meeks JC (1997) A Hormogonium Regulating Locus, hrmUA , of the Cyanobacterium *Nostoc punctiforme* Strain ATCC 29133 and its Response to an Extract of a Symbiotic Plant Partner *Anthoceros punctatus*. *Molecular Plant-Microbe Interactions*, **10**, 280–289.

Doughty DM, Hunter R, Summons RE, Newman DK (2009) 2-Methylhopanoids are maximally produced in akinetes of *Nostoc punctiforme*: geobiological implications. *Geobiology*, **7**, 524–532.

Enderlin CS, Meeks JC (1983) Pure culture and reconstitution of the *Anthoceros-Nostoc* symbiotic association. *Planta*, **158**, 157–165.

Englund E, Pattanaik B, Ubhayasekera SJK, Stensjö K, Bergquist J, Lindberg P (2014) Production of squalene in *Synechocystis* sp. PCC 6803. *PloS One*, **9**, e90270.

Flores E, Herrero A (2010) Compartmentalized function through cell differentiation in filamentous cyanobacteria. *Nature Reviews Microbiology*, **8**, 39–50.

Grant SGN, Jessee J, Bloom FR, Hanahan D (1990) Differential plasmid rescue from transgenic mouse DNAs into *Escherichia coli* methylation-restriction mutants. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 4645–4649.

Isaac G, McDonald S, Astarita G (2011) Lipid Separation using UPLC with Charged Surface Hybrid Technology. *Waters App note*, 720004107en.

Kulkarni G, Wu C-H, Newman DK (2013) The general stress response factor EcfG regulates expression of the C-2 hopanoid methylase HpnP in *Rhodopseudomonas palustris* TIE-1. *Journal of Bacteriology*, **195**, 2490–2498.

Meeks JC (2009) Physiological Adaptations in Nitrogen-fixing *Nostoc*–Plant Symbiotic Associations. In: *Microbiology Monographs: Prokaryotic Symbionts in Plants* (ed Pawlowski K), pp. 181–205. Springer Berlin Heidelberg, Berlin, Heidelberg.

Meeks JC, Campbell EL, Summers ML, Wong FC (2002) Cellular differentiation in the cyanobacterium *Nostoc punctiforme*. *Archives of Microbiology*, **178**, 395–403.

Meeks JC, Elhai J (2002) Regulation of Cellular Differentiation in Filamentous Cyanobacteria in Free-Living and Plant-Associated Symbiotic Growth States Regulation of Cellular Differentiation in Filamentous Cyanobacteria in Free-Living and Plant-Associated Symbiotic Growth States. *Microbiology and Molecular Biology Review*, **66**, 94–121.

Meeks JC, Wycoff K, Chapman J, Enderlin CS (1983) Regulation of expression of nitrate and dinitrogen assimilation by *Anabaena* species. *Applied and Environmental Mircobiology*, **45**, 1351–1359.

Neubauer C, Dalleska NF, Cowley ES, Shikuma NJ, Wu C-H, Sessions AL, Newman DK (2015) Loss of hopanoid methylation leads to lipid remodeling and differential subcellular localization in *Rhodopseudomonas palustris* TIE-1. *Geobiology*, accepted.

R Core Team (2014) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria*, URL http://www.R–project.org/No Title.

Rashby SE, Sessions AL, Summons RE, Newman DK (2007) Biosynthesis of 2-methylbacteriohopanepolyols by an anoxygenic phototroph. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 15099–15104.

Ricci JN, Coleman ML, Welander P V, Sessions AL, Summons RE, Spear JR, Newman DK (2014) Diverse capacity for 2-methylhopanoid production correlates with a specific ecological niche. *The ISME Journal*, **8**, 675–684.

Ricci JN, Michel AJ, Newman DK (2015) Phylogenetic analysis of HpnP reveals the origin of 2-methylhopanoid production in Alphaproteobacteria. *Geobiology*, **13**, 267-277.

Rippka R, Herdman M (1992) In Pasteur Culture Collection of Cyanobacteria in Axenic Culture. *Paris: Institut Pasteur*, 44–57.

Sáenz JP, Sezgin E, Schwille P, Simons K (2012) Functional convergence of hopanoids and sterols in membrane ordering. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 14236–14240.

Schmerk CL, Bernards MA, Valvano MA (2011) Hopanoid production is required for low-pH tolerance, antimicrobial resistance, and motility in *Burkholderia cenocepacia*. *Journal of Bacteriology*, **193**, 6712–6723.

Schmerk CL, Welander P V., Hamad MA, Bain KL, Bernards MA, Summons RE, Valvano MA (2015) Elucidation of the *Burkholderia cenocepacia* hopanoid biosynthesis pathway uncovers functions for conserved proteins in hopanoid-producing bacteria. *Environmental Microbiology*, **17**, 735–750.

Silipo A, Vitiello G, Gully D, Sturiale L, Chaintreuil C, Fardoux J, Gargani D, Lee H-I, Kulkarni G, Busset N, Marchetti R, Palmigiano A, Moll H, Engel R, Lanzetta R, Paduano L, Parrilli

M, Chang W-S, Holst O, Newman DK, Garozzo D, D'Errico G, Giraud E, Molinaro A (2014) Covalently linked hopanoid-lipid A improves outer-membrane resistance of a *Bradyrhizobium* symbiont of legumes. *Nature Communications*, **5**, 5106.

Smith C, Want E, O'Maille G, Abagyan R, Siuzdak G (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, **778**, 779–787.

Summers ML, Wallis JG, Campbell EL, Meeks JC (1995) Genetic evidence of a major role for glucose-6-phosphate dehydrogenase in nitrogen fixation and dark growth of the cyanobacterium *Nostoc* sp . strain ATCC 29133. *Journal of Bacteriology*, **177**, 6184–6194.

Summons RE, Jahnke LL, Hope JM, Logan GA (1999) 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature*, **400**, 554–557.

Talbot HM, Summons RE, Jahnke L, Cockell C, Rohmer M, Farrimond P (2008) Cyanobacterial bacteriohopanepolyol signatures from cultures and natural environmental settings. *Organic Geochemistry*, **39**, 232–263.

van Meer G, Voelker DR, Feigenson GW (2008) Membrane lipids: where they are and how they behave. *Nature Reviews: Molecular Cell Biology*, **9**, 112-124.

Welander P V, Coleman ML, Sessions AL, Summons RE, Newman DK (2010) Identification of a methylase required for 2-methylhopanoid production and implications for the interpretation of sedimentary hopanes. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 8537–8542.

Welander P V, Doughty DM, Wu C-H, Mehay S, Summons RE, Newman DK (2012) Identification and characterization of *Rhodopseudomonas palustris* TIE-1 hopanoid biosynthesis mutants. *Geobiology*, **10**, 163–177.

Welander P V, Hunter RC, Zhang L, Sessions AL, Summons RE, Newman DK (2009) Hopanoids play a role in membrane integrity and pH homeostasis in *Rhodopseudomonas palustris* TIE-1. *Journal of Bacteriology*, **191**, 6145–6156.

Wong FCY, Meeks JC (2002) Establishment of a functional symbiosis between the cyanobacterium *Nostoc punctiforme* and the bryophyte *Anthoceros punctatus* requires genes involved in nitrogen control and initiation of heterocyst differentiation. *Microbiology*, **148**, 315–323.

Wu C-H, Bialecka-Fornal M, Newman DK (2015a) Methylation at the C-2 position of hopanoids increases rigidity in native bacterial membranes. *eLife*, 10.7554/eL.

Wu C-H, Kong L, Bialecka-Fornal M, Park S, Thompson AL, Kulkarni G, Conway SJ, Newman DK (2015b) Quantitative hopanoid analysis enables robust pattern detection and comparison between laboratories. *Geobiology*, DOI: 10.1111/gbi.12132.

Chapter 5

CONCLUSIONS AND FUTURE DIRECTIONS

Conclusions

This thesis advances our understanding of 2-methylhopanoids, and hopanoids more broadly, in modern organisms and environments and refines their interpretation in the rock record. Prior to this body of work, the state of knowledge had discredited the hypothesis that 2-methylhopanoids were biomarkers of Cyanobacteria and their primary metabolism oxygenic photosynthesis, but it lacked a consensus for what these biomarkers were telling us about ancient life on Earth (Chapter 1). By studying *hpnP*, the C-2 hopanoid methylase gene, in various environments, we found that the capacity for 2-methylhopanoid production was correlated to plant-microbe interactions, opening up a new area of research (Chapter 2). In parallel, the evolutionary history of HpnP revealed that 2-methylhopanoids originated in the Alphaproteobacteria, suggesting that these biomarkers evolved after the rise of oxygen in the atmosphere (Chapter 3). To expand the diversity of our model organisms and further explore the link between 2-methylhopanoids and plant symbioses, we created hopanoid biosynthetic mutants in the symbiotic cyanobacterium *N. punctiforme*. We found that a mutant unable to produce hopanoids was still capable of forming functional symbioses with *A. punctatus*. Yet further experiments revealed that hopanoid-lacking *N. punctiforme* displays phenotypes at extreme temperature consistent with hopanoids rigidifying the cell membrane (Chapter 4). Together these works reveal the niches inhabited by *hpnP*-containing organisms, establish the evolutionary history of 2-methylhopanoids, and expand our knowledge of the biological function of hopanoids in Cyanobacteria.

Based on the findings presented in this thesis, it is tempting to speculate about the function of 2-methylhopanoids and the interpretation of their fossils in the rock record. While it is possible that 2-

methylhopanoids have served multiple functions over time in different taxa, the work presented here allows us to contemplate the first biological role of 2-methylhopanoids that would have likely been relevant in depositional settings immediately after the origin of these molecules. Given our current knowledge of the HpnP phylogeny, 2-methylhopanoids arose in a subset of the Rhizobiales, an order of Alphaproteobacteria. We can assume that the original function of 2-methylhopanoids was related to a feature present in these ancient organisms and is likely a common characteristic of modern organisms in this group. Two metabolisms often present in extant *hpnP*-containing Alphaproteobacteria are anoxygenic photosynthesis and nitrogen fixation, both of which are consistent with the low oxygen and low fixed nitrogen *hpnP* niche hypothesized in Chapter 2. The 2-methylhopane record contains notable links between these metabolisms. The oldest known 2-methylhopanes were deposited in the stratified marine basin of the Barney Creek Formation, an environment where anoxygenic phototrophs could have thrived, and carotenoid biomarkers indicative of anoxygenic phototrophs have been recovered there. Additionally, nitrogen fixation is thought to be important during ocean anoxic events, which cause disruptions in the nitrogen cycle. A functional role between nitrogen fixation and 2-methylhopanoids could explain increases in the 2-methylhopane index during these events. Furthermore, anoxygenic photosynthesis and nitrogen fixation share a common physiological link: they both occur in the absence of oxygen. Considering that 2-methylhopanoids likely evolved after the rise of oxygen in the atmosphere and 2-methylhopanoids have been shown to rigidify membranes, perhaps 2-methylhopanoids served to protect these processes from oxygen by strengthening the cell's membrane. Further work will be needed to determine if 2-methylhopanoids play a role in either of these metabolisms.

Future Directions

Based on the findings presented in this thesis, some follow-up studies of interest are outlined below:

- In Chapter 1, we conducted a broad survey of *hpnP* in diverse environments. A relevant follow-up study could track the abundance of *hpnP* temporally or on a fine special scale in a habitat of interest, such as a microbial mat, marine setting, or plant-associated locale. Initially, environmental samples could be screened for *hpnP* with the PCR primers described in Chapter 1. Then primers for specific *hpnP* types could be designed for use in qPCR or q-rt-PCR to allow for abundance measurements. By delving into a particular environment in detail, new *hpnP* correlations could be found that would not have been clear with a large-scale survey.

- An additional study to follow up on the results from Chapter 1 would be to identify the organisms that the unknown environmental *hpnP* sequences originate from. Attempts were made to co-localize these environmental *hpnP* sequences with 16S rRNA from their organismal source by using microfluidic digital PCR but were unsuccessful, likely due to the low abundance of *hpnP* in the environment. An alternative experimental route could be to use FACS to enrich the organisms containing the target *hpnP* and then to sequence the genomic content of the sorted cells.

- As stated in Chapter 3, the HpnP phylogeny can be updated as new sequences and phylogenetic methods become available. New HpnP sequences may be found serendipitously in genomes and metagenomes or with targeted approaches such as metagenome walking and genome sequencing of enriched *hpnP*-containing cells. Additionally, phylogenetic techniques are likely to improve in the future, which may yield a more robust HpnP phylogeny.

- A related but distinct project from the HpnP phylogeny that appears in Chapter 3 is to reconstruct the evolutionary history of Shc. There appears to be at least one duplication of *shc* because some organisms have more than one copy. It would be interesting to determine the number of duplications that exist and when they occurred; then these events could be matched to distinct functions of the gene. A similar small-scale analysis has been done in *Bradyrhizobium* spp., but a wider study to include all *shc* sequences would be worthwhile.

- How is the synteny of the hopanoid gene cluster conserved across different species? This question can provide insight into how the hopanoid biosynthetic genes have been transferred between organisms and into the selective pressures that cluster genes within the genome. We have observed that there are well-conserved hopanoid gene clusters in the Alphaproteobacteria, while these genes are more disperse throughout the genomes of Cyanobacteria. Coupled to phylogenies, synteny data could identify horizontal transfer events of gene clusters and provide a clearer picture of hopanoid biosynthesis evolution than phylogenetics alone.

- In Chapter 4, we found no symbiotic defect when hopanoid-lacking *N. punctiforme* infected *A. punctatus*. It would be worthwhile to follow up this result in two ways: 1) examine hormogonia morphology and motility induced by *A. punctatus* to identify any subtle infection phenotypes and 2) test the symbiotic efficiency of the *N. punctiforme* hopanoid mutants in other plants such as *Blasia* or *Gunnera*.

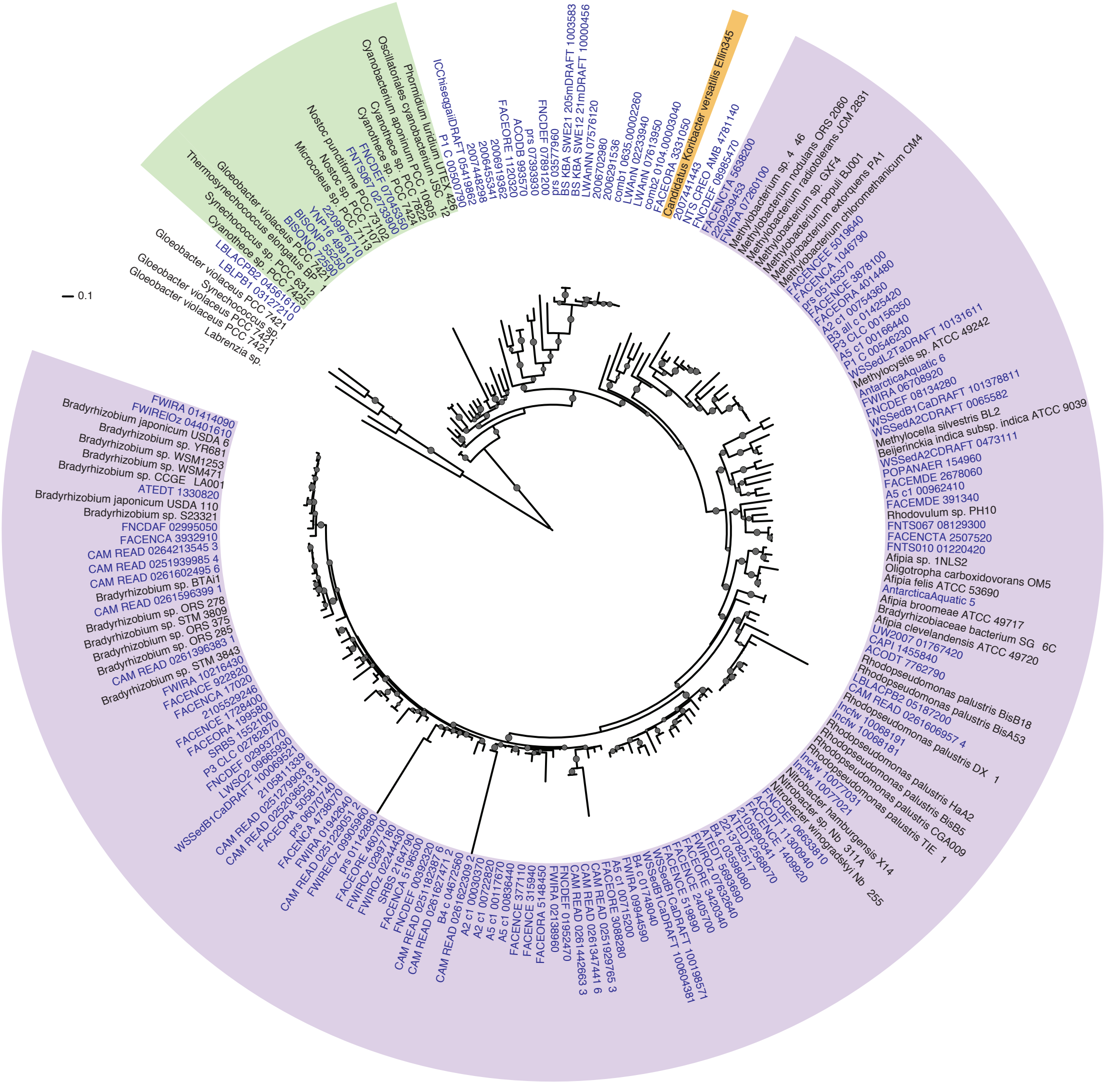
- We exposed the *N. punctiforme* hopanoid and 2-methylhopanoid mutants to multiple stress conditions when grown vegetatively (Chapter 4). There are some stressors that we did not test in *N. punctiforme* that have shown phenotypes in hopanoid mutants of other organisms, such as extreme pH and bile salts. Additionally, we only tested vegetative cells; it would be prudent to screen for phenotypes in other cell types, especially akinetes, which have high levels of hopanoids. Challenging *N. punctiforme* with these additional stresses and in other cell types will allow for us to gain a more complete picture of the role of hopanoids in *N. punctiforme*.

- We discovered changes in the lipidomes of *N. punctiforme* hopanoid mutants exposed to high and low temperature stresses (Chapter 4), but we did not identify the particular lipids that led to the changes in the lipidomes we observed. It would be worthwhile to determine the identity of the lipids that changed abundance between the mutants and WT because these lipids may play an important role in compensating for the lack of hopanoids.

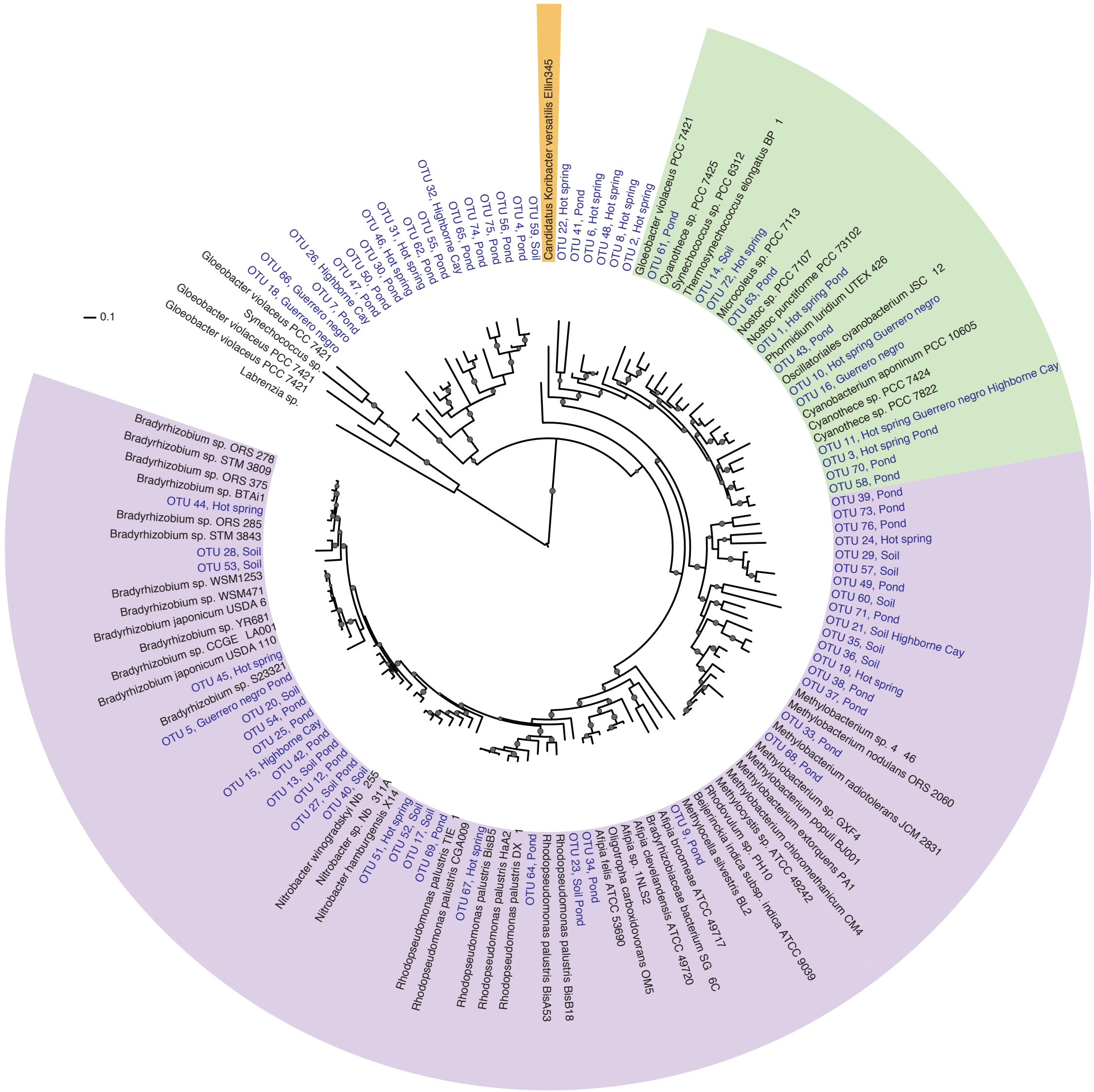
- *R. palustris* transports hopanoids to its outer membrane via HpnN and Rpal_4267, belonging to families 7 and 8 of the RND transporters. There are no known cyanobacterial homologs in either of these families, raising the question: how do Cyanobacteria transport their hopanoids to the outer membrane? *N. punctiforme* has an RND transporter homolog (locus F5285) located near other hopanoid biosynthetic genes. If this gene is responsible for hopanoid transport to the outer membrane, as could be shown genetically, it would be a hopanoid RND transporter that is evolutionarily distinct from those currently known.

Appendix A

DATASETS FROM CHAPTER II



Supplemental Figure 1 | Phylogeny of metagenomic HpnP sequences. Metagenomic HpnP sequences, blue, appear in this maximum likelihood phylogeny with known HpnP sequences from genomes. Phylogeny is the same as Figure 2.4. Additional information on metagenomic HpnP can be found in Supplemental Table 2. Clades are colored in purple for alphaproteobacteria, green for cyanobacteria, orange for acidobacteria, and colorless for unknown. Branch support was calculated with aLRT. Gray dots represent aLRT values over 0.8 with larger dot equaling larger aLRT values. The scale bar is a measure of evolutionary distance equaling 0.1 substitutions per site.


0.1
Labrenzia sp.
Gloeobacter violaceus PCC 7421
Gloeobacter violaceus PCC 7421
Synechococcus sp.
Gloeobacter violaceus PCC 7421
OTU 18, Guerrero negro
OTU 66, Guerrero negro
OTU 7, Pond
OTU 26, Highborne Cay
OTU 47, Pond
OTU 50, Pond
OTU 30, Pond
OTU 46, Hot spring
OTU 31, Hot spring
OTU 62, Pond
OTU 55, Pond
OTU 32, Highborne Cay
OTU 65, Pond
OTU 74, Pond
OTU 75, Pond
OTU 56, Pond
OTU 4, Pond
OTU 59, Soil
Candidatus Koribacter versatilis Ellin345
OTU 22, Hot spring
OTU 41, Pond
OTU 6, Hot spring
OTU 48, Hot spring
OTU 8, Hot spring
OTU 2, Hot spring
Gloeobacter violaceus PCC 7421
OTU 61, Pond
Cyanothece sp. PCC 7425
Synechococcus sp. PCC 6312
Thermosynechococcus elongatus BP 1
OTU 14, Soil
OTU 72, Hot spring
Microcoleus sp. PCC 7113
OTU 63, Pond
Nostoc sp. PCC 7107
Nostoc punctiforme PCC 73102
OTU 1, Hot spring Pond
Phormidium luridium UTEX 426
OTU 43, Pond
Oscillatoriales cyanobacterium JSC 12
OTU 10, Hot spring Guerrero negro
OTU 16, Guerrero negro
Cyanobacterium aponinum PCC 10605
Cyanothece sp. PCC 7424
Cyanothece sp. PCC 7822
OTU 11, Hot spring Guerrero negro Highborne Cay
OTU 3, Hot spring Pond
OTU 70, Pond
OTU 58, Pond
OTU 39, Pond
OTU 73, Pond
OTU 76, Pond
OTU 24, Hot spring
OTU 29, Soil
OTU 57, Soil
OTU 49, Pond
OTU 60, Soil
OTU 71, Pond
OTU 21, Soil Highborne Cay
OTU 35, Soil
OTU 36, Soil
OTU 19, Hot spring
OTU 38, Pond
OTU 37, Pond
Methylobacterium sp. 4 46
Methylobacterium nodulans ORS 2060
OTU 33, Pond
Methylobacterium radiotolerans JCM 2831
OTU 68, Pond
Methylobacterium sp. GXF4
Methylobacterium populi BJ001
Methylobacterium extorquens PA1
Methylobacterium chloromethanicum CM4
Methylocystis sp. ATCC 49242
Rhodovulum sp. PH10
Beijerinckia indica subsp. indica ATCC 9039
Methylocella silvestris BL2
OTU 9, Pond
Afipia broomeae ATCC 49717
Bradyrhizobiaceae bacterium SG 6C
Afipia clevelandensis ATCC 49720
Afipia sp. 1NLS2
Oligotropha carboxidovorans OM5
Afipia felis ATCC 53690
OTU 34, Pond
OTU 23, Soil Pond
Rhodopseudomonas palustris BisB18
Rhodopseudomonas palustris BisA53
OTU 64, Pond
Rhodopseudomonas palustris DX 1
Rhodopseudomonas palustris HaA2
Rhodopseudomonas palustris BisB5
OTU 67, Hot spring
Rhodopseudomonas palustris CGA009
Rhodopseudomonas palustris TIE 1
OTU 69, Pond
OTU 17, Soil
OTU 52, Soil
OTU 51, Hot spring
Nitrobacter hamburgensis X14
Nitrobacter sp. Nb 311A
Nitrobacter winogradskyi Nb 255
OTU 40, Soil
OTU 27, Soil Pond
OTU 12, Pond
OTU 13, Soil Pond
OTU 42, Pond
OTU 15, Highborne Cay
OTU 25, Pond
OTU 54, Pond
OTU 20, Soil
OTU 5, Guerrero negro Pond
Bradyrhizobium sp. S23321
OTU 45, Hot spring
Bradyrhizobium japonicum USDA 110
Bradyrhizobium sp. CCGE LA001
Bradyrhizobium sp. YR681
Bradyrhizobium japonicum USDA 6
Bradyrhizobium sp. WSM471
Bradyrhizobium sp. WSM1253
OTU 53, Soil
OTU 28, Soil
Bradyrhizobium sp. STM 3843
Bradyrhizobium sp. ORS 285
OTU 44, Hot spring
Bradyrhizobium sp. BTAi1
Bradyrhizobium sp. ORS 375
Bradyrhizobium sp. STM 3809
Bradyrhizobium sp. ORS 278

Supplemental Figure 2 | Phylogenetic tree comparing PCR clone library generated HpnP sequences to reference HpnP sequences from genomes. Environmental *hpnP* sequences were clustered at 95% similarity in CD-HIT. The representative sequences recovered were aligned in MAFFT. The phylogeny was generated in PhyML and edited in iTOL. Leaf names reflect all sequences that clustered with the representative sequence in blue. Clades are colored in purple for alphaproteobacteria, green for cyanobacteria, and orange for acidobacteria. Unknown group 1 is the early diverging uncolored region, and unknown group 2 is the uncolored region between the acidobacterium and cyanobacteria. Branch supports were calculated with aLRT. Gray dots represent aLRT values over 0.8 with larger dot equaling larger aLRT values. The scale bar is a measure of evolutionary distance equaling 0.1 substitutions per site.

Supplemental Table 1 | description of environmental samples that appear in Figure 2.5

Environment	Site/Sample type	Sample Name	Description	Previously Published Name
Hot springs	Carbonate	Narrow Gauge 1	Carb; floating pieces of mat from terrace pool	NG09-1
Hot springs	Carbonate	Narrow Gauge 2	Carb; thin orange mat from terrace edge	NG09-2
Hot springs	Carbonate	Narrow Gauge 3	Carb; thin orange mat from terrace overflow	NG09-3
Hot springs	Carbonate	Narrow Gauge 5	Carb; orange mat from runnoff channel	NG09-5
Hot springs	Carbonate	Narrow Gauge 6	Carb; white carbonate from spring source	NG09-6
Hot springs	Carbonate	Narrow Gauge 7	Carb; thin white streamers in source pool	NG09-7
Hot springs	Carbonate	Narrow Gauge 8	Carb; orange mat from runnoff channel	NG09-8
Hot springs	Pink streamer	Octopus Spring 1	PS; pink streamers	OS09-1
Hot springs	Pink streamer	Brain Pool 1	PS; pink streamers	BP09-1
Hot springs	Pink streamer	Bison Pool 1	PS; pink streamers	B09-1
Hot springs	Pink streamer	Ojo Caliente 1	PS; pink streamers	OC09-1
Hot springs	Pink streamer	Brain Pool 2	PS; pink streamers with green coatings forming in mixing zone between geyser and creek	BP09-2
Hot springs	Pink streamer	Narrow Gauge 4	PS; white opaque streamers from source old Narrow Gauge (carbonate system)	NG09-4
Hot springs	Pink streamer	Spent Kleenex 1	PS; white streamers	SK09-1
Hot springs	Pink streamer	Bison Pool 2	PS; yellow streamers	B09-2
Hot springs	Yellow biofilm	Bison Pool 3	YB; yellow biofilm	B09-3
Hot springs	Yellow biofilm	Octopus Spring 2	YB; yellow biofilm	OS09-2
Hot springs	Orange mat high temperature	Octopus Spring 3	OM-HT; thick green filaments growing off of stratified orange mat	OS09-3
Hot springs	Orange mat high temperature	Imperial Geyser 1	OM-HT; yellow topped stratified mat with orange and green layers below	IG09-1
Hot springs	Orange mat high temperature	Imperial Geyser 2a	OM-HT; <1mm think upper orange and green layer of stratified mat	IG09-2a
Hot springs	Orange mat high temperature	Imperial Geyser 2b	OM-HT; bright salmon color layers in stratified mat	IG09-2b
Hot springs	Orange mat high temperature	Imperial Geyser 2c	OM-HT; thick grey soft mat	IG09-2c
Hot springs	Orange mat high temperature	Imperial Geyser 3	OM-HT; thick green filaments growing off of stratified orange mat	IG09-3
Hot springs	Orange mat low temperature	Bison Pool 4a	OM-LT; upper orange layer from conoform orange mat	B09-4a
Hot springs	Orange mat low temperature	Bison Pool 4b	OM-LT; green middle layer from conoform orange mat	B09-4b
Hot springs	Orange mat low temperature	Bison Pool 4c	OM-LT; salmon colored sinter pieces at base of conoform mat	B09-4c
Hot springs	Orange mat low temperature	Octopus Spring 4a	OM-LT; upper orange layer from conoform orange mat	OS09-4a
Hot springs	Orange mat low temperature	Octopus Spring 4b	OM-LT; green middle layer from conoform orange mat	OS09-4b
Hot springs	Orange mat low temperature	Octopus Spring 4c	OM-LT; salmon colored sinter pieces at base of conoform mat	OS09-4c
Hot springs	Black sediment	Norris zygogonium mat 1	BS; zygogonium mat	NR09-1
Hot springs	Black sediment	Norris zygogonium mat 2	BS; elemental sulfur colored white mat from zygogoinum mat source spring	NR09-2
Hot springs	Black sediment	Norris zygogonium mat 3	BS; bright green biofilm from runoff channel	NR09-3
Hot springs	Black sediment	Washburn 1	BS; black gelatinous sediment	WB09-1
Hot springs	Black sediment	Washburn 2	BS; black sediment	WB09-2
Hot springs	Black sediment	Boulder Spring 1	BS; black gelatinous sediment	BS09-1
Hot springs	Black sediment	Boulder Spring 2	BS; black gelatinous sediment	BS09-2

Environment	Site/Sample type	Sample Name	Description	Previously Published Name
Terrestrial	Soil	Rhizosphere, control	Rancho Sierra Vista Newbury Park, CA ; Sage brush rhizosphere; 0 g N/m/yr	
Terrestrial	Soil	Rhizosphere, medium nitrogen	Rancho Sierra Vista Newbury Park, CA ; Sage brush rhizosphere; 1.5 g N/m/yr	
Terrestrial	Soil	Rhizosphere, high nitrogen	Rancho Sierra Vista Newbury Park, CA ; Sage brush rhizosphere; 3 g N/m/yr	
Freshwater	Man-made ponds	Baxter water, summer	Whole water	
Freshwater	Man-made ponds	Baxter leafs, summer	Lilypad leaf scrapping	
Freshwater	Man-made ponds	Baxter plant debris, summer	Decaying plant matter from the bottom of the pond	
Freshwater	Man-made ponds	Baxter water, spring	Whole water	
Freshwater	Man-made ponds	Baxter leafs, spring	Lilypad leaf scrapping	
Freshwater	Man-made ponds	Baxter plant debris, spring	Decaying plant matter from the bottom of the pond	
Freshwater	Man-made ponds	Beckman water, summer	Whole water	
Freshwater	Man-made ponds	Beckman leafs, summer	Lilypad leaf scrapping	
Freshwater	Man-made ponds	Beckman plant debris, summer	Decaying plant matter from the bottom of the pond	
Freshwater	Man-made ponds	Turtle pond water, summer	Whole water	
Freshwater	Man-made ponds	Turtle pond leafs, summer	Whole leaf	
Freshwater	Man-made ponds	Turtle pond sediment, summer	Pond sediment in contact with plant roots	
Marine	Open water	Whole seawater	SPOT 121; 33o33'N, 118o24'W; Sept. 21, 2012; 2L surface water filter on GF/F Whatman	
Marine	Open water	Plankton tow	SPOT 121; 33o33'N, 118o24'W; Sept. 21, 2012; horizontal surface tow filtered on GF/F Whatman; contained pennate diatoms, e.g., Rhizosolenia	
Marine	Hypersaline lagoon	Laguna Guerrero Negro 1	In situ Pond 4 near 1 (0-5mm); September 16, 2002	
Marine	Hypersaline lagoon	Guerrero negro 2	Flume experiment Pond 4 near 5 85 ppt salinity top 0 mm, orange	
Marine	Hypersaline lagoon	Guerrero negro 3	Flume experiment Pond 4 near 5 85 ppt salinity top 1-1.5 mm	
Marine	Hypersaline lagoon	Guerrero negro 4	Flume experiment Pond 4 near 5 diel 10:45 pm 0-1 mm	
Marine	Hypersaline lagoon	Guerrero negro 5	Flume experiment Pond 4 near 5 diel 2 pm 2-3 mm	
Marine	Hypersaline lagoon	Guerrero negro 6	Flume experiment Pond 4 near 5 diel 2 pm 0-1 mm	
Marine	Modern stromatolites	Highborne Cay 1	Subtidal mats, windward side of Highborne Cay Bahamas (76°49'W, 24°43'N), March 2010	
Marine	Modern stromatolites	Highborne Cay 2	As above	
Marine	Modern stromatolites	Highborne Cay 3	As above	

Supplemental Table 2 | List of *hpnP* sequences identified from metagenomes

Environment	Sample	Gene ID	Database
Ace Lake	AntarcticaAquatic_5 - MARINE DERIVED LAKE	724291281814423074caa13d3080ac8b_134536_836	CAMERA
Ace Lake	AntarcticaAquatic_6 - ACE LAKE, ANTARCTICA	ee13dbfa4b0778bcdead298de4342f33_223345_1049	CAMERA
Coastal	Baltic Sea site KBA sample SWE 12_21m (Baltic Sea site KBA sample SWE 12_21m, Oct 2011 Assem)	BS_KBA_SWE12_21mDRAFT_10000456	IMG
Coastal	Baltic Sea site KBB sample SWE 21_20.5m (Baltic Sea site KBB sample SWE 21_20.5m, Oct 2011 Assem)	BS_KBA_SWE21_205mDRAFT_1003583	IMG
Deep sea hydrothermal vent	Virome EPR hydrothermal vent: Extracellular RNA virome	CAM_READ_0251182387_6	CAMERA
Deep sea hydrothermal vent	Virome EPR hydrothermal vent: Extracellular RNA virome	CAM_READ_0251229051_2	CAMERA
Deep sea hydrothermal vent	Virome EPR hydrothermal vent: Extracellular RNA virome	CAM_READ_0251279903_6	CAMERA
Deep sea hydrothermal vent	Virome EPR hydrothermal vent: Induced RNA virome	CAM_READ_0251929765_3	CAMERA
Deep sea hydrothermal vent	Virome EPR hydrothermal vent: Induced RNA virome	CAM_READ_0251939985_4	CAMERA
Deep sea hydrothermal vent	Virome EPR hydrothermal vent: Induced ssDNA virome	CAM_READ_0252036513_3	CAMERA
Deep sea hydrothermal vent	Virome Guaymas hydrothermal vent: Induced ssDNA virome	CAM_READ_0261347441_6	CAMERA
Deep sea hydrothermal vent	Virome Guaymas hydrothermal vent: Induced ssDNA virome	CAM_READ_0261396383_1	CAMERA
Deep sea hydrothermal vent	Virome Guaymas hydrothermal vent: Induced ssDNA virome	CAM_READ_0261442663_3	CAMERA
Deep sea hydrothermal vent	Virome Guaymas hydrothermal vent: Induced ssDNA virome	CAM_READ_0261596399_1	CAMERA
Deep sea hydrothermal vent	Virome Guaymas hydrothermal vent: Induced ssDNA virome	CAM_READ_0261602495_6	CAMERA
Deep sea hydrothermal vent	Virome Guaymas hydrothermal vent: Induced ssDNA virome	CAM_READ_0261606957_4	CAMERA
Deep sea hydrothermal vent	Virome Guaymas hydrothermal vent: Induced ssDNA virome	CAM_READ_0261622309_2	CAMERA
Deep sea hydrothermal vent	Virome Guaymas hydrothermal vent: Induced ssDNA virome	CAM_READ_0261627471_2	CAMERA
Deep sea hydrothermal vent	Virome Guaymas hydrothermal vent: Induced RNA virome	CAM_READ_0264213545_3	CAMERA
Estuary	Soil microbial communities from sample at FACE Site 1 Maryland Estuary CO2+ (Maryland Estuary elevated)	FACEMDE_2678060	IMG
Estuary	Soil microbial communities from sample at FACE Site 1 Maryland Estuary CO2+ (Maryland Estuary elevated)	FACEMDE_391340	IMG
Estuary	Wetland microbial communities from Twitchell Island in the Sacramento Delta, sample from surface sediment Feb2011 Site A2 Cattail (Wetland Surface Sediment Feb2011 Site A2 Cattail Sept 2011 assem)	WSSedA2CDRAFT_0065582	IMG
Estuary	Wetland microbial communities from Twitchell Island in the Sacramento Delta, sample from surface sediment Feb2011 Site A2 Cattail (Wetland Surface Sediment Feb2011 Site A2 Cattail Sept 2011 assem)	WSSedA2CDRAFT_0473111	IMG
Estuary	Wetland microbial communities from Twitchell Island in the Sacramento Delta, sample from surface sediment Feb2011 Site B1 Cattail (Wetland Surface Sediment Feb2011 Site B1 Cattail, Assem Ctgs Sep 2011 assem)	WSSedB1CaDRAFT_100069521	IMG
Estuary	Wetland microbial communities from Twitchell Island in the Sacramento Delta, sample from surface sediment Feb2011 Site B1 Cattail (Wetland Surface Sediment Feb2011 Site B1 Cattail, Assem Ctgs Sep 2011 assem)	WSSedB1CaDRAFT_100198571	IMG
Estuary	Wetland microbial communities from Twitchell Island in the Sacramento Delta, sample from surface sediment Feb2011 Site B1 Cattail (Wetland Surface Sediment Feb2011 Site B1 Cattail, Assem Ctgs Sep 2011 assem)	WSSedB1CaDRAFT_100604381	IMG

Environment	Sample	Gene ID	Database
Estuary	Wetland microbial communities from Twitchell Island in the Sacramento Delta, sample from surface sediment Feb2011 Site B1 Cattail (Wetland Surface Sediment Feb2011 Site B1 Cattail, Assem Ctgs Sep 2011 assem)	WSSedB1CaDRAFT_101378811	IMG
Estuary	CECUM_4-1 (Microbiome Characterization)	WSSedL2TaDRAFT_10131611	IMG
Groundwater	Oak Ridge Pristine Groundwater FRC FW301	2007441443	IMG
Groundwater	Oak Ridge Pristine Groundwater FRC FW301	2007448298	IMG
Hot spring	5_050719P	BISONP_55250	IMG
Hot spring	4_050719Q	BISONQ_72590	IMG
Hot spring	Hot spring microbial community from Yellowstone Hot Springs, sample YNP16 from Fairy Spring Red Layer	YNP16_49910	IMG
Lake	Methylotrophic community from Lake Washington sediment Methanol enrichment	2006291536	IMG
Lake	Methylotrophic community from Lake Washington sediment Formaldehyde enrichment	2006455341	IMG
Lake	Methylotrophic community from Lake Washington sediment combined (v2)	2006702980	IMG
Lake	Methylotrophic community from Lake Washington sediment combined (v2)	2006919362	IMG
Lake	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, original sample replicate 1	comb1_0635.00002260	IMG
Lake	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, original sample replicate 1	comb2_0104.00003040	IMG
Lake	Lentic microbial communities from Lake Waban, Wellesley MA, that are anoxygenic and photosynthetic, photosynthetic consortia incandescent light (FW_incandescent_CN)	Incfw_10068181	IMG
Lake	Lentic microbial communities from Lake Waban, Wellesley MA, that are anoxygenic and photosynthetic, photosynthetic consortia incandescent light (FW_incandescent_CN)	Incfw_10068191	IMG
Lake	Lentic microbial communities from Lake Waban, Wellesley MA, that are anoxygenic and photosynthetic, photosynthetic consortia incandescent light (FW_incandescent_CN)	Incfw_10077021	IMG
Lake	Lentic microbial communities from Lake Waban, Wellesley MA, that are anoxygenic and photosynthetic, photosynthetic consortia incandescent light (FW_incandescent_CN)	Incfw_10077031	IMG
Lake	Fresh water microbial communities from LaBonte Lake, Laramie, Wyoming, sample from algal/cyanobacterial bloom material peak-bloom 2 (algal/cyano bloom peak-bloom 2)	LBLACPB2_04561610	IMG
Lake	Fresh water microbial communities from LaBonte Lake, Laramie, Wyoming, sample from peak-bloom 1 (Peak bloom metagenome 1)	LBLACPB2_05187200	IMG
Lake	Fresh water microbial communities from LaBonte Lake, Laramie, Wyoming, sample from peak-bloom 1 (Peak bloom metagenome 1)	LBLPB1_03127210	IMG
Lake	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, sample SIP 13C-methane anaerobic+nitrate (Anaerobic + nitrate SIP Nov 2010 with PE)	LWAnN_02233940	IMG
Lake	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, sample SIP 13C-methane anaerobic+nitrate (Anaerobic + nitrate SIP Nov 2010 with PE)	LWAnN_07613950	IMG
Lake	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, sample SIP 13Cmethane anaerobic no nitrate (Anaerobic no nitrate SIP Nov 2010 with PE)	LWAnNN_07576120	IMG
Lake	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, original sample replicate 2 (Original sample replicate 2 12C fraction)	LWSO2_09665930	IMG
Leaf-cutting ant waste dump	Fungus garden microbial communities from Atta colombica in Panama, sample from dump bottom (Dump bottom)	ACODB_993570	IMG
Leaf-cutting ant waste dump	Fungus garden microbial communities from Atta colombica in Panama, sample from dump top (Dump top)	ACODT_11300940	IMG

Environment	Sample	Gene ID	Database
Leaf-cutting ant waste dump	Fungus garden microbial communities from Atta colombica in Panama, sample from dump top (Dump top)	ACODT_7762790	IMG
Leaf-cutting ant waste dump	Dump top (Dump top)	ATEDT_1330820	IMG
Leaf-cutting ant waste dump	Dump top (Dump top)	ATEDT_2568070	IMG
Leaf-cutting ant waste dump	Dump top (Dump top)	ATEDT_5693690	IMG
Soil	Arabidopsis rhizosphere microbial communities from University of North Carolina, sample Mutant cpr5 (cpr5 454/Illumina combined assembly)	2105529246	IMG
Soil	Arabidopsis rhizosphere microbial communities from University of North Carolina, sample Mutant cpr5 (cpr5 454/Illumina combined assembly)	2105690341	IMG
Soil	Arabidopsis rhizosphere microbial communities from University of North Carolina, sample Mutant cpr5 (cpr5 454/Illumina combined assembly)	2105811339	IMG
Soil	Soil microbial communities sample from Dark Crust, Colorado Plateau, Green Butte (Dark Crust, Colorado Plateau, Green Butte June 2011 assem)	2209239453	IMG
Soil	Soil microbial communities sample from Dark Crust, Colorado Plateau, Green Butte (Dark Crust, Colorado Plateau, Green Butte June 2011 assem)	2209976710	IMG
Soil	Arabidopsis rhizosphere microbial communities from University of North Carolina, sample Wild type Col-0 (Arabidopsis rhizosphere microbiome-wild type Col-0 454/Illumina 2011 July Assem)	2213782517	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Active Layer A2 (A2_CLC_pe)	A2_c1_00030370	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Active Layer A2 (A2_CLC_pe)	A2_c1_00722820	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Active Layer A2 (A2_CLC_pe)	A2_c1_00754360	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Active Layer A5 (A5_CLC_pe)	A5_c1_00117670	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Active Layer A5 (A5_CLC_pe)	A5_c1_00166440	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Active Layer A5 (A5_CLC_pe)	A5_c1_00715200	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Active Layer A5 (A5_CLC_pe)	A5_c1_00836440	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Active Layer A5 (A5_CLC_pe)	A5_c1_00962410	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Bog Site B3 (B3_all_CLC)	B3_all_c_01425420	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Bog Site B4 (B4_CLC)	B4_c_01748040	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Bog Site B4 (B4_CLC)	B4_c_03598080	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Bog Site B4 (B4_CLC)	B4_c_04672560	IMG
Soil	Soil microbial communities from sample at FACE Site 2 North Carolina CO2-	FACENCA_1046790	IMG
Soil	Soil microbial communities from sample at FACE Site 2 North Carolina CO2-	FACENCA_17020	IMG
Soil	Soil microbial communities from sample at FACE Site 2 North Carolina CO2-	FACENCA_3932910	IMG
Soil	Soil microbial communities from sample at FACE Site 2 North Carolina CO2-	FACENCA_4738070	IMG
Soil	Soil microbial communities from sample at FACE Site 2 North Carolina CO2-	FACENCA_5196500	IMG
Soil	Soil microbial communities from sample at FACE Site 2 North Carolina CO2+ (North Carolina Elevated CO2)	FACENCE_1409920	IMG
Soil	Soil microbial communities from sample at FACE Site 2 North Carolina CO2+ (North Carolina Elevated CO2)	FACENCE_1728400	IMG
Soil	Soil microbial communities from sample at FACE Site 2 North Carolina CO2+ (North Carolina Elevated CO2)	FACENCE_2405700	IMG

Environment	Sample	Gene ID	Database
Soil	Soil microbial communities from sample at FACE Site 2 North Carolina CO2+ (North Carolina Elevated CO2)	FACENCE_315940	IMG
Soil	Soil microbial communities from sample at FACE Site 2 North Carolina CO2+ (North Carolina Elevated CO2)	FACENCE_377110	IMG
Soil	Soil microbial communities from sample at FACE Site 2 North Carolina CO2+ (North Carolina Elevated CO2)	FACENCE_3878100	IMG
Soil	Soil microbial communities from sample at FACE Site 2 North Carolina CO2+ (North Carolina Elevated CO2)	FACENCE_519890	IMG
Soil	Soil microbial communities from sample at FACE Site 2 North Carolina CO2+ (North Carolina Elevated CO2)	FACENCE_922820	IMG
Soil	Soil microbial communities from sample at FACE Site 3 Nevada Test Site Creosote CO2+	FACENCEE_5019640	IMG
Soil	Soil microbial communities from sample at FACE Site 4 Nevada Test Site Crust CO2-	FACENCTA_2507520	IMG
Soil	Soil microbial communities from sample at FACE Site 4 Nevada Test Site Crust CO2-	FACENCTA_5638200	IMG
Soil	Soil microbial communities from sample at FACE Site 5 Oak Ridge CO2- (Oak Ridge ambient)	FACEORA_199580	IMG
Soil	Soil microbial communities from sample at FACE Site 5 Oak Ridge CO2- (Oak Ridge ambient)	FACEORA_3331050	IMG
Soil	Soil microbial communities from sample at FACE Site 5 Oak Ridge CO2- (Oak Ridge ambient)	FACEORA_4014480	IMG
Soil	Soil microbial communities from sample at FACE Site 5 Oak Ridge CO2- (Oak Ridge ambient)	FACEORA_5058110	IMG
Soil	Soil microbial communities from sample at FACE Site 5 Oak Ridge CO2- (Oak Ridge ambient)	FACEORA_5148450	IMG
Soil	Soil microbial communities from sample at FACE Site 5 Oak Ridge CO2+ (Oak Ridge elevated CO2)	FACEORE_1120320	IMG
Soil	Soil microbial communities from sample at FACE Site 5 Oak Ridge CO2+ (Oak Ridge elevated CO2)	FACEORE_3088280	IMG
Soil	Soil microbial communities from sample at FACE Site 5 Oak Ridge CO2+ (Oak Ridge elevated CO2)	FACEORE_3420340	IMG
Soil	Soil microbial communities from sample at FACE Site 5 Oak Ridge CO2+ (Oak Ridge elevated CO2)	FACEORE_460700	IMG
Soil	Soil microbial communities from sample at FACE Site North Carolina NCD_AmbF (NCD_AmbF)	FNCDAF_02995050	IMG
Soil	Soil microbial communities from sample at FACE Site North Carolina NCD_ElevF (NCD_ElevF)	FNCDEF_00392320	IMG
Soil	Soil microbial communities from sample at FACE Site North Carolina NCD_ElevF (NCD_ElevF)	FNCDEF_01952470	IMG
Soil	Soil microbial communities from sample at FACE Site North Carolina NCD_ElevF (NCD_ElevF)	FNCDEF_02993770	IMG
Soil	Soil microbial communities from sample at FACE Site North Carolina NCD_ElevF (NCD_ElevF)	FNCDEF_06633810	IMG
Soil	Soil microbial communities from sample at FACE Site North Carolina NCD_ElevF (NCD_ElevF)	FNCDEF_07045350	IMG
Soil	Soil microbial communities from sample at FACE Site North Carolina NCD_ElevF (NCD_ElevF)	FNCDEF_07891200	IMG
Soil	Soil microbial communities from sample at FACE Site North Carolina NCD_ElevF (NCD_ElevF)	FNCDEF_08134280	IMG
Soil	Soil microbial communities from sample at FACE Site North Carolina NCD_ElevF (NCD_ElevF)	FNCDEF_08985470	IMG
Soil	Soil microbial communities from sample at Multiple FACE and OTC sites (NTS_010)	FNTS010_01220420	IMG
Soil	Soil microbial communities from sample at FACE Site NTS_067 Nevada Test Site (NTS_067)	FNTS067_02733900	IMG
Soil	Soil microbial communities from sample at FACE Site NTS_067 Nevada Test Site (NTS_067)	FNTS067_08129300	IMG
Soil	Soil microbial communities from sample at FACE Site Metagenome WIR_Amb2 (WIR_Amb2)	FWIRA_01414090	IMG
Soil	Soil microbial communities from sample at FACE Site Metagenome WIR_Amb2 (WIR_Amb2)	FWIRA_01942640	IMG
Soil	Soil microbial communities from sample at FACE Site Metagenome WIR_Amb2 (WIR_Amb2)	FWIRA_02138960	IMG
Soil	Soil microbial communities from sample at FACE Site Metagenome WIR_Amb2 (WIR_Amb2)	FWIRA_06708920	IMG

Environment	Sample	Gene ID	Database
Soil	Soil microbial communities from sample at FACE Site Metagenome WIR_Amb2 (WIR_Amb2)	FWIRA_07260100	IMG
Soil	Soil microbial communities from sample at FACE Site Metagenome WIR_Amb2 (WIR_Amb2)	FWIRA_09944590	IMG
Soil	Soil microbial communities from sample at FACE Site Metagenome WIR_Amb2 (WIR_Amb2)	FWIRA_10216430	IMG
Soil	Soil microbial communities from sample at FACE Site Metagenome WIR_ElevOz2 (WIR_ElevOz2)	FWIRElOz_04401610	IMG
Soil	Soil microbial communities from sample at FACE Site Metagenome WIR_ElevOz2 (WIR_ElevOz2)	FWIRElOz_09905960	IMG
Soil	Soil microbial communities from sample at FACE Site Metagenome WIR_Oz2 (WIR_Oz2)	FWIROz_02244430	IMG
Soil	Soil microbial communities from sample at FACE Site Metagenome WIR_Oz2 (WIR_Oz2)	FWIROz_02997180	IMG
Soil	Soil microbial communities from sample at FACE Site Metagenome WIR_Oz2 (WIR_Oz2)	FWIROz_07632640	IMG
Soil	Soil microbial communities from Great Prairies, sample from Iowa, Continuous Corn soil (Iowa, Continuous Corn soil, Jan 2012 Assem MSU hiseq+gaii)	ICChiseqgaiiDRAFT_054 19862	IMG
Soil	Soil microbial communities from sample at FACE Site 3 Nevada Test Site Creosote CO2-	NTS_CREO_AMB_4781 140	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Permafrost Layer P1 (P1_CLC_pe)	P1_C_00500790	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Permafrost Layer P1 (P1_CLC_pe)	P1_C_00546230	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Permafrost Layer P3 (P3_CLC)	P3_CLC_00156350	IMG
Soil	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Permafrost Layer P3 (P3_CLC)	P3_CLC_02782870	IMG
Soil	Luquillo Experimental Forest Soil, Puerto Rico	prs_01142880	IMG
Soil	Luquillo Experimental Forest Soil, Puerto Rico	prs_03577960	IMG
Soil	Luquillo Experimental Forest Soil, Puerto Rico	prs_05145370	IMG
Soil	Luquillo Experimental Forest Soil, Puerto Rico	prs_06070740	IMG
Soil	Luquillo Experimental Forest Soil, Puerto Rico	prs_07393930	IMG
Soil	Switchgrass rhizosphere microbial community from Michigan, US, sample from East Lansing bulk soil	SRBS_1552100	IMG
Soil	Switchgrass rhizosphere microbial community from Michigan, US, sample from East Lansing bulk soil	SRBS_2164790	IMG
Wastewater	Candidatus Accumulibacter phosphatis Type I	CAPI_1455840	IMG
Wastewater	Candidatus Accumulibacter phosphatis Type I (Sanger/454/Illimina Metagenome Assembly < 97% CAP2UW1)	UW2007_01767420	IMG
Wood compost	Poplar biomass bioreactor microbial communities from Brookhaven National Lab, NY sample from anaerobic community	POPANAER_154960	IMG

Supplemental Table 3 | List of metagenomes that appear in Table 2.1

Environment	Project Name	Sample Name	ID	Database
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP11 from Octopus Springs	2014031007	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial communities from Yellowstone National Park, One Hundred Springs Plain, sample OSP_C (OSP_C)	2077657024	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP4 from Joseph's Coat Springs	2013843003	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP19 from Cistern Spring	2015219000	IMG
Hot spring	Sediment and Water microbial communities from Great Boiling Spring	Sediment microbial communities from Great Boiling Spring, Nevada, sample from cellulolytic enrichment CS 77C (GBS Cellulolytic enrichment CS 77C sediment, Combined June 2011 assem)	3300000106	IMG
Hot spring	Sediment and Water microbial communities from Great Boiling Spring	Sediment microbial communities from Great Boiling Spring, Nevada, sample from Cellulolytic enrichment CS 85C (GBS Cellulolytic enrichment CS 85C sediment, Feb 2012 assem)	3300000084	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP14 from OSP Spring	2013954001	IMG
Hot spring	Sediment and Water microbial communities from Great Boiling Spring	Sediment microbial communities from Great Boiling Spring, Nevada, sample from cellulolytic enrichment CS 77C (GBS Cellulolytic enrichment CS 77C sediment, Feb 2012 assem)	3300000085	IMG
Hot spring	Hot spring microbial communities from Yellowstone Bison Hot Spring Pool	1_050719N	2009439003	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP18 from Washburn Springs #1	2016842004	IMG
Hot spring	Sediment and Water microbial communities from Great Boiling Spring	Sediment microbial communities from Great Boiling Spring, Nevada, sample from Cellulolytic enrichment CS 85C (GBS Cellulolytic enrichment CS 85C sediment)	2100351009	IMG
Hot spring	Hot spring microbial communities from Yellowstone Bison Hot Spring Pool	4_050719Q	2010170002	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Bath Hot Springs, filamentous community	2007309000	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Bath Hot Springs, planktonic community	2007309001	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP10 from Narrow Gauge	2015391001	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP1 from Alice Springs, Crater Hills	2014031002	IMG
Hot spring	Sediment and Water microbial communities from Great Boiling Spring	Water microbial communities from Great Boiling Spring, Nevada, sample from cellulolytic enrichment S 77C (Cellulolytic enrichment S 77C water)	2149837005	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP9 from Dragon Spring, Norris Geyser Basin	2014031004	IMG
Hot spring	Sediment and Water microbial communities from Great Boiling Spring	Water microbial communities from Great Boiling Spring, Nevada, sample 1 (13 Aug 2010 assembly with PE data)	2077657003	IMG

Environment	Project Name	Sample Name	ID	Database
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP3 from Monarch Geyser, Norris Geyser Basin	2014031003	IMG
Hot spring	Sediment and Water microbial communities from Great Boiling Spring	Sediment microbial communities from Great Boiling Spring, Nevada, sample from Cellulolytic enrichment CS 85C (GBS Cellulolytic enrichment CS 85C sediment, Combined June 2011 assem)	3300000109	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial communities from Yellowstone National Park, One Hundred Springs Plain, sample OSP_D (OSP_D)	2140918001	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP2 from Nymph Lake 10	2015219001	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP17 from Obsidian Pool Prime	2016842005	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP20 from Bath Lake Vista Annex - Purple-Sulfur Mats	2016842008	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial communities from Yellowstone National Park, One Hundred Springs Plain, sample OSP_B (OSP_B)	2077657023	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Beowulf Spring, Yellowstone National Park, sample YNP_Beowulf Spring_E (YNP_Beowulf Spring_E)	2100351008	IMG
Hot spring	Sediment and Water microbial communities from Great Boiling Spring	Sediment microbial communities from Great Boiling Spring, Nevada, sample from cellulolytic enrichment CS 77C (Cellulolytic enrichment CS 77C sediment)	2088090027	IMG
Hot spring	Sediment and Water microbial communities from Great Boiling Spring	Sediment microbial communities from Great Boiling Spring, Nevada, sample from surface sediment (Surface sediment)	2053563014	IMG
Hot spring	Hot spring microbial communities from Yellowstone Bison Hot Spring Pool	5_050719P	2010170003	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP15 from Mushroom Spring	2015219002	IMG
Hot spring	Hot spring microbial communities from Yellowstone Bison Hot Spring Pool	3_050719R	2010170001	IMG
Hot spring	Hot spring microbial communities from Yellowstone Bison Hot Spring Pool	2_050719S	2009439000	IMG
Hot spring	Sediment and Water microbial communities from Great Boiling Spring	Sediment microbial communities from Great Boiling Spring, Nevada, sample from Cellulolytic enrichment Sediment 77C (Cellulolytic enrichment S 77C sediment)	2149837004	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP6 from White Creek Site 3	2013515000	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP13 from Bechler Spring	2013515002	IMG
Hot spring	Sediment and Water microbial communities from Great Boiling Spring	Water microbial communities from Great Boiling Spring, Nevada, sample from cellulolytic enrichment S 77C (GBS Cellulolytic enrichment S 77C water, Feb 2012 assem)	3300000083	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Beowulf Spring, Yellowstone National Park, sample YNP_Beowulf Spring_D (YNP_Beowulf Spring_D)	2119805007	IMG

Environment	Project Name	Sample Name	ID	Database
Hot spring	Hot Spring microbial communities from Yellowstone Obsidian Hot Spring	Hot Spring microbial communities from Yellowstone Obsidian Hot Spring, Sample 10594	2010170004	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP7 from Chocolate Pots	2014031006	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP8 from OSP Spring	2013515001	IMG
Hot spring	Sediment and Water microbial communities from Great Boiling Spring	Water microbial communities from Great Boiling Spring, Nevada, sample 1 (Water borne 27 Oct 2010 assembly)	2084038020	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP12 from Calcite Springs, Tower Falls Region	2014031005	IMG
Hot spring	Sediment and Water microbial communities from Great Boiling Spring	Water viral communities from Great Boiling Spring, Nevada (Water borne viral community)	2058419004	IMG
Hot spring	Sediment and Water microbial communities from Great Boiling Spring	Sediment microbial communities from Great Boiling Spring, Nevada, sample from Cellulolytic enrichment Sediment 77C (GBS Cellulolytic enrichment 77S sediment, Feb 2012 assem)	3300000082	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP5 from Bath Lake Vista Annex	2013954000	IMG
Hot spring	Hot spring microbial communities from Yellowstone National Park, US	Hot spring microbial community from Yellowstone Hot Springs, sample YNP16 from Fairy Spring Red Layer	2016842003	IMG
Hot spring	Yellowstone National Park Octopus/Mushroom Hot Springs Metagenome	MushroomSpringsMatCoreA	4443745.3	myMGDB/ MG-RAST
Hot spring	Yellowstone National Park Octopus/Mushroom Hot Springs Metagenome	MushroomSpringsMatCoreD	4443747.3	myMGDB/ MG-RAST
Hot spring	Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses	Octopus spring		myMGDB/ MG-RAST
Hot spring	Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses	Mushroom spring		myMGDB/ MG-RAST
Hot spring	Yellowstone National Park Octopus/Mushroom Hot Springs Metagenome	OctopusSpringsMatCoreF	4443749.3	myMGDB/ MG-RAST
Hot spring	Yellowstone National Park Octopus/Mushroom Hot Springs Metagenome	OctopusSpringsMatCoreR	4443750.3	myMGDB/ MG-RAST
Hot spring	Yellowstone National Park Octopus/Mushroom Hot Springs Metagenome	MushroomSpringsMatCoreB	4443746.3	myMGDB/ MG-RAST
Hot spring	Yellowstone National Park Octopus/Mushroom Hot Springs Metagenome	MushroomSpringsMatCoreF	4443762.3	myMGDB/ MG-RAST
Soil and rhizosphere	Soil microbial communities from Great Prairies (Kansas, Wisconsin and Iowa)	Soil microbial communities from Great Prairies, sample from Iowa, Continuous Corn soil (Iowa, Continuous Corn soil, Jan 2012 Assem MSU hiseq+gaii)	3300000033	IMG
Soil and rhizosphere	Soil microbial communities from four geographically distinct crusts in the Colorado Plateau and Sonoran desert	Soil microbial communities sample from Light Crust, Colorado Plateau, Green Butte (Light Crust, Colorado Plateau, Green Butte 2 June 2011 assem)	2199352006	IMG

Environment	Project Name	Sample Name	ID	Database
Soil and rhizosphere	Permafrost microbial communities from Centeral Alaska	Permafrost field sample	2067725009	IMG
Soil and rhizosphere	Soil microbial communities from Miscanthus in Kellogg Biological Station, MSU	Miscanthus rhizosphere microbial communities from Kellogg Biological Station, MSU, sample from Bulk Soil Replicate 2: eDNA_1 (Bulk soil 2 January 2011 combined assembly)	2124908025	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site North Carolina NCD_ElevF (NCD_ElevF)	2124908001	IMG
Soil and rhizosphere	Soil microbial communities from four geographically distinct crusts in the Colorado Plateau and Sonoran desert	Soil microbial communities sample from Dark Crust, Colorado Plateau, Green Butte (Dark Crust, Colorado Plateau, Green Butte June 2011 assem)	2209111000	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site Metagenome WIR_Elev2 (WIR_Elev2)	2124908008	IMG
Soil and rhizosphere	Soil microbial communities from four geographically distinct crusts in the Colorado Plateau and Sonoran desert	Soil microbial communities sample from Light Crust, Colorado Plateau, Green Butte 2 (Light Crust Colorado Plateau Green Butte 2, Oct 2011 assem)	3300000095	IMG
Soil and rhizosphere	Soil microbial communities from Puerto Rico rain forest, that decompose switchgrass	Soil microbial communities from Puerto Rico rain forest, that decompose switchgrass, sample from feedstock-adapted consortia SG only (SG only, May 2011 assembly)	2189573022	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site Metagenome WIR_Oz2 (WIR_Oz2)	2124908006	IMG
Soil and rhizosphere	Soil microbial communities from Puerto Rico rain forest, that decompose switchgrass	Soil microbial communities from Puerto Rico rain forest, that decompose switchgrass, sample from Feedstock-adapted consortia SG + Fe (SG + Fe)	2119805008	IMG
Soil and rhizosphere	Switchgrass rhizosphere microbial community from Michigan, US	Switchgrass rhizosphere microbial community from Michigan, US, sample from Rose Lake bulk soil RL3 (Bulk soil RL3 January 2011 combined assembly)	2124908021	IMG
Soil and rhizosphere	N/A	Switchgrass rhizosphere microbial community from Michigan, US, sample from East Lansing bulk soil (Bulk soil GOTP January 2011 combined assembly)	2124908023	IMG
Soil and rhizosphere	Soil microbial communities from permafrost in Bonanza Creek, Alaska	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Permafrost Layer P1 (P1_CLC_pe)	2140918006	IMG
Soil and rhizosphere	Soil microbial communities from permafrost in Bonanza Creek, Alaska	Permafrost metatranscriptome cDNA-P1	3300000338	IMG
Soil and rhizosphere	Soil microbial communities from permafrost in Bonanza Creek, Alaska	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Active Layer A2 (A2_CLC_pe)	2124908043	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site 4 Nevada Test Site Crust CO2-	2032320003	IMG
Soil and rhizosphere	Soil microbial communities from Puerto Rico rain forest, that decompose switchgrass	Soil microbial communities from Puerto Rico rain forest, that decompose switchgrass, sample from Feedstock-adapted consortia SG + Fe (SG + Fe, May 2011 assembly)	2162886008	IMG
Soil and rhizosphere	Soil microbial communities from permafrost in Bonanza Creek, Alaska	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Bog Site B3 (B3_all_CLC)	2124908038	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site Metagenome WIR_ElevOz2 (WIR_ElevOz2)	2124908007	IMG
Soil and rhizosphere	Soil microbial communities from permafrost in Bonanza Creek, Alaska	Permafrost metatranscriptome cDNA-P3	3300000337	IMG

Environment	Project Name	Sample Name	ID	Database
Soil and rhizosphere	Soil microbial communities from permafrost in Bonanza Creek, Alaska	Thermokarst Bog cDNA B3	3300000336	IMG
Soil and rhizosphere	Soil microbial communities from permafrost in Bonanza Creek, Alaska	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Permafrost Layer P3 (P3_CLC)	2124908041	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site 5 Oak Ridge CO2- (Oak Ridge ambient)	2032320005	IMG
Soil and rhizosphere	Soil microbial communities from Waseca County, Minnesota Farm	Soil microbial communities from Minnesota Farm	2001200001	IMG
Soil and rhizosphere	Soil microbial communities from switchgrass rhizosphere	Maize field bulk soil microbial communities from University of Illinois Energy Farm, Urbana, IL (Bulk soil sample from field growing corn (Zea mays))	2044078000	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at Multiple FACE and OTC sites (NTS_010)	2119805011	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site Metagenome WIR_Amb2 (WIR_Amb2)	2124908009	IMG
Soil and rhizosphere	Switchgrass rhizosphere microbial community from Michigan, US	Switchgrass rhizosphere microbial community from Michigan, US, sample from Rose Lake bulk soil RL2 (Bulk soil RL2 April 2011 assembly)	2162886013	IMG
Soil and rhizosphere	Soil microbial communities from permafrost in Bonanza Creek, Alaska	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Active Layer A5 (A5_CLC_pe)	2124908044	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site 2 North Carolina CO2-	2035918004	IMG
Soil and rhizosphere	Soil microbial communities from Miscanthus in Kellogg Biological Station, MSU	Miscanthus rhizosphere microbial communities from Kellogg Biological Station, MSU, sample Bulk Soil Replicate 1 : eDNA_1 (Bulk soil 1 April 2011 assembly)	2162886012	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site 3 Nevada Test Site Creosote CO2-	2029527002	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site NTS_007 Nevada Test Site (NTS_007)	2119805009	IMG
Soil and rhizosphere	Soil microbial communities from permafrost in Bonanza Creek, Alaska	Thermokarst Bog cDNA B4	3300000334	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site 5 Oak Ridge CO2+ (Oak Ridge elevated CO2)	2032320006	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site 3 Nevada Test Site Creosote CO2+	2032320002	IMG
Soil and rhizosphere	Green-waste compost microbial community from soild state bioreactor	Luquillo Experimental Forest Soil, Puerto Rico	2070309004	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site NTS_067 Nevada Test Site (NTS_067)	2119805012	IMG
Soil and rhizosphere	Soil microbial communities from switchgrass rhizosphere	Miscanthus field bulk soil microbial communities from University of Illinois Energy Farm, Urbana, IL (Bulk soil sample from field growing Miscanthus x giganteus)	2044078003	IMG
Soil and rhizosphere	Soil microbial pyrene-degrading mixed culture	Bacterial pyrene-degrading mixed culture	2021593002	IMG
Soil and rhizosphere	Soil microbial communities from permafrost in Bonanza Creek, Alaska	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Bog Site B4 (B4_CLC)	2124908040	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site 2 North Carolina CO2+ (North Carolina Elevated CO2)	2040502001	IMG

Environment	Project Name	Sample Name	ID	Database
Soil and rhizosphere	Soil microbial community from Bioreactor with Chloroethene contaminated sediment	Soil microbial community from bioreactor at Alameda Naval Air Station, CA, contaminated with Chloroethene, Sample 196	2014730001	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site NTS_071 Nevada Test Site (NTS_071)	2081372006	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site North Carolina NCD_AmbF (NCD_AmbF)	2119805010	IMG
Soil and rhizosphere	Soil microbial communities from Puerto Rico rain forest, that decompose switchgrass	Soil microbial communities from Puerto Rico rain forest, that decompose switchgrass, sample from feedstock-adapted consortia SG only (SG only)	2088090029	IMG
Soil and rhizosphere	Soil microbial community from Bioreactor with Chloroethene contaminated sediment	Soil microbial community from bioreactor at Alameda Naval Air Station, CA, contaminated with Chloroethene, Sample 196 (Jan 2009 assem)	2199034002	IMG
Soil and rhizosphere	Soil microbial communities from switchgrass rhizosphere	Switchgrass field bulk soil microbial communities from University of Illinois Energy Farm, Urbana, IL (Bulk soil sample from field growing switchgrass (Panicum virgatum))	2044078005	IMG
Soil and rhizosphere	N/A	Switchgrass rhizosphere microbial community from Michigan, US, sample from East Lansing bulk soil	2021593004	IMG
Soil and rhizosphere	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site 4 Nevada Test Site Crust CO2+ (NTS Crust elevated CO2)	2035918005	IMG
Soil and rhizosphere	Soil microbial communities from permafrost in Bonanza Creek, Alaska	Soil microbial communities from permafrost in Bonanza Creek, Alaska, sample from Active Layer A5	3300000335	IMG
Soil and rhizosphere	Soil microbial communities from switchgrass rhizosphere	Miscanthus rhizosphere soil microbial communities from University of Illinois Energy Farm, Urbana, IL (Rhizosphere soil sample of Miscanthus x giganteus)	2044078002	IMG
Soil and rhizosphere	Switchgrass rhizosphere microbial community from Michigan, US	Switchgrass rhizosphere microbial community from Michigan, US, sample from Rose Lake RL3 (Rhizosphere RL3 April 2011 assembly)	2162886006	IMG
Soil and rhizosphere	Switchgrass rhizosphere microbial community from Michigan, US	Switchgrass rhizosphere microbial community from Michigan, US, sample from Rose Lake rhizosphere BV2.2 (BV2.2 January 2011 combined assembly)	2124908018	IMG
Soil and rhizosphere	Soil microbial communities from Miscanthus in Kellogg Biological Station, MSU	Miscanthus rhizosphere microbial communities from Kellogg Biological Station, MSU, sample Rhizosphere Soil Replicate 1: eDNA_1 (Rhizosphere replicate 1 April 2011 assembly)	2162886011	IMG
Soil and rhizosphere	Microbial communities from Arabidopsis rhizosphere	Arabidopsis rhizosphere microbial communities from University of North Carolina, sample from Arabidopsis Col-0 old rhizosphere (Arabidopsis Col-0 old rhizosphere, Nov 2011 assem)	3300000045	IMG
Soil and rhizosphere	N/A	Switchgrass rhizosphere microbial community from Michigan, US, sample from East Lansing, 10341 (454/Illumina contigs)	2040502002	IMG
Soil and rhizosphere	Switchgrass rhizosphere microbial community from Michigan, US	Switchgrass rhizosphere microbial community from Michigan, US, sample from Rose Lake rhizosphere RL2 (Rhizosphere RL2 April 2011 assembly)	2162886007	IMG
Soil and rhizosphere	Soil microbial communities from switchgrass rhizosphere	Switchgrass soil microbial communities from University of Illinois Energy Farm, Urbana, IL (Rhizosphere soil sample from switchgrass (Panicum virgatum))	2044078004	IMG

Environment	Project Name	Sample Name	ID	Database
Soil and rhizosphere	Microbial communities from Arabidopsis rhizosphere	Arabidopsis rhizosphere microbial communities from University of North Carolina, sample Wild type Col-0 (Arabidopsis rhizosphere microbiome-wild type Col-0 454/Illumina 2011 July Assem)	2209111006	IMG
Soil and rhizosphere	Microbial communities from Arabidopsis rhizosphere	Arabidopsis rhizosphere microbial communities from University of North Carolina, sample from Arabidopsis soil young (Arabidopsis soil young, Nov 2011 assem)	3300000042	IMG
Soil and rhizosphere	Switchgrass rhizosphere microbial community from Michigan, US	Switchgrass rhizosphere microbial community from Michigan, US, sample from Buena Vista Grasslands Wildlife Area, Rhizosphere BV2.1 (BV2.1 January 2011 combined assembly)	2124908019	IMG
Soil and rhizosphere	Soil microbial communities from switchgrass rhizosphere	Maize rhizosphere soil microbial communities from University of Illinois Energy Farm, Urbana, IL (Soil sample from rhizosphere of corn (Zea mays))	2044078001	IMG
Soil and rhizosphere	Microbial communities from Arabidopsis rhizosphere	Arabidopsis rhizosphere microbial communities from University of North Carolina, sample from Arabidopsis cpr5 young rhizosphere (Arabidopsis cpr5 young rhizosphere, Nov 2011 assem)	3300000043	IMG
Soil and rhizosphere	Soil microbial communities from Miscanthus in Kellogg Biological Station, MSU	Miscanthus rhizosphere microbial communities from Kellogg Biological Station, MSU, sample Rhizosphere Soil Replicate 2: eDNA_1 (Rhizo 2 January 2011 combined assembly)	2124908027	IMG
Soil and rhizosphere	Microbial communities from Arabidopsis rhizosphere	Arabidopsis rhizosphere microbial communities from University of North Carolina, sample Mutant cpr5 (cpr5 454/Illumina combined assembly)	2100351005	IMG
Soil and rhizosphere	Microbial communities from Arabidopsis rhizosphere	Arabidopsis rhizosphere microbial communities from University of North Carolina, sample from Arabidopsis soil old (Arabidopsis soil old, Nov 2011 assem)	3300000044	IMG
Soil and rhizosphere	Microbial communities from Arabidopsis rhizosphere	Arabidopsis rhizosphere microbial communities from University of North Carolina, sample from Arabidopsis cpr5 old rhizosphere (Arabidopsis cpr5 old rhizosphere, Nov 2011 assem)	3300000041	IMG
Insect fungal gardens	Fungus garden microbial communities from Acromyrmex echinatior in Panama	Fungus garden combined (combined)	2035918000	IMG
Insect fungal gardens	N/A	Dump bottom (Dump bottom)	2032320007	IMG
Insect fungal gardens	N/A	Dump top (Dump top)	2030936006	IMG
Insect fungal gardens	Fungus gallery microbial communities from Dendroctonus ponderosae	Dendroctonus ponderosae beetle community (MPB hybrid beetle) (Lodgepole pine)	2032320008	IMG
Insect fungal gardens	Fungus gallery microbial communities from Dendroctonus ponderosae	Dendroctonus ponderosae fungus gallery (Hybrid pine) (MPB hybrid gallery)	2029527007	IMG
Insect fungal gardens	N/A	Fungus garden microbial communities from Atta colombica in Panama, sample from dump bottom (Dump bottom)	2040502000	IMG
Insect fungal gardens	Xyleborus affinis microbiome from Bern, Switzerland	Xyleborus affinis microbiome from Bern, Switzerland, sample of gallery community (Gallery community)	2084038008	IMG
Insect fungal gardens	Fungus-growing Termite Fungus Garden	Fungus garden microbial community from termites in South Africa, sample from Oerleman's Farm	2065487014	IMG
Insect fungal gardens	N/A	Fungus garden microbial communities from Atta colombica in Panama, sample from fungus garden top	2029527005	IMG

Environment	Project Name	Sample Name	ID	Database
Insect fungal gardens	Mountain Pine Beetle microbial communities from Grand Prairie, Alberta	Mountain Pine Beetle microbial communities from McBride, British Columbia, Canada, sample from Lodgepole pine (Lodgepole pine)	2035918003	IMG
Insect fungal gardens	Mountain Pine Beetle microbial communities from Grand Prairie, Alberta	Mountain Pine Beetle microbial communities from Grand Prairie, Alberta, sample from Hybrid pine (MPB hybrid beetle)	2032320009	IMG
Insect fungal gardens	Fungus garden microbial communities from Atta cephalotes	Atta cephalotes fungus garden (ACEF)	2029527004	IMG
Insect fungal gardens	Fungus garden microbial communities from Apterostigma dentigerum	Apterostigma fungus garden Combined	2029527003	IMG
Insect fungal gardens	Xyleborus affinis microbiome from Bern, Switzerland	Xyleborus affinis microbiome from Bern, Switzerland, sample of adult community (Ambrosia beetle adult)	2043231000	IMG
Insect fungal gardens	Fungus garden microbial communities from Trachymyrmex in Gamboa, Panama	Trachymyrmex fungus garden	2084038018	IMG
Insect fungal gardens	N/A	Fungus garden microbial communities from Atta colombica in Panama, sample from fungus garden bottom (Fungus garden bottom)	2029527006	IMG
Insect fungal gardens	Xyleborus affinis microbiome from Bern, Switzerland	Xyleborus affinis microbiome from Bern, Switzerland, sample of larvae (Larvae community)	2044078011	IMG
Insect fungal gardens	N/A	Fungus garden microbial communities from Atta colombica in Panama, sample from dump top (Dump top)	2038011000	IMG
Insect fungal gardens	N/A	Dendroctonus frontalis Fungal community	2044078007	IMG
Insect fungal gardens	Fungus garden microbial communities from Cyphomyrmex longiscapus	Cyphomyrmex longiscapus fungus garden	2030936005	IMG
Wood compost	Poplar biomass decaying microbial community	Poplar biomass bioreactor microbial communities from Brookhaven National Lab, NY sample from anaerobic community	2010388001	IMG
Wood compost	Poplar biomass decaying microbial community	Poplar biomass bioreactor microbial communities from Brookhaven National Lab, NY, sample from pooled GH fosmids	2020627002	IMG
Wood compost	Poplar biomass decaying microbial community	Poplar biomass bioreactor microbial communities from Brookhaven National Lab, NY, sample from total biomass decay community (13 April 2010 assembly with 454 paired-end)	2048955003	IMG
Wood compost	Decomposing wood compost microbial communities from rain forest habitat in Puerto Rico, that are thermophilic	Thermal compost enrichment from Puerto Rico rainforest (Compost thermophiles - Biomass metagenome May 2011 assem)	2199352035	IMG
Lentic and lotic	N/A	Lotic microbial communities from Mississippi River at two locations in the state of Minnesota, sample from River Site 1, Mississippi Headwaters	3300000206	IMG
Lentic and lotic	N/A	Lotic microbial communities from Mississippi River at two locations in the state of Minnesota, sample from River Site 7, Mississippi Headwaters (Site 7)	3300000269	IMG
Lentic and lotic	N/A	Lentic microbial communities from Lake Waban, Wellesley MA, that are anoxygenic and photosynthetic, photosynthetic consortia incandescent light (FW_incandescent_CN)	3300000497	IMG
Lentic and lotic	N/A	Lentic microbial communities from Wellesley MA, that are anoxygenic and photosynthetic, sample Photosynthetic Consortia 720nm	3300000496	IMG

Environment	Project Name	Sample Name	ID	Database
Lentic and lotic	Fresh water microbial communities from LaBonte Lake	Fresh water microbial communities from LaBonte Lake, Laramie, Wyoming, sample from pre-bloom (pre-bloom)	2149837010	IMG
Lentic and lotic	Fresh water microbial communities from LaBonte Lake	Fresh water microbial communities from LaBonte Lake, Laramie, Wyoming, sample from algal/cyanobacterial bloom material peak-bloom 1 (algal/cyano bloom peak-bloom 1)	2166559022	IMG
Lentic and lotic	Fresh water microbial communities from LaBonte Lake	Fresh water microbial communities from LaBonte Lake, Laramie, Wyoming, sample from peak-bloom 2 (Peak bloom metagenome 2)	2166559021	IMG
Lentic and lotic	Fresh water microbial communities from LaBonte Lake	Fresh water microbial communities from LaBonte Lake, Laramie, Wyoming, sample from algal/cyanobacterial bloom material peak-bloom 2 (algal/cyano bloom peak-bloom 2)	2189573023	IMG
Lentic and lotic	Fresh water microbial communities from LaBonte Lake	Fresh water microbial communities from LaBonte Lake, Laramie, Wyoming, sample from post-bloom (post-bloom)	2149837011	IMG
Lentic and lotic	Fresh water microbial communities from LaBonte Lake	Fresh water microbial communities from LaBonte Lake, Laramie, Wyoming, sample from peak-bloom 1 (Peak bloom metagenome 1)	2166559023	IMG
Lentic and lotic	Freshwater microbial communities from Lake Kinneret	Aquatic microbial communities from Lake Kinneret (02)	2010483006	IMG
Lentic and lotic	Freshwater microbial communities from Lake Kinneret	Aquatic microbial communities from Lake Kinneret (08)	2010483000	IMG
Lentic and lotic	Freshwater microbial communities from Lake Kinneret	Aquatic microbial communities from Lake Kinneret (03)	2010483002	IMG
Lentic and lotic	Freshwater microbial communities from Lake Kinneret	Aquatic microbial communities from Lake Kinneret (01)	2010483005	IMG
Lentic and lotic	Freshwater microbial communities from Lake Kinneret	Aquatic microbial communities from Lake Kinneret (06)	2010483001	IMG
Lentic and lotic	Freshwater microbial communities from Lake Kinneret	Aquatic microbial communities from Lake Kinneret (07)	2010483007	IMG
Lentic and lotic	Freshwater microbial communities from Lake Kinneret	Aquatic microbial communities from Lake Kinneret (04)	2010483003	IMG
Lentic and lotic	Freshwater microbial communities from Lake Kinneret	Aquatic microbial communities from Lake Kinneret (05)	2010483004	IMG
Lentic and lotic	Freshwater microbial communities from Lake Sakinaw in Canada	Sakinaw Lake metagenomics (120m) (Sakinaw Lake metagenomic (120m), Feb 2012 assem)	2263328000	IMG
Lentic and lotic	Freshwater microbial communities from Lake Sakinaw in Canada	Sakinaw Lake 454 metagenomics (120m): eDNA_2 (120 m)	2088090031	IMG
Lentic and lotic	Freshwater microbial communities from Lake Vostok at Ice accretion	Freshwater microbial communities from Lake Vostok at Ice accretion (5G Core, April 2011 assem (454, Ilumina combined))	2222084007	IMG
Lentic and lotic	Freshwater microbial communities from Mississippi River	Lake Itasca #1 (Itasca #1)	2077657006	IMG
Lentic and lotic	Freshwater microbial communities from Mississippi River	Minneapolis Minnesota #1 (Minneapolis #1)	2077657007	IMG
Lentic and lotic	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, IL	Freshwater microbial communities from Trout Bog Lake, WI, sample from Practice 18AUG2009 hypolimnion (Trout Bog Practice 18AUG2009 hypolimnion June 2011 assem)	2199352002	IMG
Lentic and lotic	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, IL	Freshwater microbial communities from Trout Bog Lake, WI, sample from Practice 03JUN2009 hypolimnion (Trout Bog Practice 03JUN2009 hypolimnion June 2011 assem)	2199352000	IMG

Environment	Project Name	Sample Name	ID	Database
Lentic and lotic	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, IL	Freshwater microbial communities from Trout Bog Lake, WI sample from Practice 18AUG2009 epilimnion (Trout Bog Practice 18AUG2009 epilimnion June 2011 assem)	2199352001	IMG
Lentic and lotic	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, IL	Freshwater microbial communities from Trout Bog Lake, WI sample from Practice 03JUN2009 epilimnion (Test dataset: Trout Bog Practice 03JUN2009 epilimnion 1000 subsample, June 2011)	3300000220	IMG
Lentic and lotic	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, IL	Freshwater microbial communities from Lake Mendota, WI, sample from Practice 20APR2010 epilimnion (Lake Mendota Practice 20APR2010 epilimnion June 2011 assem)	2199352003	IMG
Lentic and lotic	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, IL	Freshwater microbial communities from Trout Bog Lake, WI, sample from Practice 18AUG2009 hypolimnion (Trout Bog Practice 18AUG2009 hypolimnion June 2011 assem)	3300000177	IMG
Lentic and lotic	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, IL	Freshwater microbial communities from Trout Bog Lake, WI sample from Practice 03JUN2009 epilimnion (Trout Bog Practice 03JUN2009 epilimnion June 2011 assem)	3300000176	IMG
Lentic and lotic	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, IL	Freshwater microbial communities from Trout Bog Lake, WI, sample from Practice 03JUN2009 hypolimnion (Trout Bog Practice 03JUN2009 hypolimnion June 2011 assem)	3300000162	IMG
Lentic and lotic	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, IL	Freshwater microbial communities from Trout Bog Lake, WI sample from Practice 03JUN2009 epilimnion (Trout Bog Practice 03JUN2009 epilimnion June 2011 assem)	2199034001	IMG
Lentic and lotic	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, IL	Freshwater microbial communities from Lake Mendota, WI, sample from Practice 15JUN2010 epilimnion (Lake Mendota Practice 15JUN2010 epilimnion June 2011 assem)	2199352004	IMG
Lentic and lotic	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, IL	Freshwater microbial communities from Trout Bog Lake, WI sample from Practice 18AUG2009 epilimnion (Trout Bog Practice 18AUG2009 epilimnion June 2011 assem)	3300000203	IMG
Lentic and lotic	Freshwater microbial communities from Trout Bog Lake, WI and Lake Mendota, IL	Freshwater microbial communities from Lake Mendota, WI, sample from Practice 29OCT2010 epilimnion (Lake Mendota Practice 29OCT2010 epilimnion June 2011 assem)	2199352005	IMG
Lentic and lotic	Sediment methylotrophic communities from Lake Washington	Methylotrophic community from Lake Washington sediment Formaldehyde enrichment	2006207003	IMG
Lentic and lotic	Sediment methylotrophic communities from Lake Washington	Methylotrophic community from Lake Washington sediment Methanol enrichment	2006207001	IMG
Lentic and lotic	Sediment methylotrophic communities from Lake Washington	Methylotrophic community from Lake Washington sediment combined (v2)	2006543005	IMG
Lentic and lotic	Sediment methylotrophic communities from Lake Washington	Methylotrophic community from Lake Washington sediment Formate enrichment	2006207004	IMG
Lentic and lotic	Sediment methylotrophic communities from Lake Washington	Methylotrophic community from Lake Washington sediment Methylamine enrichment	2006207002	IMG
Lentic and lotic	Sediment methylotrophic communities from Lake Washington	Methylotrophic community from Lake Washington sediment Methane enrichment	2006207000	IMG

Environment	Project Name	Sample Name	ID	Database
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, sample SIP 13C methane aerobic+nitrate (Aerobic with added nitrate, 13C SIP)	2046860006	IMG
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, original sample replicate 2 (Original sample replicate 2 12C fraction)	2088090006	IMG
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, sample SIP 13Cmethane anaerobic no nitrate (Anaerobic without added nitrate, 13C SIP)	2046860008	IMG
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, original sample replicate 1 (Original sample replicate 1)	2088090005	IMG
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, sample Microc enrich af exp to meth lab w 13Ccarbon-no added nitrate (Aerobic without added nitrate, 13C SIP)	2046860007	IMG
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, sample from SIP 13C-methane aerobic no nitrate additional fraction (Aerobic without added nitrate, SIP additional fraction)	2046860005	IMG
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, sample from flow sorted aerobic no nitrate (Flow sorted aerobic no nitrate)	2084038009	IMG
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, sample from flow sorted anaerobic no nitrate (Flow sorted anaerobic no nitrate Feb 2011 assembly)	2140918012	IMG
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, sample SIP 13C-methane anaerobic+nitrate (Anaerobic + nitrate SIP Nov 2010 with PE)	2088090009	IMG
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, sample from flow sorted anaerobic no nitrate (WGA anaerobic no nitrate)	2124908000	IMG
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, original sample replicate 1	2149837030	IMG
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, sample SIP 13Cmethane anaerobic no nitrate (Anaerobic no nitrate SIP Nov 2010 with PE)	2088090013	IMG
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, sample SIP 13C-methane anaerobic+nitrate (Anaerobic with added nitrate, 13C SIP)	2046860004	IMG
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, sample from flow sorted anaerobic plus nitrate (Flow sorted anaerobic plus nitrate)	2088090007	IMG

Environment	Project Name	Sample Name	ID	Database
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, sample from flow sorted aerobic plus nitrate (Flow sorted aerobic plus nitrate)	2100351007	IMG
Lentic and lotic	Sediment microbial communities from Lake Washington for Methane and Nitrogen Cycles	Sediment microbial communities from Lake Washington, Seattle, for Methane and Nitrogen Cycles, original sample replicate 1	2149837029	IMG
Lentic and lotic	Lake Huron sinkhole microbial mat community	Sinkhole freshwater microbial communities from Lake Huron, US, Sample 419	2049941002	IMG
Lentic and lotic	Lake Huron sinkhole microbial mat community	Sinkhole freshwater microbial communities from Lake Huron, US, Ph40x	2065487012	IMG
Groundwater	Groundwater dechlorinating community (KB-1) from synthetic mineral medium	Groundwater dechlorinating community (KB-1) from synthetic mineral medium in Toronto, ON, sample from Site contaminated with chlorinated ethenes	2013843002	IMG
Groundwater	Groundwater dechlorinating microbial community from Kitchener, Ontario, containing dehalobacter	TCA/MEAL culture (TCA/MEAL culture Nov 2010 assembly with PE data)	2100351010	IMG
Groundwater	Ground water	Oak Ridge Pristine Groundwater FRC FW301	2007427000	IMG
Groundwater	Ground water	Uranium Contaminated Groundwater FW106	2006543007	IMG
Groundwater	Soil microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy	Groundwater microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy, sample from two redox zones- GS10_10 (Targeted Biofilm samples from two redox zones-GS10_10, Oct 2011 Assem)	3300000230	IMG
Groundwater	Soil microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy	Groundwater microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy, sample from two redox zones- LI09_4 (Targeted Biofilm samples from two redox zones-LI09_4, Oct 2011 Assem)	3300000231	IMG
Groundwater	Soil microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy	Groundwater microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy, sample from two redox zones- PC08_64 (Targeted Biofilm samples from two redox zones-PC08_64, Feb 2012 Assem)	3300000232	IMG
Groundwater	Soil microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy	Groundwater microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy, sample from two redox zones- PC08_66 (Targeted Biofilm samples from two redox zones-PC08_66, Dec 2011 Assem)	3300000228	IMG
Groundwater	Soil microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy	Groundwater microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy, sample from two redox zones- GS09_5 (Targeted Biofilm samples from two redox zones-GS09_5, Oct 2011 Assem)	3300000234	IMG
Groundwater	Soil microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy	Groundwater microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy, sample from two redox zones- FS06_10 (Targeted Biofilm samples from two redox zones-FS06_10, Dec 2011 Assem)	3300000233	IMG
Groundwater	Soil microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy	Groundwater microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy, sample from two redox zones- AS07_7 (Targeted Biofilm samples from two redox zones-AS07_7, Dec 2011 Assem)	3300000227	IMG

Environment	Project Name	Sample Name	ID	Database
Groundwater	Soil microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy	Groundwater microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy, sample from two redox zones- FS08_3 (Targeted Biofilm samples from two redox zones-FS08_3, Oct 2011 Assem)	3300000236	IMG
Groundwater	Soil microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy	Groundwater microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy, sample from two redox zones- LI09_3 (Targeted Biofilm samples from two redox zones-LI09_3, Jan 2012 Assem)	3300000229	IMG
Groundwater	Soil microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy	Groundwater microbial communities from subsurface biofilms in sulfidic aquifier in Frasassi Gorge, Italy, sample from two redox zones- PC08_3 (Targeted Biofilm samples from two redox zones-PC08_3, Oct 2011 Assem)	3300000235	IMG
Wastewater	Wastewater Terephthalate-degrading communities from Bioreactor	TA reactor DNA contigs from 4 sample (Terephthalate degrading reactor metagenome contigs from 4 samples)	2081372008	IMG
Wastewater	Freshwater propionate Anammox bacterial community from bioreactor in Nijmegen, The Netherlands	Bioreactor Anammox bacterial community from Nijmegen, The Netherlands, sample from Brocadia fulgida enrichment	2030936003	IMG
Wastewater	Wastewater treatment Type I Accumulibacter community from EBPR Bioreactor	Candidatus Accumulibacter phosphatis Type I	2022004001	IMG
Wastewater	Freshwater propionate Anammox bacterial community from bioreactor in Nijmegen, The Netherlands	Bioreactor Anammox bacterial community from Nijmegen, The Netherlands, sample from Scalindua species enirchment	2017108002	IMG
Wastewater	US sludge - combined Sanger 454 assembly	Sludge/Australian, Phrap Assembly	2000000001	IMG
Wastewater	N/A	Sludge/US Virion (fgenesb)	2007300000	IMG
Wastewater	Wastewater Terephthalate-degrading communities from Bioreactor	Wastewater Terephthalate-degrading communities from Bioreactor	2007915000	IMG
Wastewater	Biofuel metagenome	Biofuel metagenome 3 (Biofuel metagenome 3 Illumina assembly)	2166559024	IMG
Wastewater	Wastewater treatment plant plasmid pool from Switzerland	Activated sludge plasmid pool Visp-2009 (Newbler)	2035918001	IMG
Wastewater	Freshwater propionate Anammox bacterial community from bioreactor in Nijmegen, The Netherlands	Bioreactor Anammox bacterial community from Nijmegen, The Netherlands, sample from Brocadia fulgida enrichment (Brocadia contigs)	2225789020	IMG
Wastewater	Freshwater propionate Anammox bacterial community from bioreactor in Nijmegen, The Netherlands	Bioreactor Anammox bacterial community from Nijmegen, The Netherlands, sample from Scalindua species enirchment	2022004002	IMG
Wastewater	Wastewater treatment plant plasmid pool from Switzerland	Activated sludge plasmid pool Morges (MIRA contigs) (MIRA contigs, 5x coverage)	2209111023	IMG
Wastewater	Wastewater treatment plant plasmid pool from Switzerland	Activated sludge plasmid pool Morges-2007 (PGA)	2013843001	IMG
Wastewater	Wastewater treatment Type I Accumulibacter community from EBPR Bioreactor	Candidatus Accumulibacter phosphatis Type I (Sanger/454/Illimina Metagenome Assembly < 97% CAP2UW1)	2100351003	IMG
Wastewater	US sludge - combined Sanger 454 assembly	Sludge/US, Phrap Assembly	2000000000	IMG
Wastewater	N/A	Sludge/US, Jazz Assembly	2001000000	IMG
Wastewater	Biofuel metagenome	Biofuel metagenome 3 (Biofuel metagenome 3 July 2011 assem)	2199352027	IMG
Open Ocean	Global Ocean Sampling Expedition	GS041 Shotgun - Open Ocean - Tropical South Pacific - Tropical South Pacific - International	4441126.3	myMGDB/ MG-RAST

Environment	Project Name	Sample Name	ID	Database
Open Ocean	Global Ocean Sampling Expedition	GS122b Shotgun - Open Ocean - Indian Ocean - International waters between Madagascar and South Africa - International	4441139.4	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS044 Shotgun - Open Ocean - Tropical South Pacific - 600 miles from F. Polynesia - International	4441129.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS122a Shotgun - Open Ocean - Indian Ocean - International waters between Madagascar and South Africa - International	4441615.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS115 Shotgun - Open Ocean - Indian Ocean - Indian Ocean - International	4441150.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS039 Shotgun - Open Ocean - Tropical South Pacific - Tropical South Pacific - International	4441136.3	myMGDB/ MG-RAST
Open Ocean	The Sorcerer II Global Ocean Sampling expedition	GS00a		myMGDB/ MG-RAST
Open Ocean	The Sorcerer II Global Ocean Sampling expedition	GS000b_11	4441572.3	myMGDB/ MG-RAST
Open Ocean	The Sorcerer II Global Ocean Sampling expedition	GS000c Shotgun - Open Ocean - Sargasso Sea - Sargasso Stations 3 - Bermuda	4441574.3	myMGDB/ MG-RAST
Open Ocean	The Sorcerer II Global Ocean Sampling expedition	GS000d Shotgun - Open Ocean - Sargasso Sea - Sargasso Station 13 - Bermuda	4441575.3	myMGDB/ MG-RAST
Open Ocean	The Sorcerer II Global Ocean Sampling expedition	GS001a Shotgun - Open Ocean - Sargasso Sea - Hydrostation S - Bermuda	4441576.3	myMGDB/ MG-RAST
Open Ocean	The Sorcerer II Global Ocean Sampling expedition	GS001b Shotgun - Open Ocean - Sargasso Sea - Hydrostation S - Bermuda	4441577.3	myMGDB/ MG-RAST
Open Ocean	The Sorcerer II Global Ocean Sampling expedition	GS001c Shotgun - Open Ocean - Sargasso Sea - Hydrostation S - Bermuda	4441578.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS045 Shotgun - Open Ocean - Tropical South Pacific - 400 miles from F. Polynesia - International	4441130.3	myMGDB/ MG-RAST
Open Ocean	The Sorcerer II Global Ocean Sampling expedition	GS017 Shotgun - Open Ocean - Caribbean Sea - Yucatan Channel - Mexico	4441587.3	myMGDB/ MG-RAST
Open Ocean	The Sorcerer II Global Ocean Sampling expedition	GS018 Shotgun - Open Ocean - Caribbean Sea - Rosario Bank - Honduras	4441588.3	myMGDB/ MG-RAST
Open Ocean	The Sorcerer II Global Ocean Sampling expedition	GS022 Shotgun - Open Ocean - Eastern Tropical Pacific - 250 miles from Panama City - Panama	4441592.3	myMGDB/ MG-RAST
Open Ocean	The Sorcerer II Global Ocean Sampling expedition	GS023 Shotgun - Open Ocean - Eastern Tropical Pacific - 30 miles from Cocos Island - Costa Rica	4441661.3	myMGDB/ MG-RAST
Open Ocean	The Sorcerer II Global Ocean Sampling expedition	GS026 Shotgun - Open Ocean - Galapagos Islands - 134 miles NE of Galapagos - Ecuador	4441594.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS109 Shotgun - Open Ocean - Indian Ocean - Indian Ocean - International	4441155.3	myMGDB/ MG-RAST
Open Ocean	The Sorcerer II Global Ocean Sampling expedition	GS037 Shotgun - Open Ocean - Eastern Tropical Pacific - Equatorial Pacific TAO Buoy - International	4441145.3	myMGDB/ MG-RAST
Open Ocean	The Sorcerer II Global Ocean Sampling expedition	GS047 Shotgun - Open Ocean - Tropical South Pacific - 201 miles from F. Polynesia - French Polynesia	4441146.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS110b Shotgun - Open Ocean - Indian Ocean - Indian Ocean - International	4441608.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS120 Shotgun - Open Ocean - Indian Ocean - Madagascar Waters - Madagascar	4441135.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS112a Shotgun - Open Ocean - Indian Ocean - Indian Ocean - International	4441609.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS040 Shotgun - Open Ocean - Tropical South Pacific - Tropical South Pacific - International	4441125.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	"GS116 Shotgun - Open Ocean - Indian Ocean - Outside Seychelles, Indian Ocean - Seychelles"	4441149.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS042 Shotgun - Open Ocean - Tropical South Pacific - Tropical South Pacific - International	4441127.3	myMGDB/ MG-RAST

Environment	Project Name	Sample Name	ID	Database
Open Ocean	Global Ocean Sampling Expedition	GS112b Shotgun - Open Ocean - Indian Ocean - Indian Ocean - International	4441147.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS119 Shotgun - Open Ocean - Indian Ocean - International Water Outside of Reunion Island - International	4441568.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS111 Shotgun - Open Ocean - Indian Ocean - Indian Ocean - International	4441156.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS110a Shotgun - Open Ocean - Indian Ocean - Indian Ocean - International	4441607.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS123 Shotgun - Open Ocean - Indian Ocean - International water between Madagascar and South Africa - International	4441616.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS121 Shotgun - Open Ocean - Indian Ocean - International water between Madagascar and South Africa - International	4441614.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS113 Shotgun - Open Ocean - Indian Ocean - Indian Ocean - International	4441610.3	myMGDB/ MG-RAST
Open Ocean	Global Ocean Sampling Expedition	GS114 Shotgun - Open Ocean - Indian Ocean - 500 Miles west of the Seychelles in the Indian Ocean - International	4441611.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Baltic Sea site KBA sample SWE 07_21m (Baltic Sea site KBA sample SWE 07_21m, Oct 2011 Assem)	3300000134	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Tierra del Fuego site OR sample ARG 06_12.3m (Tierra del Fuego site OR sample ARG 06_12.3m, Oct 2011 Assem)	3300000118	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Tierra del Fuego site MC sample ARG 02_11.3m (Tierra del Fuego site MC sample ARG 02_11.3m, Jan 2012 Assem)	3300000131	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Svalbard Archipelago station 1 sample NOR 02_45m (Svalbard Archipelago station 1 sample NOR 02_45m, Jan 2012 Assem)	3300000133	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	King George Island site S1 sample ANT 02_9.5m (King George Island site S1 sample ANT 02_9.5m, Dec 2011 Assem)	3300000136	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Svalbard Archipelago station 1 sample NOR 05_45m (Svalbard Archipelago station 1 sample NOR 05_45m, Nov 2011 Assem)	3300000127	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	King George Island site S2 sample ANT 06_23.45m (King George Island site S2 sample ANT 06_23.45m, Oct 2011 Assem)	3300000123	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Tierra del Fuego site OR sample ARG 05_12.3m (Tierra del Fuego site OR sample ARG 05_12.3m, Oct 2011 Assem)	3300000242	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Tierra del Fuego site MC sample ARG 03_11.3m (Tierra del Fuego site MC sample ARG 03_11.3m, Oct 2011 Assem)	3300000121	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Tierra del Fuego site MC sample ARG 01_11.3m (Tierra del Fuego site MC sample ARG 01_11.3m, Nov 2011 Assem)	3300000125	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Svalbard Archipelago station 1 sample NOR 08_45m (Svalbard Archipelago station 1 sample NOR 08_45m, Dec 2011 Assem)	3300000128	IMG

Environment	Project Name	Sample Name	ID	Database
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	King George Island site S2 sample ANT 04_23.45m (King George Island site S2 sample ANT 04_23.45m, Dec 2011 Assem)	3300000129	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Svalbard Archipelago station 2 sample NOR 15_50m (Svalbard Archipelago station 2 sample NOR 15_50m, Dec 2011 Assem)	3300000130	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	King George Island site S1 sample ANT 03_9.5m (King George Island site S1 sample ANT 03_9.5m, Dec 2011 Assem)	3300000135	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	King George Island site S1 sample ANT 01_9.5m (King George Island site S1 sample ANT 01_9.5m, Oct 2011 Assem)	3300000119	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	King George Island site S2 sample ANT 05_23.45m (King George Island site S2 sample ANT 05_23.45m, Jan 2012 Assem)	3300000132	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Baltic Sea site KBB sample SWE 21_20.5m (Baltic Sea site KBB sample SWE 21_20.5m, Oct 2011 Assem)	3300000241	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Baltic Sea site KBB sample SWE 26_20.5m (Baltic Sea site KBB sample SWE 26_20.5m, Nov 2011 Assem)	3300000126	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Svalbard Archipelago station 2 sample NOR 18_50m (Svalbard Archipelago station 2 sample NOR 18_50m, Dec 2011 Assem)	3300000243	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Svalbard Archipelago station 2 sample NOR 13_50m (Svalbard Archipelago station 2 sample NOR 13_50m, Oct 2011 Assem)	3300000120	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Tierra del Fuego site OR sample ARG 04_12.3m (Tierra del Fuego site OR sample ARG 04_12.3m, Oct 2011 Assem)	3300000122	IMG
Coastal, upwelling, and harbor	Marine microbial communities from chronically polluted sediments in four geographic locations	Baltic Sea site KBA sample SWE 12_21m (Baltic Sea site KBA sample SWE 12_21m, Oct 2011 Assem)	3300000124	IMG
Coastal, upwelling, and harbor	Marine microbial communities from Deepwater Horizon Oil Spill	Marine microbial communities from Deepwater Horizon Oil Spill, sample BP Oil Spill BM58: eDNA_1 (BM58 Illumina assembly)	2088090017	IMG
Coastal, upwelling, and harbor	Marine microbial communities from Delaware Coast	Marine microbial communities from Delaware Coast, sample from Delaware MO Spring March 2010 (Delaware MO Spring March 2010, Nov 2011 assem)	3300000116	IMG
Coastal, upwelling, and harbor	Marine microbial communities from Delaware Coast	Marine microbial communities from Delaware Coast, sample from Delaware MO Summer July 2011 (Delaware MO Summer July 2011, Nov 2011 assem)	3300000115	IMG
Coastal, upwelling, and harbor	Marine microbial communities from Delaware Coast	Late spring/early summer (stable) metatranscriptome	3300000368	IMG
Coastal, upwelling, and harbor	Marine microbial communities from Delaware Coast	Marine microbial communities from Delaware Coast, sample from Delaware MO Winter December 2010 (Delaware MO Winter December 2010, Nov 2011 assem)	3300000117	IMG
Coastal, upwelling, and harbor	Marine microbial communities from Delaware Coast	Marine microbial communities from Delaware Coast, sample from Delaware MO Early Summer May 2010 (Delaware MO Early Summer May 2010, Feb 2012 assem)	3300000101	IMG

Environment	Project Name	Sample Name	ID	Database
Coastal, upwelling, and harbor	Marine microbial communities from Near-Shore Anoxic Basin of Saanich Inlet of Vancouver	Saanich Inlet	2006543006	IMG
Coastal, upwelling, and harbor	Methane oxidizing archaeal communities in the Santa Barbara Basin	Marine sediment archaeal communities from Santa Barbara Basin, CA, that are methane-oxidizing, sample 15-18 cm (ANME Sed A12 15-18 cm)	2140918004	IMG
Coastal, upwelling, and harbor	Methane oxidizing archaeal communities in the Santa Barbara Basin	Marine sediment archaeal communities from Santa Barbara Basin, CA, that are methane-oxidizing, sample 9-12 cm (ANME Sed A12 9-12 cm)	2084038021	IMG
Coastal, upwelling, and harbor	Methane oxidizing archaeal communities in the Santa Barbara Basin	Marine sediment archaeal communities from Santa Barbara Basin, CA, that are methane-oxidizing, sample 3-6 cm (ANME Sed A12 3-6 cm)	2077657019	IMG
Coastal, upwelling, and harbor	Methane oxidizing archaeal communities in the Santa Barbara Basin	Marine sediment archaeal communities from Santa Barbara Basin, CA, that are methane-oxidizing, sample 12-15 cm (ANME Sed A12 12-15 cm)	2140918003	IMG
Coastal, upwelling, and harbor	Methane oxidizing archaeal communities in the Santa Barbara Basin	Marine sediment archaeal communities from Santa Barbara Basin, CA, that are methane-oxidizing, sample 6-9 cm (ANME Sed A12 6-9 cm)	2077657014	IMG
Coastal, upwelling, and harbor	Methane oxidizing archaeal communities in the Santa Barbara Basin	Marine sediment archaeal communities from Santa Barbara Basin, CA, that are methane-oxidizing, sample 0-3 cm (ANME Sed A12 0-3 cm)	2077657018	IMG
Coastal, upwelling, and harbor	Sediment archaeal communities from Eel River Basin	Anaerobic methane oxidation (AOM) community from Eel River Basin sediment, California	2004175001	IMG
Coastal, upwelling, and harbor	Monterey Bay Microbial Study	mb2001jd115_1	4443714.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	GAFFA	Sample 2D	4442589.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	Botany Bay Metagenomic	BBAY01	4443688.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	Monterey Bay Microbial Study	mb2000jd298_1	4443713.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	Monterey Bay Microbial Study	mb2001jd135_2	4443717.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	Monterey Bay Microbial Study	mb2001jd135_1	4443716.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	Monterey Bay Microbial Study	mb2000jd298_2	4443712.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	Botany Bay Metagenomic	BBAY15	4443693.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	Botany Bay Metagenomic	BBAY04	4443691.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	Monterey Bay Microbial Study	mb2001jd115_2	4443715.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	Botany Bay Metagenomic	BBAY02	4443689.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	Global Ocean Sampling Expedition	"GS117b Shotgun - Coastal sample - Indian Ocean - St. Anne Island, Seychelles - Seychelles"	4441148.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	Global Ocean Sampling Expedition	"GS049 Shotgun - Coastal - Polynesia Archipelagos - Moorea, Outside Cooks Bay - Fr. Polynesia"	4441605.3	myMGDB/ MG-RAST

Environment	Project Name	Sample Name	ID	Database
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	GS002 Shotgun - Coastal - North American East Coast - Gulf of Maine - Canada	4441579.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	"GS003 Shotgun - Coastal - North American East Coast - Browns Bank, Gulf of Maine - Canada"	4441580.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	"GS004 Shotgun - Coastal - North American East Coast - Outside Halifax, Nova Scotia - Canada"	4441152.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	GS007 Shotgun - Coastal - North American East Coast - Northern Gulf of Maine - Canada	4441153.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	"GS008 Shotgun - Coastal - North American East Coast - Newport Harbor, RI - USA"	4441583.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	"GS009 Shotgun - Coastal - North American East Coast - Block Island, NY - USA"	4441143.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	"GS010 Shotgun - Coastal - North American East Coast - Cape May, NJ - USA"	4441144.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	"GS013 Shotgun - Coastal - North American East Coast - Off Nags Head, NC - USA"	4441585.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	"GS014 Shotgun - Coastal - North American East Coast - South of Charleston, SC - USA"	4441659.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	"GS015 Shotgun - Coastal - Caribbean Sea - Off Key West, FL - USA"	4441586.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	GS016 Shotgun - Coastal Sea - Caribbean Sea - Gulf of Mexico - USA	4441660.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	GS019 Shotgun - Coastal - Caribbean Sea - Northeast of Colon - Panama	4441589.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	GS021 Shotgun - Coastal - Eastern Tropical Pacific - Gulf of Panama - Panama	4441591.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	"GS027 Shotgun - Coastal - Galapagos Islands - Devil's Crown, Floreana Island - Ecuador"	4441595.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	GS028 Shotgun - Coastal - Galapagos Islands - Coastal Floreana - Ecuador	4441596.4	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	"GS029 Shotgun - Coastal - Galapagos Islands - North James Bay, Santigo Island - Ecuador"	4441596.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	"GS030 Shotgun - Warm Seep - Galapagos Islands - Upwelling, Fernandina Island"	4442626.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	GS034 Shotgun - Coastal - Galapagos Islands - North Seamore Island - Ecuador	4441600.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	GS035 Shotgun - Coastal - Galapagos Islands - Wolf Island - Ecuador	4441601.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	The Sorcerer II Global Ocean Sampling expedition	"GS036 Shotgun - Coastal - Galapagos Islands - Cabo Marshall, Isabella Island - Ecuador"	4441602.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	Global Ocean Sampling Expedition	Sample 32	4442591.4	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	Global Ocean Sampling Expedition	"GS117a Shotgun - Coastal sample - Indian Ocean - St. Anne Island, Seychelles - Seychelles"	4441613.3	myMGDB/ MG-RAST
Coastal, upwelling, and harbor	Global Ocean Sampling Expedition	"GS149 Shotgun - Harbor - Indian Ocean - West coast Zanzibar (Tanzania), harbour region - Tanzania"	4441618.3	myMGDB/ MG-RAST

Environment	Project Name	Sample Name	ID	Database
Estuary	Soil microbial communities from Twitchell Island in the Sacramento Delta	Wetland microbial communities from Twitchell Island in the Sacramento Delta, sample from surface sediment Feb2011 Site B2 Tule (Wetland Surface Sediment Feb2011 Site B2 Tule Oct 2011 assem)	3300000078	IMG
Estuary	Soil microbial communities from Twitchell Island in the Sacramento Delta	Wetland microbial communities from Twitchell Island in the Sacramento Delta, sample from surface sediment Feb2011 Site A2 Cattail (Wetland Surface Sediment Feb2011 Site A2 Cattail Sept 2011 assem)	3300000067	IMG
Estuary	Soil microbial communities from Twitchell Island in the Sacramento Delta	Wetland microbial communities from Twitchell Island in the Sacramento Delta, sample from surface sediment Feb2011 Site B1 Cattail (Wetland Surface Sediment Feb2011 Site B1 Cattail, Assem Ctgs Sep 2011 assem)	3300000313	IMG
Estuary	Soil microbial communities from Twitchell Island in the Sacramento Delta	Wetland microbial communities from Twitchell Island in the Sacramento Delta, sample from surface sediment Feb2011 Site B1 Bulk (Wetland Surface Sediment Feb2011 Site B1 Bulk Feb 2012)	3300000077	IMG
Estuary	Soil microbial communities from Twitchell Island in the Sacramento Delta	Wetland microbial communities from Twitchell Island in the Sacramento Delta, sample from surface sediment Feb2011 Site B2 Bulk (Wetland Surface Sediment Feb2011 Site B2 Bulk, Assem Ctgs Oct 2011 assem)	3300000312	IMG
Estuary	Soil microbial communities from Twitchell Island in the Sacramento Delta	Wetland microbial communities from Twitchell Island in the Sacramento Delta, sample from surface sediment Feb2011 Site A1 Tule (Wetland Surface Sediment Feb2011 Site A1 Tule Jan 2012 assem)	3300000076	IMG
Estuary	Soil microbial communities from Twitchell Island in the Sacramento Delta	CECUM_4-1 (Microbiome Characterization)	3300000317	IMG
Estuary	Soil microbial communities from Twitchell Island in the Sacramento Delta	Wetland microbial communities from Twitchell Island in the Sacramento Delta, sample from surface sediment Feb2011 Site A1 Bulk (Wetland Surface Sediment Feb2011 Site A1 Bulk, Assem Ctgs IBYY 2011 Sep Assem)	3300000309	IMG
Estuary	Soil microbial communities from Twitchell Island in the Sacramento Delta	Wetland microbial communities from Twitchell Island in the Sacramento Delta, sample from surface sediment Feb2011 Site L1 Bulk (Wetland Surface Sediment Feb2011 Site L1 Bulk Jan 2012 assem)	3300000068	IMG
Estuary	Soil microbial communities from Twitchell Island in the Sacramento Delta	Blue Grama Grass Combined Assembly	3300000316	IMG
Estuary	Soil microbial communities from Twitchell Island in the Sacramento Delta	Wetland microbial communities from Twitchell Island in the Sacramento Delta, sample from surface sediment Feb2011 Site L1 Cattail (Wetland Surface Sediment Feb2011 Site L1 Cattail, Assem Ctgs Sep 2011 assem)	3300000318	IMG
Estuary	Sediment microbial communities from Kolumbo Volcano mats	Marine sediment microbial communities from Kolumbo Volcano mats, Greece, sample red mat (Red mat combined assembly)	2088090030	IMG
Estuary	Sediment microbial communities from Kolumbo Volcano mats	Marine sediment microbial communities from Kolumbo Volcano mats, Greece, sample white/grey mat (white/grey mat, combined 454/Illumina assembly)	2084038012	IMG
Estuary	N/A	Microbial Communities from Little Sippewissett Salt Marsh, Woods Hole, MA that are anoxygenic and photosynthetic, Marine photosynthetic community that grows at 940nm (Marine_940nm_cellulose)	3300000504	IMG

Environment	Project Name	Sample Name	ID	Database
Estuary	N/A	Microbial Communities from Little Sippewissett Salt Marsh, Woods Hole, MA that are anoxygenic and photosynthetic, Marine photosynthetic community that grows at 940nm with malate (Marine_940nm_malate)	3300000499	IMG
Estuary	N/A	Microbial Communities from Little Sippewissett Salt Marsh, Woods Hole, MA that are anoxygenic and photosynthetic, Photosynthetic Consortia grown using light of 590nm sample 1 (Marine_590nm_sample1)	3300000470	IMG
Estuary	N/A	Microbial Communities from Little Sippewissett Salt Marsh, Woods Hole, MA that are anoxygenic and photosynthetic, Photosynthetic Consortia grown using light of 590nm sample 2 (Marine_590nm_sample2)	3300000502	IMG
Estuary	N/A	Microbial Communities from Little Sippewissett Salt Marsh, Woods Hole, MA that are anoxygenic and photosynthetic, Photosynthetic Consortia grown using light of 750nm (Marine_750nm)	3300000503	IMG
Estuary	N/A	Microbial Communities from Little Sippewissett Salt Marsh, Woods Hole, MA that are anoxygenic and photosynthetic, sample photosynthetic consortia 740nm (Marine_740nm)	3300000500	IMG
Estuary	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site 1 Maryland Estuary CO2- (Maryland Estuary ambient)	2032320004	IMG
Estuary	Soil microbial communities from FACE and OTC sites	Soil microbial communities from sample at FACE Site 1 Maryland Estuary CO2+ (Maryland Estuary elevated)	2035918006	IMG
Estuary	The Sorcerer II Global Ocean Sampling expedition	"GS005 Shotgun - Embayment - North American East Coast - Bedford Basin, Nova Scotia - Canada"	4441581.3	myMGDB/ MG-RAST
Estuary	Global Ocean Sampling Expedition	"MOVE858 Shotgun - Estuary - North American East Coast - Chesapeake Bay, MD - USA"	4441132.3	myMGDB/ MG-RAST
Estuary	The Sorcerer II Global Ocean Sampling expedition	"GS006 Shotgun - Estuary - North American East Coast - Bay of Fundy, Nova Scotia - Canada"	4441582.3	myMGDB/ MG-RAST
Estuary	The Sorcerer II Global Ocean Sampling expedition	"GS011 Shotgun - Estuary - North American East Coast - Delaware Bay, NJ - USA"	4441658.3	myMGDB/ MG-RAST
Estuary	The Sorcerer II Global Ocean Sampling expedition	"GS012 Shotgun - Estuary - North American East Coast - Chesapeake Bay, MD - USA"	4441584.3	myMGDB/ MG-RAST
High latitude	Marine Bacterioplankton communities from Antarctic	Marine Bacterioplankton communities from Antarctic, Sample 10335 (Summer fosmids)	2040502005	IMG
High latitude	Marine Bacterioplankton communities from Antarctic	Marine Bacterioplankton communities from the Antarctic, sample from Summer (Summer fosmids Sept 2010 assemblies)	2077657013	IMG
High latitude	Marine Bacterioplankton communities from Antarctic	Marine Bacterioplankton communities from Antarctic, sample from Summer (Summer fosmid end sequences)	2008193000	IMG
High latitude	Marine Bacterioplankton communities from Antarctic	Marine Bacterioplankton communities from Antarctic, sample from Winter (Winter fosmid end sequences)	2008193001	IMG
High latitude	Marine Bacterioplankton communities from Antarctic	Marine Bacterioplankton communities from Antarctic, Sample 10334 (Winter fosmids)	2040502004	IMG
High latitude	Marine Bacterioplankton communities from Antarctic	Marine Bacterioplankton communities from the Antarctic, sample from Summer	2264265093	IMG
High latitude	Marine Bacterioplankton communities from Antarctic	Marine Bacterioplankton communities from the Antarctic, sample from Winter (Winter fosmids Sept 2010 assemblies)	2077657020	IMG
High latitude	Marine microbial communities from six Antarctic regions	DNA Fragments from Six Antarctic Marine environments	2012990003	IMG

Environment	Project Name	Sample Name	ID	Database
High latitude	Coastal water and sediment microbial communities from Arctic	Sediment microbial communities from Arctic Ocean, off the coast from Alaska, sample from low methane PC12-247-20cm (Low methane PC12-247-20cm)	2100351001	IMG
High latitude	Coastal water and sediment microbial communities from Arctic	Sediment microbial communities from Arctic Ocean, off the coast from Alaska, sample from medium methane PC12-240-170cm (Medium methane PC12-240-170cm Sept2010 assembly)	2100351011	IMG
High latitude	Coastal water and sediment microbial communities from Arctic	Sediment microbial communities from Arctic Ocean, off the coast from Alaska, sample from high methane PC12-225-485cm (High methane PC12-225-485cm Dec 2010 assembly)	2100351006	IMG
High latitude	Coastal water and sediment microbial communities from Arctic	Sediment microbial communities from Arctic Ocean, off the coast from Alaska, sample from high methane PC12-225-485cm (High methane PC12-225-485cm Jan 2011 assembly)	2140918005	IMG
High latitude	Coastal water and sediment microbial communities from Arctic	Sediment microbial communities from Arctic Ocean, off the coast from Alaska, sample from low methane PC12-244-90cm (Low methane PC12-244-90cm Sept2010 assembly)	2100351012	IMG
High latitude	Coastal water and sediment microbial communities from Arctic	Sediment microbial communities from Arctic Ocean, off the coast from Alaska, sample from high methane PC12-236-260cm (High methane PC12-236-260cm)	2088090012	IMG
Ace Lake	Freshwater microbial communities from Antarctic Deep Lake	Freshwater microbial communities from Antarctic Deep Lake, sample 13m 0.1um (13m 0.1um 454 only)	2100351014	IMG
Ace Lake	Freshwater microbial communities from Antarctic Deep Lake	Freshwater microbial communities from Antarctic Deep Lake, sample 24m 0.1um (24 m 0.1 um 454 only)	2084038011	IMG
Ace Lake	Freshwater microbial communities from Antarctic Deep Lake	Freshwater microbial communities from Antarctic Deep Lake, sample 24m 0.8um (24 m 0.8 um 454/Illumina combined Jan 2011)	2140918017	IMG
Ace Lake	Freshwater microbial communities from Antarctic Deep Lake	Freshwater microbial communities from Antarctic Deep Lake, sample 24m 3.0um (24 m 3.0 um Illumina only)	2061766008	IMG
Ace Lake	Freshwater microbial communities from Antarctic Deep Lake	Freshwater microbial communities from Antarctic Deep Lake, sample 24m 3.0um (24 m 3.0 um Sept 2010 combined)	2081372007	IMG
Ace Lake	Freshwater microbial communities from Antarctic Deep Lake	Freshwater microbial communities from Antarctic Deep Lake, sample 36m 3.0um, 0.8um, 0.1um pool (36m 3, 0.8 and 0.1 um pool)	2100351015	IMG
Ace Lake	Freshwater microbial communities from Antarctic Deep Lake	Freshwater microbial communities from Antarctic Deep Lake, sample 36m 3.0um, 0.8um, 0.1um pool (HWGG+HTSY Jan 2011)	2140918027	IMG
Ace Lake	Freshwater microbial communities from Antarctic Deep Lake	Freshwater microbial communities from Antarctic Deep Lake, sample 5mRS 0.1um (5 mRS 0.1um 454 only)	2084038019	IMG
Ace Lake	Antarctica Aquatic Microbial Metagenome	AntarcticaAquatic_9	4443687.3	myMGDB/ MG-RAST
Ace Lake	Antarctica Aquatic Microbial Metagenome	AntarcticaAquatic_3 - MARINE DERIVED LAKE	4443679.3	myMGDB/ MG-RAST
Ace Lake	Antarctica Aquatic Microbial Metagenome	AntarcticaAquatic_5 - MARINE DERIVED LAKE	4443682.3	myMGDB/ MG-RAST
Ace Lake	Antarctica Aquatic Microbial Metagenome	AntarcticaAquatic_4 - MARINE DERIVED LAKE	4443681.3	myMGDB/ MG-RAST
Ace Lake	Antarctica Aquatic Microbial Metagenome	AntarcticaAquatic_7 - ORGANIC LAKE, ANTARCTICA	4443685.3	myMGDB/ MG-RAST
Ace Lake	Antarctica Aquatic Microbial Metagenome	AntarcticaAquatic_2 - ACE LAKE, ANTARCTICA	4443680.3	myMGDB/ MG-RAST

Environment	Project Name	Sample Name	ID	Database
Ace Lake	Antarctica Aquatic Microbial Metagenome	AntarcticaAquatic_6 - ACE LAKE, ANTARCTICA	4443684.3	myMGDB/ MG-RAST
Ace Lake	Antarctica Aquatic Microbial Metagenome	AntarcticaAquatic_1 - MARINE DERIVED LAKE	4443683.3	myMGDB/ MG-RAST
Ace Lake	Antarctica Aquatic Microbial Metagenome	AntarcticaAquatic_8	4443686.3	myMGDB/ MG-RAST
Deep sea hydrothermal vent	Guaymas Basin hydrothermal plume	Guaymas Basin hydrothermal plume	2061766003	IMG
Deep sea hydrothermal vent	Guaymas Basin hydrothermal plume	GB plume transcript assembly	2236347000	IMG
Deep sea hydrothermal vent	Moore Marine Phage/Virus Metagenomes	Virome EPR hydrothermal vent: Extracellular ssDNA	CAM_SMPL_000719	CAMERA
Deep sea hydrothermal vent	Moore Marine Phage/Virus Metagenomes	Virome EPR hydrothermal vent: Induced RNA virome	CAM_SMPL_000720	CAMERA
Deep sea hydrothermal vent	Moore Marine Phage/Virus Metagenomes	Virome EPR hydrothermal vent: Induced ssDNA virome	CAM_SMPL_000721	CAMERA
Deep sea hydrothermal vent	Moore Marine Phage/Virus Metagenomes	Virome Guaymas hydrothermal vent: Extracellular ssDNA	CAM_SMPL_000804	CAMERA
Deep sea hydrothermal vent	Moore Marine Phage/Virus Metagenomes	Virome Guaymas hydrothermal vent: Induced RNA virome	CAM_SMPL_000809	CAMERA
Deep sea hydrothermal vent	Moore Marine Phage/Virus Metagenomes	Virome Guaymas hydrothermal vent: Extracellular RNA	CAM_SMPL_000834	CAMERA
Deep sea hydrothermal vent	Moore Marine Phage/Virus Metagenomes	Virome Guaymas hydrothermal vent: Induced ssDNA virome	CAM_SMPL_000822	CAMERA
Deep sea hydrothermal vent	Moore Marine Phage/Virus Metagenomes	Virome EPR hydrothermal vent: Extracellular RNA virome	CAM_SMPL_000718	CAMERA
Deep sea hydrothermal vent	Epibiont Metagenome	EPIBIONTMETAGENOME_SMPL_EPIBIONT	CAM_PROJ_EpibiontMetagenome	CAMERA
Deep sea hydrothermal vent	Alvinella pompejana Epibiont Metagenome	ALVINELLA_SMPL_20041130	CAM_PROJ_AlvinellaPompejana	CAMERA
Deep sea hydrothermal vent	Hydrothermal Vent Metagenome	HYDROTHERMALVENT_SMPL_HCM	CAM_PROJ_HydrothermalVent	CAMERA
Reef	Global Ocean Sampling Expedition	"GS048b Shotgun - Coral Reef - Polynesia Archipelagos - Moorea, Cooks Bay - Fr. Polynesia"	4441167.3	myMGDB/ MG-RAST
Reef	Global Ocean Sampling Expedition	GS050 Shotgun - Coral Atoll - Polynesia Archipelagos - Tikehau Lagoon - Fr. Polynesia	4441121.3	myMGDB/ MG-RAST
Reef	The Sorcerer II Global Ocean Sampling expedition	"GS025 Shotgun - Fringing Reef - Eastern Tropical Pacific - Dirty Rock, Cocos Island - Costa Rica"	4441593.3	myMGDB/ MG-RAST
Reef	The Sorcerer II Global Ocean Sampling expedition	GS051 Shotgun - Coral Reef Atoll - Polynesia Archipelagos - Rangirora Atoll - Fr. Polynesia	4441604.3	myMGDB/ MG-RAST
Reef	Global Ocean Sampling Expedition	"GS148 Shotgun - Fringing Reef - Indian Ocean - East coast Zanzibar (Tanzania), offshore Paje lagoon - Tanzania"	4441617.3	myMGDB/ MG-RAST
Reef	Global Ocean Sampling Expedition	"GS108b Shotgun - Lagoon Reef - Indian Ocean - Coccos Keeling, Inside Lagoon - Australia"	4441133.3	myMGDB/ MG-RAST
Reef	Global Ocean Sampling Expedition	"GS048a Shotgun - Coral Reef - Polynesia Archipelagos - Moorea, Cooks Bay - Fr. Polynesia"	4441603.3	myMGDB/ MG-RAST
Reef	Global Ocean Sampling Expedition	"GS108a Shotgun - Lagoon Reef - Indian Ocean - Coccos Keeling, Inside Lagoon - Australia"	4441139.3	myMGDB/ MG-RAST
Mangrove	The Sorcerer II Global Ocean Sampling expedition	GS032 Shotgun - Mangrove - Galapagos Islands - Mangrove on Isabella Island - Ecuador	4441598.3	myMGDB/ MG-RAST

Environment	Project Name	Sample Name	ID	Database
Hypersaline	Solar Saltern	MedSalternSDbayMic20051110	4440435.3	myMGDB/ MG-RAST
Hypersaline	Solar Saltern	SaltonSeaMic20060823	4440329.3	myMGDB/ MG-RAST
Hypersaline	Solar Saltern	LowSalternSDbayMic20051110	4440324.3	myMGDB/ MG-RAST
Hypersaline	Solar Saltern	LowSalternSDbayMic20051128	4440426.3	myMGDB/ MG-RAST
Hypersaline	Solar Saltern	LowSalternSDbayMic200407	4440437.3	myMGDB/ MG-RAST
Hypersaline	Solar Saltern	HighSalternSDbayMicC200407	4440433.3	myMGDB/ MG-RAST
Hypersaline	Solar Saltern	HighSalternSDbayMic20051128	4440419.3	myMGDB/ MG-RAST
Hypersaline	Solar Saltern	MedSalterSDbayMic20051128	4440416.3	myMGDB/ MG-RAST
Hypersaline	Solar Saltern	MedSalternSDbayMic20051116	4440425.3	myMGDB/ MG-RAST
Hypersaline	Solar Saltern	HighSalternSDbayMicB200407	4440429.3	myMGDB/ MG-RAST
Hypersaline	Solar Saltern	MedSalternSDbayMic20051111	4440434.3	myMGDB/ MG-RAST
Hypersaline	Solar Saltern	HighSalternSDbayMicA200407	4440430.3	myMGDB/ MG-RAST
Hypersaline	Solar Saltern	HighSalternSDbayMicD200407	4440438.3	myMGDB/ MG-RAST
Hypersaline	Marine NaCl-Saturated Brine	Marine NaCl-Saturated Brine	4441050.3	myMGDB/ MG-RAST
Hypersaline	Hypersaline Guerro Negro	Guerrero Negro 4-5mm	4440967.3	myMGDB/ MG-RAST
Hypersaline	Hypersaline Guerro Negro	Guerrero Negro 5-6mm	4440969.3	myMGDB/ MG-RAST
Hypersaline	Hypersaline Guerro Negro	Guerrero Negro 0-1mm	4440964.3	myMGDB/ MG-RAST
Hypersaline	Hypersaline Guerro Negro	Guerrero Negro 6-10mm	4440970.3	myMGDB/ MG-RAST
Hypersaline	Hypersaline Guerro Negro	Guerrero Negro 2-3mm	4440965.3	myMGDB/ MG-RAST
Hypersaline	Hypersaline Guerro Negro	Guerrero Negro 34-49mm	4440972.3	myMGDB/ MG-RAST
Hypersaline	Hypersaline Guerro Negro	Guerrero Negro 22-34mm	4440971.3	myMGDB/ MG-RAST
Hypersaline	Hypersaline Guerro Negro	Guerrero Negro 10-22mm	4440968.3	myMGDB/ MG-RAST
Hypersaline	Hypersaline Guerro Negro	Guerrero Negro 3-4mm	4440966.3	myMGDB/ MG-RAST
Hypersaline	Hypersaline Guerro Negro	Guerrero Negro 1-2mm	4440963.3	myMGDB/ MG-RAST
Hypersaline	The Sorcerer II Global Ocean Sampling expedition	"GS033 Shotgun - Hypersaline - Galapagos Islands - Punta Cormorant, Hypersaline Lagoon, Floreana Island - Ecuador"	4441599.3	myMGDB/ MG-RAST
Equitorial upwelling	Marine Bacterioplankton Metagenomes	S_35155 - Pacific Equatorial Divergence Province	4443697.3	myMGDB/ MG-RAST
Equitorial upwelling	Marine Bacterioplankton Metagenomes	S_35163 - Pacific North Equatorial	4443699.3	myMGDB/ MG-RAST
Equitorial upwelling	Marine Bacterioplankton Metagenomes	S_35162 - Pacific North Equatorial	4443698.3	myMGDB/ MG-RAST
Equitorial upwelling	Marine Bacterioplankton Metagenomes	S_35171 - Pacific North Equatorial Countercurrent	4443700.3	myMGDB/ MG-RAST
Spring bloom	Metagenomic Analysis of the North Atlantic Spring Bloom	174-1	4443725.3	myMGDB/ MG-RAST
Spring bloom	Metagenomic Analysis of the North Atlantic Spring Bloom	179-2	4443732.3	myMGDB/ MG-RAST
Spring bloom	Metagenomic Analysis of the North Atlantic Spring Bloom	174-2	4443729.3	myMGDB/ MG-RAST
Spring bloom	Metagenomic Analysis of the North Atlantic Spring Bloom	179-1	4443731.3	myMGDB/ MG-RAST

Environment	Project Name	Sample Name	ID	Database
Trichodesmium bloom	Marine Trichodesmium cyanobacterial communities from the Bermuda Atlantic Time-Series	Marine Trichodesmium cyanobacterial communities from the Bermuda Atlantic Time-Series	2156126005	IMG
Trichodesmium bloom	N/A	Marine Trichodesmium cyanobacterial communities from the North Pacific Subtropical Gyre outside Oahu, HI, sample from new species B colonies	2264265224	IMG

Appendix B

DATASET FROM CHAPTER III

Supplemental Table 1 | List of sequences used in study

Dataset	GI	Species Name
Outgroup 1	320160959	*Anaerolinea thermophila* UNI-1
	39995655	*Geobacter sulfurreducens* PCA
	57233930	*Dehalococcoides ethenogenes* 195
	523470333	*Desulfovibrio* sp. X2
	193212385	*Chlorobaculum parvum* NCIB 8327
Outgroup 2	645069916	*Opitutaceae bacterium* TAV5
	501344729	*Opitutus terrae*
	496472502	*Desulfovibrio* sp. FW1012B
	501524087	*Geobacter bemidjiensis*
	506253585	*Desulfomicrobium baculatum*
	647376358	Dehalococcoidia bacterium DscP2
	504856057	*Dehalobacter* sp. DCA
	654862234	*Desulfatibacillum aliphaticivorans*
	655124705	*Desulfonatronum lacustre*
	501443562	*Chlorobium limicola*
HpnP ingroup	492877294	*Afipia broomeae* ATCC 49717
	488798710	*Afipia clevelandensis* ATCC 49720
	488803967	*Afipia felis* ATCC 53690
	639244164	*Afipia* sp. (639244164)
	640480562	*Afipia* sp. (640480562)
	496697395	*Afipia* sp. 1NLS2
	571918263	*Afipia* sp. P52-10
	501352804	*Beijerinckia indica* subsp. indica ATCC 9039
	497421586	Bradyrhizobiaceae bacterium SG-6C
	499398312	*Bradyrhizobium diazoefficiens* USDA 110
	517081761	*Bradyrhizobium elkanii* (517081761)
	654709199	*Bradyrhizobium elkanii* (654709199)
	654879908	*Bradyrhizobium elkanii* (654879908)
	654886985	*Bradyrhizobium elkanii* (654886985)
	654899386	*Bradyrhizobium elkanii* (54899386)
	504309727	*Bradyrhizobium japonicum* USDA 6
	636813563	*Bradyrhizobium japonicum* (636813563)
	648621071	*Bradyrhizobium japonicum* (648621071)
	654676491	*Bradyrhizobium japonicum* (654676491)
	654684416	*Bradyrhizobium japonicum* (654684416)
	654688277	*Bradyrhizobium japonicum* (654688277)
	654699060	*Bradyrhizobium japonicum* (654699060)
	654711923	*Bradyrhizobium japonicum* (654711923)
	654714597	*Bradyrhizobium japonicum* (654714597)

Dataset	GI	Species Name
	654723587	*Bradyrhizobium japonicum* (654723587)
	505481583	*Bradyrhizobium oligotrophicum* S58
	653496802	*Bradyrhizobium* sp. (653496802)
	653552068	*Bradyrhizobium* sp. Ai1a-2
	639168996	*Bradyrhizobium* sp. ARR65
	500989536	*Bradyrhizobium* sp. BTAi1
	404267169	*Bradyrhizobium* sp. CCGE-LA001
	653477533	*Bradyrhizobium* sp. Cp5.3
	544645606	*Bradyrhizobium* sp. DFCI-1
	608613371	*Bradyrhizobium* sp. DOA9
	640606658	*Bradyrhizobium* sp. DOA9
	653423211	*Bradyrhizobium* sp. Ec3.3
	639269491	*Bradyrhizobium* sp. OHSU_III
	500305918	*Bradyrhizobium* sp. ORS 278
	493662555	*Bradyrhizobium* sp. ORS 285
	496317759	*Bradyrhizobium* sp. ORS 375
	505725203	*Bradyrhizobium* sp. S23321
	496249482	*Bradyrhizobium* sp. STM 3809
	496253242	*Bradyrhizobium* sp. STM 3843
	656043459	*Bradyrhizobium* sp. th.b2
	639177735	*Bradyrhizobium* sp. Tv2a-2
	653447364	*Bradyrhizobium* sp. URHA0002
	653520650	*Bradyrhizobium* sp. URHA0013
	657881192	*Bradyrhizobium* sp. URHD0069
	494867534	*Bradyrhizobium* sp. WSM1253
	653392484	*Bradyrhizobium* sp. WSM1417
	653528340	*Bradyrhizobium* sp. WSM1743
	653462358	*Bradyrhizobium* sp. WSM2254
	648509787	*Bradyrhizobium* sp. WSM2793
	653436691	*Bradyrhizobium* sp. WSM3983
	648541515	*Bradyrhizobium* sp. WSM4349
	494882022	*Bradyrhizobium* sp. WSM471
	495419321	*Bradyrhizobium* sp. YR681
	518325440	*Calothrix* sp. PCC 7103
	499842390	Candidatus *Koribacter versatilis* Ellin345
	515386899	*Chlorogloeopsis fritschii*
	505032897	Cyanobacterium aponinum PCC 10605
	506435442	*Cyanothece* sp. PCC 7424
	501729632	*Cyanothece* sp. PCC 7425
	503089159	*Cyanothece* sp. PCC 7822
	517207725	filamentous cyanobacterium ESFC-1
	515365955	*Fischerella muscicola*
	652337927	*Fischerella* sp. PCC 9605
	515865344	*Geminocystis herdmanii*
	554673747	*Gloeobacter kilaueensis* JS1
	499454847	*Gloeobacter violaceus* PCC 7421
	493573935	*Gloeocapsa* sp. PCC 73106

Dataset	GI	Species Name
	648457296	*Leptolyngbya boryana*
	557879987	*Leptolyngbya boryana* CCAP 1462/2
	648365781	*Mastigocladopsis repens*
	489693355	*Methylobacterium extorquens* PA1
	506304389	*Methylobacterium extorquens* CM4
	498372741	*Methylobacterium mesophilicum* SR1.6/6
	506408858	*Methylobacterium nodulans* ORS 2060
	501431528	*Methylobacterium populi* BJ001
	501277757	*Methylobacterium radiotolerans* JCM 2831
	501287335	*Methylobacterium* sp. 4-46
	652919088	*Methylobacterium* sp. 10
	518940350	*Methylobacterium* sp. 285MFTsu5.1
	651602142	*Methylobacterium* sp. 77
	516687765	*Methylobacterium* sp. 88A
	494839951	*Methylobacterium* sp. GXF4
	516053564	*Methylobacterium* sp. MB200
	651611548	*Methylocapsa acidiphila*
	501586616	*Methylocella silvestris* BL2
	648235176	*Methylocystis parvus*
	519020310	*Methyloferula stellata*
	504998593	*Microcoleus* sp. PCC 7113
	499830400	*Nitrobacter hamburgensis* X14
	497485859	*Nitrobacter* sp. Nb-311A
	499634763	*Nitrobacter winogradskyi* Nb-255
	501377424	*Nostoc punctiforme* PCC 73102
	504925828	*Nostoc* sp. PCC 7107
	501559316	*Oligotropha carboxidovorans* OM5
	497453748	*Oscillatoriales cyanobacterium* JSC-12
	516314663	*Prochlorothrix hollandica*
	499472646	*Rhodopseudomonas palustris* CGA009
	499759880	*Rhodopseudomonas palustris* HaA2
	499791441	*Rhodopseudomonas palustris* BisB18
	499823235	*Rhodopseudomonas palustris* BisB5
	499982521	*Rhodopseudomonas palustris* BisA53
	501488665	*Rhodopseudomonas palustris* TIE-1
	503266706	*Rhodopseudomonas palustris* DX-1
	653026104	*Rhodopseudomonas palustris*
	653046421	Rhodospirillales bacterium URHD0088
	495664219	*Rhodovulum* sp. PH10
	648443022	*Scytonema hofmanni*
	657933118	*Scytonema hofmanni* UTEX 2349
	504936600	*Synechococcus* sp. PCC 6312
	499369037	*Thermosynechococcus elongatus* BP-1
	571030898	*Thermosynechococcus* sp. NK55a