ECOLOGICAL AND EVOLUTIONARY STRATEGIES OF ARCHAEAL, BACTERIAL, AND VIRAL COMMUNITIES IN DEEP-SEA HYDROTHERMAL VENTS

RIKA ELIZABETH ANDERSON

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

University of Washington
2013

Reading Committee:
John Baross, Chair
Jody Deming
Robert Morris

Program Authorized to Offer Degree:
School of Oceanography

University of Washington

Abstract

Ecological and evolutionary strategies of archaeal, bacterial, and viral communities in deep-sea hydrothermal vents

Rika Elizabeth Anderson

Chair of the Supervisory Committee:
Professor John A. Baross
School of Oceanography

The deep-sea hydrothermal vent habitat, formed by subsurface water-rock reactions that create high-temperature hydrothermal fluid, is dominated by physical, chemical, and mineralogical gradients. The mixing of cold, oxidized seawater with hot, reduced hydrothermal fluid produces environments that span a range of temperatures, pH, redox potential, chemical composition, and mineralogy, with constant fluid flux between these regions. Communities of archaea, bacteria, and viruses live across the gradients within these systems and are both exposed to and transported by these fluids. Since these conditions can push the boundaries of the limits for life, may represent conditions found on other planetary bodies, and are thought to have been important for the early evolution of life on this planet, the study of microbial adaptation to hydrothermal vents is of great astrobiological importance. This dissertation explores how these extreme gradients structure hydrothermal vent microbial and viral communities, and what evolutionary strategies are used by both cells and viruses in hydrothermal systems to adapt to these extremes.

The first part of this dissertation address adaptation on the community level by examining microbial community structuring in various niches within the vent environment. First, I explore microbial niche partitioning across diffuse flow and plumes in hydrothermal vent systems, using a combination of microbial community profiling techniques and qPCR to demonstrate that certain microbial lineages are found in high abundance in particular conditions, but are far less abundant in other regions of the gradient. Second, I use 16S pyrotag sequencing to compare the structures of the rare and abundant biospheres across several hydrothermal vent systems worldwide. Through this I demonstrate that archaeal communities exhibit fundamentally different biogeographic patterning compared to bacterial communities. Whereas bacterial rare and abundant groups show similar biogeographic patterning, abundant archaeal groups are generally cosmopolitan and abundant everywhere but rare archaeal groups are biogeographically restricted.

The second part of my dissertation focuses on adaptive strategies among viruses and their microbial hosts. I first demonstrate a novel method by which to identify potential hosts of a viral assemblage using metagenomics, showing that viruses in the vent system have the potential to infect a wide range of hosts. Finally, I use comparative metagenomics to demonstrate that the viral fraction in a high-temperature hydrothermal system is relatively enriched in energy-metabolizing genes, and present evidence suggesting that these genes are transferred by viruses as an adaptive strategy to enhance host metabolic plasticity in a dynamic environment. Taken together, this work indicates that the gradient-dominated nature of vent systems fosters a diverse microbial community through adaptation to particular niches, and that virally-mediated transfer of genes between these diverse hosts creates genomic plasticity to facilitate adaptation to the vent environment. In this sense niche partitioning drives these microbial lineages apart, while horizontal gene transfer allows them to borrow adaptive strategies from each other.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The twelfth century philosopher Bernard of Chartres said that we are like dwarves perched on the shoulders of giants—we may see more and farther than the giants, but it is not because we ourselves are strong or perceptive, but because we are carried aloft by the giants themselves. This dissertation is no exception to this rule. I am aware that this quote has been used before, and in fact it seems downright presumptive for me, with this small contribution to the general body of scientific knowledge, to use the same quote that Isaac Newton appropriated in describing his own work. Nevertheless, the sentiment is the same.

The first and foremost giant upon whose shoulders I have stood is John Baross, a giant in the world of microbiology and also quite a large man when standing in your office doorway. Frankly, he terrified me the first time I met him. But in John I have found the best advisor I could possibly have asked for. John's big-picture thinking and his crazy ideas, combined with his willingness to let me explore and make the mistakes I needed to make while doing so, have helped me focus on fundamental concepts without fretting over insubstantial details. John isn't afraid to ask big questions: How did life originate? Is there life elsewhere? How might an evolutionary perspective help us cure cancer? Then, as if to counter these contemplations, he will wander into my office to talk about the pasta he cooked last night, or the tomatoes in his garden, or his favorite opera, with occasional forays into the relationship between science, philosophy, and religion or the origin of life. John's enduring humility in the face of big, unanswered questions combined with his humor and knack for storytelling make him one of the men I most admire, as well as a real friend. He may claim to possess only three synaptic cells, but I know better. I feel indescribably fortunate to have spent six or so years absorbing John's crazy genius. I can only hope some of it has sunk in.

I have also learned much from my committee: Jody Deming, Victoria Meadows, David Butterfield, and Robert Morris. Jody has provided unique insights into my work and asks tough questions that have made me think harder about the nature of my samples and the interpretation of my data. Vikki, from her astrobiological perspective, helps me put my work in context and has introduced me to many of the leaders in the world of astrobiology. Dave's knowledge of the chemistry and geology of hydrothermal systems has been invaluable, as has been his assistance with sample collection at sea. And I've enjoyed and learned a lot from my discussions of both sulfur-oxidizing *Gammaproteobacteria* as well as metagenomics methods with Bob, and I hope to continue those discussions in the future.

The person from whom I've learned the most hands-on methods, both computationally and in the lab, is William Brazelton. Thank you, Billy, for letting me invade your office almost hourly for the first couple years I was here, and as the years went on, for acting as a sounding board for new ideas as I worked through my analyses. I couldn't have done any of this without your insights (and patience).

As a budding scientist, I had tremendous support and encouragement from various mentors before I even entered graduate school. Randy Villahermosa, Kai-Uwe Hinrichs, David Smith, Anthony Poole and Euan Nisbet kindly (or perhaps crazily) took me into their labs to learn the tools of the trade despite my utter lack of experience in their fields. Mark McKone and Phil Camill, both at Carleton College, provided mentorship early on. They all set me on the path that has led me here.

I have also received tremendous amounts of shipboard help while collecting these samples, especially the giant barrel sample which played into every single chapter of this dissertation. Paul Johnson, my office neighbor, gave me insights based on his experiences when he first deployed one himself (and has been a wonderful neighbor for these past six years). Tor Bjorklund's genius was behind the design and construction of the barrel sampler itself. I feel fortunate to be part of the Baross family, which has provided a great deal of support as well—Jim Holden was Chief Scientist of the cruises on which I obtained my samples, Jon Kaye assisted with financial support through the Moore Foundation, and Julie Huber has helped tremendously with sampling and with general mentorship. Thank you also to the captain and crews of the *R/V Atlantis* and *DSV Alvin* for their logistical support.

One of the most important parts of the graduate school experience is to cultivate a community for mutual support, learning, and commiseration. The astrobiology community has been a wonderful source of entertainment and encouragement, and gave me the opportunity to learn about concepts ranging from sulfur isotopes to atmospheric temperature structure. The oceanography community has also been wonderful to work and play with, and I've enjoyed First Fridays and Brewery Bike Tours together, and playing soccer with a bunch of oceanographers with the FC Flowbees. And of course, the ocean admin staff has been wonderful and patient for the past six years, especially Su, Chanthavy, and April; and Kathy Newell has been great fun to teach with the past couple years.

Financially, I have been supported by the National Science Foundation through a Graduate Research Fellowship, the IGERT Fellowship through the UW Astrobiology Program, and the NASA Astrobiology Institute through the Carnegie and VPL nodes. I also received an ARCS Fellowship, generously supported by Jill and William Ruckleshaus, and the ARCS Foundation has been a wonderful source of encouragement throughout my entire graduate school experience.

Finally, the most crucial source of moral support has been my friends and family. My parents are an everlasting source of support, always sending me cards or care packages. They were the ones I turned to when things got rough, and they were always there for me, urging me to think positive. They put me through college so that I could get to this point. I have always been acutely aware of how fortunate I have been to have such strong moral and financial support from my parents, and for that I am exceedingly grateful. One of the strongest motivating forces in my life has been to make you proud, and I hope I am achieving that, bit by bit. It's also been wonderful to have my sister in the Pacific Northwest as we've treaded the halls of academia in parallel, and she and I are now both moving on to new phases in life. My great uncle Ray, who is 94 this year, has always challenged me to think creatively and outside the box, and that it's never too late to learn something new. To my friends, from high school, Carleton, UW, and all the places in between: you know who you are, and all I can say is thank you. And finally, words simply cannot describe how important Tony's constant, forgiving, and generous support has been to me for the past decade. He is one of the most genuine people I have ever known, and inspires me with his excitement every time he finds something new in his research. I don't know where I would be without you by my side each day, and I look forward to seeing what our future holds.

INTRODUCTION[1]

It is often stated that "nothing in biology makes sense except in the light of evolution" (Dobzhansky, 1964). A corollary to this statement, however, is that nothing in evolution makes sense except in light of its environment. Evolution does not take place in a vacuum. Interactions with the physical environment and with the organisms within it are the primary determinants of a given species' evolutionary trajectory, and the multiplicity of available ecological niches is responsible for the vast diversity of life we observe on Earth.

Deep-sea hydrothermal vents, where superheated water created by subterranean water-rock interactions meets the frigid bottom waters of our oceans, are host to a diverse array of ecological niches. These are created by strong gradients in temperature, pH, fluid chemistry, reduction potential, and mineralogy, which are produced by the mixing of high-temperature hydrothermal fluid with cold seawater. The multiplicity of environmental conditions begets a high diversity of microorganisms, each adapted to a different set of optimal growing conditions. Moreover, the environmental conditions found in hydrothermal systems include those that we might label "extreme," which for our purposes will be defined as those environmental conditions that approach the known limits for life, particularly for temperature and pressure. Therefore, hydrothermal vents provide an ideal observatory from which to study adaptations to these extreme conditions, as well as the evolutionary processes that have produced the diversity of life on this planet. Finally, hydrothermal vents are thought to represent the most ancient continuously inhabited ecosystems on the planet, and were abundant on the Hadean Earth when life was first gaining a foothold on the planet. Thus, hydrothermal vents may provide a window into the earliest evolutionary steps for life on Earth.

This dissertation represents an attempt to unravel some of the evolutionary processes that occur within the context of hydrothermal vent systems, through the lens of microbial ecology. One of the most important concepts of evolutionary theory is adaptation: what traits or strategies do lineages develop through natural selection in order

[1] Previously published, in slightly modified form, as "The Deep Viriosphere: Assessing the Viral Impact on Microbial Community Dynamics in the Deep Subsurface" in Reviews in Mineralogy and Geochemistry, Vol. 75, p. 649-675, 2013.

to have greater fitness in a given ecological context? That ecological context includes both the physical environment and the biological community. Therefore, in the first half of this dissertation I examine microbial adaptations to the physical environment of hydrothermal vents. The large-scale questions I sought to investigate include:

- *How do archaea and bacteria partition across fluid gradients in hydrothermal systems?*
- *Which microbial lineages tend to be rare, and which are abundant?*
- *Are the rare lineages always rare, or do they bloom in different regions of the system?*

I begin with an exploration of how archaea and bacteria have spread into the many ecological niches available within the vent environment. I do this first by contrasting microbial community structure in diffuse flow hydrothermal fluids and in cooler hydrothermal plumes using TRFLP, clone libraries and qPCR (Chapter 1). Then, I use next-generation sequencing technology—pyrotag sequencing of the v4v6 region of the 16S rRNA gene—to probe more deeply into patterns of microbial community structure by examining niche partitioning and biogeography of the rare biosphere, which cannot be targeted by using more traditional techniques (Chapter 2).

In the second half I focus on microbial adaptations to interactions with the biological community, with a particular focus on viruses. Until now, the role of viruses in influencing the ecology and evolution of deep-sea hydrothermal vent microbial communities has been largely unexplored. Primary questions include:

- *Do viruses infect a wide diversity of archaea and bacteria in vent systems, or only certain types of organisms?*
- *To what degree do viruses mediate gene transfer between hosts?*
- *Do viruses help their hosts adapt to the environment by carrying or expressing genes that enhance their fitness?*

I first attempt to establish the extent to which viruses influence the microbial community at vents by determining the host range of the local viral assemblage (Chapter 3). I do this by developing a new method that takes advantage of the natural library found in CRISPRs (Clustered Regularly Interspaced Palindromic Repeats) found on microbial genomes, and compare CRISPR libraries from genomes of locally isolated organisms with a viral metagenome from a hydrothermal vent. Finally, I use comparative metagenomics of a viral and a cellular metagenome to examine how viruses manipulate the genomic landscape of the bacteria and archaea they infect, and in doing so, influence their evolutionary trajectory (Chapter 4).

The remainder of this introduction will focus firstly on characterizing the subsurface habitat that will be the focus of this dissertation, and will briefly touch on concepts related to microbial structuring in these regions. Having established this important ecological context, I will then move into an introduction to the viral world, touching upon their most important impacts upon microbial communities.

THE SUBSURFACE BIOSPHERE

Much of the ocean crust experiences fluid flux to a certain degree; it is estimated that at least 60% of the ocean crust is hydrologically active (Edwards *et al.* 2011). The volume of fluid fluxing through the crust is at its highest near active hydrothermal systems at mid-ocean ridges, but does not immediately dissipate. The degree of fluid flux varies depending on the sediment cover, as sediments tend to restrict fluid flow (Edwards *et al.* 2005). This is illustrated schematically in Figure 0.1. Most fluid flux occurs through connected channels in ocean crust, such as around breccia zones and around pillow basalts or flow boundaries (Fisher and Becker, 2000). Seawater flows through seamounts, ridge flanks and recharge zones away from the ridge axis, with residence times ranging from days to years, depending on the location (Johnson and Pruis 2003). Thus a substantial portion of the ocean subsurface biosphere is exposed to dynamic fluid flux, and hydrothermal vents tend to have the highest degree of fluid flux.

Hydrothermal vent systems

Hydrothermal vent systems are found at mid-ocean ridge spreading centers, ridge flanks, or seamounts. Hydrothermal systems can be broadly classified into magma-driven or basalt-hosted systems and serpentinization-driven or peridotite-hosted systems, though some hydrothermal fields have been found that exhibit characteristics of both these types of systems, such as Rainbow Field. Lost City is currently the best-known example of a serpentinization-driven, peridotite-hosted hydrothermal system, but the remainder of this dissertation will focus on microbial processes at basalt-hosted systems.

In magma-driven systems, seawater comes into close contact with the cracking front of a magma chamber, resulting in high-temperature water rock reactions. The resulting buoyant hydrothermal fluid rises through fissures in the porous crust, and when it emerges at the seafloor it comes into contact with the colder, more alkaline seawater, precipitating iron sulfide to create large sulfide towers. These basalt-hosted systems are characterized by high-temperature, low-pH fluids that are enriched in reduced compounds, transition metals, sulfide, CO_2, helium, methane, and hydrogen, and are depleted in magnesium (Von Damm, 1990). The sulfide chimneys play host to complex microbial communities, which in turn form the trophic basis of macrofaunal communities hosting worms, mussels, crabs, and shrimp. One of the defining characteristics of these systems is the dominance of chemical, physical, and mineralogical gradients that shape the structure of the microbial communities inhabiting these systems (Baross & Hoffman 1985; Schrenk *et al.* 2003). As seawater mixes with hydrothermal fluid, this results in temperatures that range from 2˚ to 400˚C, acidities that range from 2 to 8, and chemical composition that spans the spectrum of reduced hydrothermal fluid to oxidized seawater. As a result, vents play host to a wide range of archaeal and bacterial species: hyperthermophiles, thermophiles and psychrophiles, heterotrophs and autotrophs. These organisms partition themselves along the physiochemical gradients present in the vent system. In Chapter 1, I investigate the ways in which the structure of archaeal and bacterial communities changes across these gradients, ranging from deep seawater to hydrothermal plumes to diffuse flow fluids.

Diversity in hydrothermal systems

As a result of the multitude of ecological niches available in hydrothermal systems, these regions host tremendous microbiological diversity. Sulfide structures are dominated by thermophilic and hyperthermophilic microbial communities, particularly Crenarchaea and Archaeoglobaceae (Kelley *et al.*, 2002; Schrenk *et al.*, 2003; Slobodkin *et al.*, 2001; Takai *et al.*, 2001; Takai & Horikoshi, 1999); a study by Schrenk *et al.* (2003) observed an abundance of uncultured Crenarchaea in the center of vent sulfide structures. Cooler diffuse flow fluids, characterized by a mixture of high-temperature hydrothermal fluid from the subsurface and background seawater, tend to be bacterially-dominated, particularly by Gamma- and Epsilonproteobacteria (Huber *et al.*, 2003, 2007; Deming and Baross, 1993), though archaea are found in these fluids as well, including Marine Group I Crenarchaea, as well as thermophilic Thermoprotei, Thermococcales and Methanococcales groups (Huber *et al.*, 2002). Though diffuse flow fluids are often sampled at around 5-20˚C, culturable thermophiles are often isolated from these fluids (Summit and Baross, 2001; Holden *et al.*, 1998), and therefore these fluids are thought to represent "windows" to a deep, hot subsurface biosphere microbial community (Deming and Baross, 1993).

Deep sequencing of microbial communities in diffuse flow fluids found almost 20,000 different bacterial operational taxonomic units (OTUs) and nearly 2,000 archaeal OTUs (Huber *et al.*, 2007). Rarefaction analysis indicated that saturation had not been reached for the bacterial sequences, indicating that the total diversity of these systems had not yet been sampled. Moreover, deep sequencing of these samples revealed that thousands of low-abundance populations dominated these samples, for which the term "rare biosphere" was coined (Sogin *et al.*, 2006). It has been suggested that the rare biosphere acts as a seed bank for dormant cells that become more abundant in different conditions (Jones and Lennon, 2010; Lennon and Jones, 2011; Gibbons *et al.*, 2013), or that it acts as a bank of "genetic memory" from past events that molded the microbial community (Brazelton *et al.*, 2010). However, much remains unknown regarding the ecological role of the rare biosphere: are these lineages always rare, or do they "bloom" when conditions are favorable? What is their biogeographic distribution? What impact do they have on local biogeochemical cycles, and what is the nature of their interaction with the microbial community? In Chapter 2, I explore the biogeography and ecology of the rare and abundant biosphere of both the bacterial and archaeal domains in hydrothermal systems across the globe.

One of the most important factors governing microbial community structure and diversity is viral infection. Since viral infection is dependent upon host density, viral populations may be responsible for maintaining even diversity by killing the more abundant populations (Thingstad & Lignell, 1997). Viruses may also be responsible for the existence of rare lineages: some rare populations are actively growing, but are particularly susceptible to viral lysis (Pedrós-Alió, 2012; Bouvier and del Giorgio, 2007). In addition to their impacts on microbial diversity, viruses can also manipulate biogeochemical cycles as well as the evolutionary trajectories of their microbial hosts. Our focus will now turn to the viruses.

THE VIRAL WORLD

All regions of Earth's biosphere that we have studied—the waters of Earth's oceans, the soil beneath our feet, and even the air we breathe—teem with viruses. Viral particles are among the smallest biological entities on the planet, with the average viral particle measuring about 100 nm in length: a size so small that five thousand viruses, lined end to end, would fit across the thickness of a human fingernail. What they lack in size, though, they compensate with sheer abundance. If we were to line up all the viruses in the ocean, they would stretch across the diameter of the Milky Way galaxy one hundred times (Suttle 2007). Those viruses are responsible for up to 10^{23} infections per second in the oceans (Suttle 2007). With each new infection, viruses can have a profound impact on their hosts: they can alter the structure of a microbial population, break up cellular biomass into its constituent organic matter, or introduce new genes into their hosts. Through this, viruses play a role in top-down as well as bottom-up processes, and can potentially alter the course of evolution.

The importance of viruses in the surface oceans is now well recognized, and research is increasingly dedicated to improving our understanding of their role in important marine processes. The viral role in the deep subsurface, however, is rarely considered. Deep within the crust and sediment below the ocean, viruses may play a profound role in altering biogeochemical cycles, structuring microbial diversity, and manipulating genetic content. Yet many questions remain unanswered: Are certain species or strains in the deep subsurface more susceptible to viral infection than others? What role do viruses play in driving natural selection and evolution in the deep biosphere? Is it more common for viruses to persist as protein-bound virion particles, or do they more commonly incorporate their genomes into that of their hosts? What impact do viruses have on their hosts while incorporated as stable symbionts? Can viruses provide their hosts with the keys to survival in the extreme environments of our planet?

The remainder of this introduction will focus on ecological and evolutionary interactions between viruses and their microbial hosts. I begin with a review of viral diversity by briefly describing the diversity of viral morphologies, nucleic acid types, and genetic content, and provide an overview of viral life cycles. I briefly discuss what is known of the viral impact on microbial biogeochemistry, microbial population structure and diversity, and on genetic content and expression patterns. I then apply these concepts to hydrothermal systems, a region where unique attributes such as extreme temperature, high diversity and fluid flux may combine to produce an environment in which viruses play a significant role in manipulating the genetic landscape of deep subsurface microbial communities. Finally, I ask whether these viruses may have been involved in the origin of life in the subsurface. Viruses of the deep may play an important role in altering the evolutionary trajectory of their microbial hosts, and in doing so they complicate the concepts of parasitism and symbiosis in the microbial world, both now and in life's deep past. Ultimately, it is possible that the smallest biological entities on the planet have their most profound influence in its deepest realms, both now and in Earth's early history.

Diversity in the viral world

Viruses infect all three domains of life, and in doing so they adopt a wide variety of morphologies, lifestyle strategies, and genetic materials. These differences in viral types can have important implications for the nature of the virus-host relationship, and for the ways in which viruses can manipulate microbial community structure and evolution. By understanding the types of viruses that predominate in a given system, we can predict the nature of their impact on the host community. Here, I provide a brief overview of different viral types and life cycles, and then describe what types of viruses we might expect to predominate in the deep subsurface, given the environmental conditions.

Viral life cycles

Viruses infecting archaea and bacteria assume two different lifestyle strategies, each with significant implications for the viral relationship with the host and for the nature of virus-host co-evolution. Here, I provide a very simplified overview of viral life cycles; these are illustrated schematically in Figure 0.2. In the *lytic cycle*, viral particles attach to the outside of the host and inject their genetic material into the host cytoplasm. This genetic material then mounts a takeover of cell machinery for immediate synthesis of viral particles, which accumulate within the cell until it bursts, or lyses, releasing the viral particles into the surrounding medium, ready to infect a new host (Figure 0.2A). Viruses employing the *lysogenic cycle*, in contrast, incorporate their genome into the host genome upon infection. Incorporated viral genomes are known as "prophage" or "proviruses," and can lie latent within the cellular genome for many generations. Cells can maintain one or several prophage, which can sometimes provide immunity from superinfection by other viruses. Viral genes can be expressed while integrated into the host genome, and can thereby influence the cellular phenotype. Generally, these viruses are induced, or triggered to enter the lytic cycle, in response to an environmental stimulus. At this point, the viral genome removes itself from the host genome and takes over the cellular machinery to create new viral particles, which then lyse the cell to begin the infection cycle anew (Figure 0.2B). It is likely that the lysogenic cycle predominates among viruses in the deep biosphere, for reasons discussed below.

Viral sizes and morphologies

Viruses can range in size from 20 nm to well over 800 nm, and adopt myriad shapes, genome sizes, and replication strategies. Most viruses possess genomes ranging between a few to ~100 kb, but recently the giant amoeba-infecting Mimivirus was discovered to possess a genome of 1,185 kb, and the virus structure itself is larger than some of the smallest cells (La Scola *et al.*, 2003; Raoult *et al.*, 2004). On the other end of the spectrum, the tiny Sputnik virus possesses a genome of only 18 kb, and parasitizes not a cell, but the Mimivirus itself (La Scola *et al.*, 2008). RNA viruses are often among the smallest of the viruses, with some RNA viruses possessing genomes of only about 2 kb. Giant viruses continue to be discovered in various biomes of the globe (Fischer *et al.* 2010), and much remains to be learned about their lifestyles, replication mechanisms, and their evolutionary and ecological impacts on their hosts.

Viruses of the archaea and bacteria, our focus here, are represented by a wide variety of morphologies, including filamentous, icosahedral, and head-tail viruses. Many of the archaeal viruses possess particularly unusual shapes that have only recently been discovered. The most commonly observed phages (bacterial viruses) in the oceans are the head-tail viruses (Suttle 2005), all of which have linear double-stranded DNA genomes. Among the dsDNA viruses, morphology can give an indication of lifestyle and host range. In the marine realm the most abundant viruses are from the *Podoviridae* family, which have short, non-contractile tails and tend to infect only a narrow range of hosts, usually only particular strains within a species (Suttle 2005). In contrast, the members of the *Myoviridae* family, with contractile tails, and the *Siphoviridae*, with long non-contractile tails, tend to have a broader host range. Consequently, environments dominated by *Myoviridae* or *Siphoviridae* are more likely to be sites of interspecies viral infections.

However, viruses are not limited to the use of double-stranded DNA. Viruses also use single-stranded DNA (ssDNA) as their genetic material, and ssDNA viruses are increasingly found to be important members of the marine viral community. A recent study found that *Microviridae,* a family of ssDNA viruses, is one of the most common viral types in marine waters (Angly *et al.*, 2006). Viruses also use RNA as their genetic material, and can be double- or single-stranded with plus or minus sense RNA strands. RNA viruses have been found to be important constituents of the marine ecosystem (Culley *et al.* 2003, 2006), infecting members across the trophic levels, from bacteria to whales. Retroviruses are one type of RNA virus that use an enzyme called reverse transcriptase to produce DNA from their RNA genomes, and then integrate this DNA into the genome of the host. These also occur in both double-stranded and single-stranded forms. Interestingly, while retroviruses are common in eukaryotes, none have yet been found to naturally infect either the archaea or the bacteria.

While much is known about the morphologies and nucleic acid types of bacteriophages, very little is known about archaeal viruses. The few archaeal viruses isolated thus far have morphologies vastly different from those seen in bacterial viruses (Pina *et al.* 2011, Prangishvili *et al.* 2006). Some archaeal viruses possess a never-before-seen ability to change their morphology outside of the host, extruding tails on each end of an initially lemon-shaped viral capsid after release from the host (Häring *et al.*, 2005).

The unusual viral shapes encountered among the archaeal viruses are occasionally accompanied by unique release mechanisms from the host cell, such as the formation of pyramid-like structures in archaeal membranes that serve as virus outlet sites (Bize *et al.* 2009; Brumfield *et al.* 2009). Most archaeal viruses studied to date are double-stranded DNA viruses, with only a single ssDNA archaeal virus discovered thus far (Pietilä *et al.* 2009). However, these results almost certainly reflect the nature of the detection techniques that have been utilized thus far and not the true diversity of archaeal viruses in natural environments. Considering the similarities between the eukaryotic and archaeal transcription apparatus, the discovery of archaeal RNA viruses and retroviruses seems imminent and may have great potential for yielding important insights into viral evolution. In the deep subsurface biosphere, where archaea constitute a larger proportion of the community than in surface oceans (Biddle *et al.*, 2006), archaeal viruses may dominate, and further study may reveal as yet unknown morphologies or life strategies. Finally, a recent metagenomics study in an acidic, high-temperature lake in Lassen Volcanic Park, USA, uncovered a viral genome sequence suggesting recombination between an RNA and a DNA virus (Diemer and Stedman, 2012). While the host of this particular virus was most likely eukaryotic, this study points to the possibility of such recombination events, which may occur between bacterial or archaeal RNA and DNA viruses as well.

Genetic diversity

An important question in viral ecology is the degree to which viral types are restricted to a given environment, or whether there is movement across biomes. In this sense viruses represent a further test of the null hypothesis of microbial biogeography: "Everything is everywhere, but the environment selects" (O'Malley, 2007). One of the great challenges in assessing viral diversity and biogeography is the lack of a universal "barcoding" gene, analogous to the 16S small ribosomal subunit among the archaea and bacteria, which might be used to compare across all groups. Therefore, other techniques are used to assess virus biogeography. Steward *et al.* (2000) compared the relative genome sizes of viruses using pulsed-field gel electrophoresis, and found that certain genome size classes are found in many different marine environments. Similarly, Breitbart *et al.* (2004) investigated the environmental distribution of the T7 phage DNA polymerase gene, and found that the same sequences were found in a wide variety of diverse biomes, indicating a ubiquity of similar viruses across diverse environmental types. In this scenario, viral diversity is high locally, but viral types are distributed globally. The observation of globally-distributed viral types implies extensive movement among biomes and potential infection of (and sharing genes between) a wide array of hosts (Breitbart and Rohwer, 2005).

Other studies present a contrasting picture of viral biogeography. For example, genomic analysis of a thermophilic virus of *Sulfolobus* revealed that viruses and their hosts tend to be spatially restricted in hot springs (Held and Whitaker, 2009). Metagenomic analysis of viruses in stromatolites and thrombolites found a similar geographic restriction (Desnues *et al.*, 2008), and metagenomic characterization of viruses in soil found distinctions between viral assemblages in soil samples and those in marine or fecal samples (Fierer *et al.*, 2007). On a larger scale, metagenomic studies have revealed that while certain types of viruses, such as the myoviruses, were ubiquitous across sample sites, others, such as podoviruses and siphoviruses, had more site-specific distributions (Williamson *et al.* 2008b). Thus, an opposing paradigm suggests that distinct groups of viruses are tied closely to specific hosts, resulting in spatial restriction (Thurber, 2009). While further study will provide greater insight into this story, it seems that some viral types are globally distributed, while others are much more spatially restricted. Furthermore, spatial distribution is likely to be determined by host specificity, but this relationship is mostly unexplored. Future work aimed at distinguishing between widely distributed viral types and more locally restricted (and presumably more host-specific) viral types may give insight into which viral types are most likely to facilitate gene flow between biomes.

In this context, the viruses of the deep subsurface represent an interesting case. It might be expected that viruses in the deep subsurface, on the one hand, should have reduced mobility as a result of being restricted within a sediment or rock matrix, and therefore have limited and patchy geographic distribution. This may be particularly the case in sedimented regions away from the ridge axis. On the other hand, fluid flux within the subsurface in regions closer to the ridge axis, as well as allochthonous input from above, might facilitate movement of hosts and therefore of viruses from one locality to the next (Anderson *et al.* 2011a). It seems entirely possible that some viral types are restricted to particular regions of the subsurface, while others are more ubiquitous across the deep biosphere, perhaps in biogeographic correlation with their hosts. Further study will be required to resolve these questions.

Viral impacts on host ecology and evolution

Viruses are a peculiarly potent force in that they can impact host community structure through both bottom-up and top-down control, and they can influence host genetic content through horizontal gene transfer and lysogenic conversion. Here, I provide a brief overview of what is known thus far of the viral impact on host microbial communities, with the aim of better understanding the ecological and evolutionary dynamics of the deep subsurface habitat.

Bottom-up effects: The biogeochemical impact

Through lysis of microbial hosts, viruses convert biomass to dissolved and particulate organic matter (Proctor and Fuhrman 1990). Estimates show that viral lysis removes approximately 20-40% of prokaryotic biomass in the ocean daily, though quantifying mortality rates due to viral lysis is difficult (Suttle 2007). This has tremendous biogeochemical implications, as viral lysis converts organic matter from biomass into the pool of dissolved organic matter (DOM), redirecting it from higher trophic levels and effectively short-circuiting the microbial loop. This phenomenon has been dubbed the "viral shunt," and has the effect of stimulating bacterial production by providing a source of DOM and thus stimulating respiration (Suttle 2007). Moreover, the "viral shunt" is thought to stimulate the ocean's biological pump by accelerating sinking rates of lysed cells or transforming bacterial biomass into dissolved organic matter, though it is unclear what percentage of this lysed material is recalcitrant or labile (Jiao *et al.*, 2010). This is depicted schematically in Figure 0.3A.

The impact of the viral shunt on the deep biosphere naturally depends on virus-to-cell ratios, which impact the rate of infection. As this varies according to depth and location, it is difficult to calculate the net impact of viruses on prokaryotic mortality in the deep subsurface. Danovaro *et al.* (2008) showed that viruses become the predominant source of prokaryotic mortality as depth increases in the sediments; in continental margin sediments off of Chile, it was estimated that viruses were responsible for mortality of 38-144% of bacterial net production (Middelboe *et al.* 2006). In mud volcanoes, viruses account for up to 33% of cells killed daily, and also contributed 49 mg C m^{-2} d^{-1}, a substantial contribution to the total carbon budget (Corinaldesi *et al.*, 2011). Thus it can be expected that viruses in the deep biosphere will have a significant, if poorly constrained, impact on microbial mortality and, by extension, biogeochemical cycles. The extent of viral impact will also necessarily depend upon the predominant life cycle of viruses in the subsurface: if lysis predominates, the virus to cell ratio will be the most important factor in determining the importance of viruses in microbial mortality and trophic cycling; whereas if lysogeny predominates, the viral impact on mortality will also be dependent on the frequency and pattern of induction events within each environment.

Top-down effects: Altering population structure

As predators, viruses also control population structure from the top-down; in the deep subsurface, where other predators such as grazers are likely to be absent, viruses may constitute the sole inducer of cell mortality, aside from natural decay. The question that then arises is how the dynamics of viral host range, lifestyle and infection frequency can alter the structure of host microbial communities.

One of the most influential ideas related to viral control of population structure is the notion of “kill the winner” (Thingstad and Lignell, 1997). Several authors have observed that viral infection rates are dependent upon cell density and growth rate (*e.g.* Middelboe 2000); as most viruses have a fairly limited host range, this implies that if a particular microbial group becomes dominant in a population, these cells are most susceptible to viral infection as a result of their increased density. This is depicted schematically in Figure 0.3B. Consequently, viruses may act as a homogenizing agent on the diversity of microbial communities, effectively maintaining high species evenness. Studies have shown that viruses are instrumental in the termination of certain types of plankton blooms (Bratbak *et al.* 1993). Moreover, Rodriguez-Valera *et al.* (2009) found that regions with the greatest variability within a given species’ genome were regions coding for surface receptors, which are potential phage-recognition targets. They argue that viruses maintain high diversity in a system through kill-the-winner-like purges of ecotypes carrying the same surface receptors, which they coined the "constant-diversity" (CD) model. These viral purges contrast with the natural selection purges in the theory of "periodic selection," in which occasional changes in environmental conditions drastically reduce diversity by eliminating all groups not adapted to those conditions (Cohan, 2002). Thus, viruses can contribute to ecosystem stability by maintaining high levels of diversity, even though they are agents of mortality.

Should viruses be a potent force in the deep subsurface, these impacts on population structure should not be discounted, and the impact is likely to vary depending on the nature of the environment. In stagnant sediments with little fluid flux, for example, environmental conditions may be fairly stable, and thus the CD model posited by Rodriguez-Valera *et al.* may be a primary mechanism for maintaining diversity among strains in subsurface communities. However, in more dynamic environments, such as hydrothermal vent systems, community diversity may be structured through a synergistic combination of periodic selective sweeps through environmental change as well as CD dynamics through viral predation.

Viral manipulation of genetic content and expression

It is thus clear that viruses play a crucial role in molding microbial ecology and evolution: viruses are agents of cell mortality and nutrient recycling, they stimulate co-evolution with their hosts through the virus-host arms race, and they play a hand in the structuring of communities and therefore in the generation of new ecotypes and species. Additionally, viruses are known to manipulate genetic content and expression through horizontal gene transfer and lysogenic conversion. Through these mechanisms, viruses may facilitate adaptation to specific niches within a given ecosystem and thereby exert profound impacts on the evolution of their hosts.

1. Transduction.

Viruses facilitate horizontal gene transfer through the process of transduction, which occurs when a virus introduces foreign genetic material into a host during the course of infection. This can occur during the course of lytic infection in a process known as *generalized transduction*, depicted schematically in Figure 0.3C. During the lytic cycle, a virus degrades host DNA and synthesizes viral particles. In this process, host DNA can be accidentally incorporated into a new virus capsid. The resulting *transducing particles* can then infect a new host, introducing genetic material from a previous host into a new one, which can then recombine into the genome of the new host. In this process, almost any region of genetic material may be transferred from the donor cell to the recipient.

In the process of *specialized transduction*, lysogenic viruses excise their genomes incorrectly, incorporating a small region of adjacent host genetic material into the viral genome. This is shown schematically in Figure 0.3D. Combined, generalized and specialized transduction can have a significant impact on the genetic content of viral hosts: one study estimated that up to 10^{14} transduction events occur per year in Tampa Bay Estuary alone (Jiang and Paul, 1998).

A more recent discovery may increase the estimated rates for transduction even further. Gene transfer agents, or GTAs, are viral-like transducing particles, most likely defective phages, which seem to have been usurped by the host to facilitate the process of horizontal gene transfer. While most have been found in Alphaproteobacteria such as *Rhodobacter* (Lang and Beatty 2000) or *Brachyspira* (Matson *et al.* 2005), GTAs have also been found in methanogens and other groups (Stanton 2007) and may be widespread throughout the archaeal and bacterial domains. A recent study suggested that GTA transduction rates may be over one million times higher than previously reported viral transduction rates in the marine environment (McDaniel *et al.*, 2010). Because of their small size (Matson *et al.* 2005) GTA particles should be well-represented in “viral” metagenomes, but positive identifications of GTAs are difficult due to the scarcity of sequenced GTAs thus far (Kristensen *et al.* 2009). Nevertheless, GTAs may constitute a crucial source of genetic innovation in all biomes of the planet, including the deep subsurface. Isolation of strains encoding GTAs may be necessary to enable metagenomic identifications and to increase our knowledge of the scope of their impact.

2. Expression of genes during the course of infection.

Viruses can also carry genes that are expressed during the course of infection. These genes often serve to improve host fitness, which in turn improves virus fitness while the virus is dependent on the host. Some of these genes are expressed by the virus during the lytic cycle, presumably as a means to support host machinery while viral genomes and capsids are replicating within the cell. The most well-known examples of this are photosynthesis genes expressed by cyanophage infecting *Prochlorococcus* and *Synechococcus* (Mann *et al.* 1993). Genes encoding the photosystem core reaction center D1 are expressed during the course of infection in *Prochlorococcus* phage, and it was proposed that these genes improve phage fitness by supporting host photosynthesis during infection (Lindell *et al.* 2005).

Lysogenic viruses can also manipulate host genetic content while integrated as prophage in the cell. As prophage, lysogenized viruses depend on the host for survival over a longer term, and thus benefit from improving host fitness while integrated in the host genome. In some lysogenic viruses, selection has favored the maintenance of genes known as "fitness factors," genes that are encoded and expressed by prophage that alter host phenotype and enhance fitness. One of the most well-known examples of this is the production of cholera toxin by a filamentous bacteriophage integrated into the genomes of virulent *Vibrio cholera*e strains (Waldor and Mekalanos, 1996). Studies have shown that infection by a prophage can drastically alter the phenotypic range for a given species (Vidgen *et al.* 2006). In some cases, phage can also alter host phenotype by suppressing certain metabolic capabilities; it has been suggested that these phage can act to slow host metabolism to shut down unnecessary pathways and conserve energy in environments with low nutrient or energy resources (Paul 2008)—conditions that are expected to be quite common in many deep subsurface ecosystems.

Thus, in addition to influencing host evolution through top-down or bottom-up control of microbial communities, viruses can directly manipulate host genetic content in multiple ways: through generalized or specialized transduction, or through expression of genes either during the lytic cycle or as prophage. Next, I will focus on how the unique attributes of the deep subsurface biosphere, and vents in general, may create an environment ripe for viral manipulation of host genetic content.

Viral manipulation of the deep subsurface biosphere

While the significant role played by viruses is becoming increasingly apparent in the surface oceans, few studies have been conducted on viruses inhabiting hydrothermal systems. Ortmann and Suttle (2005) found that the abundance of viral-like particles in diffuse flow fluids was approximately 10^6 per milliliter, about ten times that of cells, a ratio that is also typical of surface seawater. Williamson *et al.* (2008a) found that a higher abundance of viral-like particles were induced from hydrothermal vent microbial communities exposed to a mutagen compared to those from the upper water column, suggesting that lysogeny is a more predominant lifestyle at vents than in the upper water column. Metagenomics has revealed that the marine vent viral assemblage has the potential to infect a wide variety of bacterial and archaeal hosts from a range of thermal regimes, reflecting the gradient-dominated nature of the environment (Anderson *et al.* 2011b). Together, these studies suggest that many different archaeal and bacterial groups may have prophage integrated into their genomes that potentially introduce novel genetic material or express fitness factors. If this is the case, then genomic analyses of subsurface archaea and bacteria that contain prophage should be a fairly efficient, though clearly biased, approach for exploring the diversity of subsurface viruses. Ideally, such analyses would be coupled with a metagenomic census of free viral particles (*e.g.* Anderson *et al.* 2011b) in order to compare lysogenic and lytic viruses. Clearly, much more research remains to be done to better understand the nature of viral roles in the subsurface.

Deeply buried sediments

Regions of the deep subsurface with more restricted fluid flux, particularly in regions with high sedimentation such as on continental margins, present a drastically different set of conditions for microbial inhabitants. Within the sediment matrix, viral mobility may be reduced, resulting in a potentially lower host contact rate. This would be the case especially if cell abundances are low in deeply buried sediments, such as in the sediments of the South Pacific Gyre (D'Hondt *et al.*, 2009). On the other hand, viruses that form small but hardy particles may be less affected by restrictions on fluid flux than other mechanisms of genetic exchange, which would accentuate the importance of viruses in the ecology and evolution of these isolated communities. The challenges faced by microbial communities inhabiting these regions include limitations on nutrient and energy levels, particularly in deeply buried sediments, where organic carbon and potential oxidizing agents are scarce (Jørgensen and D'Hondt, 2006), and metabolic rates have been shown to be extremely low (Røy *et al.* 2012). The low activity and long doubling times of cells in these regions likely provides further resistance to viral infection, which is largely dependent on the density and activity of the host (Fuhrman 2009). As with cellular abundances, viral abundance decreases with depth in the sediments. Middelboe *et al.* (2011) quantified viral abundance with depth on the eastern margin of the Porcupine Seabight and observed approximately 10^8 VLPs/cm^{-3} at 4 meters below the seafloor (mbsf), to about 10^6 VLPs/cm^{-3} at 96 mbsf. However, another study by Engelhardt *et al.* (2012) found that the virus-to-cell ratio increased with depth in the sediments, potentially indicating continued viral production at these depths in sediments, rather than long-term viral preservation, as had been suggested previously.

Lysogeny is expected to be a common viral lifestyle in the deep subsurface, resulting from selection for a viral lifestyle that limits the necessity for finding hosts in a sparse soil matrix and in harsh environmental conditions. Work by Engelhardt *et al.* (2011) has demonstrated that nearly half of the bacterial isolates tested from a deep-sea sediment core harbored prophage, and a subsequent study found that all deeply buried isolates of a common deep-sea species, *Rhizobium radiobacter,* were lysogenic (Engelhardt *et al.* 2012). The ubiquity of lysogeny could have interesting implications for cellular survival in these energy-limited systems, where archaea and bacteria are likely to be under strong selection pressure to harness alternate forms of energy when they are available, and to minimize energy use when energy sources are limiting. Previous studies have shown that certain lysogenic phage can actively repress host metabolic genes, and therefore repress wasteful host metabolic processes when conditions are not favorable (Paul 2008), a trait that would be particularly useful in the energy-limited subsurface. Further work in deeply buried sediments will reveal what genes are expressed by these lysogenized phage, perhaps revealing that the relationship between virus and host transcends the parasitic, becoming instead a mutualistic symbiosis.

Viral impacts on surface-attached communities

Regardless of whether a deep subsurface habitat is hydrologically active or not, most of the inhabitants probably live as biofilms (*i.e.* communities attached to some sort of hard surface, which could be hard rock, vent chimney deposit, or sediment). Generally, biofilms have much higher cell density than the surrounding medium, so there is potential for biofilms to be hotspots of viral activity. Viruses are known to accumulate in biofilms growing in drinking water systems (Skraber *et al.* 2005), and viral lysis is frequent during biofilm development in *Staphylococcus aureus* (Resch *et al.* 2005). Metagenomic sequencing of biofilms from an acid mine drainage site recovered complete viral genomes and extensive evidence that bacterial genomes are continually influenced by viral infection (Andersson and Banfield 2008). Viral genes are highly expressed in *Pseudomonas aeruginosa* biofilms (Whiteley *et al.*, 2001), and viral-mediated cell death is a normal component of biofilm development (Webb *et al.* 2003). Beyond these basic detections of viruses and viral genes in biofilm habitats, however, surprisingly little is known about the molecular mechanisms and ecological impacts of virus-biofilm interactions, especially in the subsurface.

The polysaccharide-rich extracellular matrix of biofilms is probably a barrier against infection, but it is clear that viruses can penetrate the barrier, in some cases via enzymatic digestion (Weinbauer, 2004). Another complication is the recent finding that a human virus can generate its own biofilm-like matrix (Thoulouze and Alcover, 2011). The prevalence of viruses encased within extracellular matrices is entirely unexplored in subsurface ecosystems, and it is likely that such surface-attached viral populations can evade detection and depress counts of viral-like particles in fluid samples. Therefore, interpretations of viral abundance and activity data from subsurface fluid samples must consider how well the fluid samples represent the rocks and sediments that provide habitat for most bacteria, archaea, and viruses in the subsurface.

In addition to high cell density, many biofilm communities also have high genetic and phenotypic diversity, resulting in complex interactions among many species on microscopic spatial scales (Stoodley *et al.* 2002). One potential consequence is that viruses with high host specificity may have greater difficulty finding their host in a tightly-packed, diverse biofilm, resulting in a large total number of viruses, each capable of infecting only a tiny proportion of the diverse biofilm community. This scenario is one possible explanation of the “infectivity paradox”: the observation that many habitats have high viral abundance but low infectivity (Weinbauer 2004). Preliminary data (Filippini *et al.* 2006) suggest that biofilms exemplify the infectivity paradox, but no such studies have been conducted in the deep subsurface.

Many studies have demonstrated the importance of lateral gene transfer in biofilm communities (Molin and Tolker-Nielsen, 2003), and in some cases, viruses have been identified as the agents of transfer (Webb *et al.* 2003; Whiteley *et al.* 2001). It is clear that in biofilms, gene transfer is not a rare curiosity but a fundamental aspect of biofilm formation and development, notably as a mechanism for a phenomenon known as “phenotype switching.” In *Staphylococcus epidermidis* biofilms, for example, genomic insertion of a mobile genetic element results in a stable population of variant cells unable to produce the biofilm matrix (Ziebuhr *et al.*, 1999). The effect is reversible because the inserted DNA is frequently excised, restoring biofilm production. Other species also exhibit reversible phenotype switching associated with biofilm formation, and viruses have been implicated in at least one case (Webb *et al.* 2004). The evolutionary dynamics of such processes have not been explored experimentally, but one simulation study predicted that the coexistence of multiple phenotypes in a biofilm community can be promoted by continual gene transfer. If two phenotypes are linked, as in a syntrophic partnership, the fitness of each member is dependent on the fitness of the other. Therefore, natural selection of such cells living in a dense community could result in complex inter-species relationships that are dependent on (potentially viral-mediated) transfer of genetic content. In summary, future research is likely to reveal that biofilms in subsurface habitats exemplify the concept described above that viral activity in the subsurface is likely to have complex and varied evolutionary consequences that extend beyond just cell mortality.

Tools for analysis: Viral metagenomics in the deep subsurface

Given the current state of knowledge about viruses in the deep subsurface, how can we gain further insight into the role they play in manipulating geochemical cycles, altering diversity, and influencing the course of evolution in their hosts? One method by which we can probe the viral world is through metagenomics, in which a sample of community DNA is extracted and sequenced directly from the environment. While metagenomics in the microbial realm has traditionally focused on asking “who is there?” and “what are they doing?”, viral metagenomics presents a unique set of challenges. Viruses are generally separated from the microbial fraction through size fractionation, which may exclude large viral particles or include small cells, so contamination is an issue of concern. Moreover, one of the primary challenges facing viral metagenomics is the large proportion of unknown sequences. The average percentage of viral metagenomic sequences with no match to existing databases ranges from about 60 to over 90%, depending on the read length (i.e.. Anderson *et al.* 2011b; Angly *et al.* 2006; Breitbart *et al.* 2002; Desnues *et al.* 2008; Rosario and Breitbart 2011). The vast number of sequences with no match to existing databases presents a challenge to viral ecologists seeking to understand who viruses infect, what impacts they have on their hosts, and what types of genes they encode and transfer.

One goal of viral ecology in any environment is the identification of which archaeal or bacterial groups play host to those viruses. This information is key to understanding how viruses may impact a given microbial community. If only certain groups are most susceptible to viral attack, this may have further implications for microbial population structure or biogeochemistry. Some information about potential hosts can be gleaned by identification of known viral groups: *Rudiviridae* and *Fuselloviridae*, for example, are only known to infect the archaea. As stated previously, classification of viral metagenomic sequences is tremendously challenging, and even if it is successful, only limited information is gained because many families of viruses infect wide ranges of hosts. One method that has been used to identify potential hosts of a viral assemblage is to identify clustered regularly interspaced palindromic repeats (CRISPRs), an immune system used by archaea and bacteria to combat invasive genetic material, including viruses and plasmids. Chapter 3 outlines my approach to using CRISPR spacers as a means of identifying the potential hosts of a viral assemblage.

Another outstanding question regarding viral roles in the subsurface is the degree to which viruses mediate horizontal gene transfer, or manipulate archaeal and bacterial genomes through incorporation as prophage. One way to address this question is to examine viral and cellular metagenomes for sequences potentially associated with mobile elements, such as those that encode transposases, integrases, and recombinases. The presence of abundant genes that encode such enzymes provides one piece of evidence that the organisms in a particular community actively exchange genes. Chapter 4 describes results from a comparison of viral and cellular metagenomes, and suggests that the vent viral gene pool may be enriched in genes that facilitate host adaptation to the hydrothermal vent environment.

Metagenomics holds great potential for illuminating the virus-host relationship, and further sequencing of both viral and cellular metagenomes from regions of the deep subsurface will contribute much to our understanding of which organisms are most susceptible to viral infection, whether the lytic or lysogenic lifestyle is more common in the subsurface, and the nature of the role viruses play in facilitating horizontal gene transfer in the deep subsurface.

VENTS, VIRUSES AND THE ORIGIN OF LIFE: AN ASTROBIOLOGICAL PERSPECTIVE

The evidence appears to be clear that viruses can have a substantial impact on the evolution of their hosts, and have most likely been doing so for billions of years. But for how long has this mutual evolutionary relationship persisted? When and how did viruses originate? The question is particularly relevant here because the deep subsurface, and hydrothermal systems in particular, are often considered the most ancient continuously inhabited ecosystems on the planet (Reysenbach and Shock 2002), and indeed, are often thought to have been an important setting for the origin of life on Earth.

Hydrothermal vents and the deep subsurface: key settings in the origin of life

On the Hadean Earth, about 4 billion years ago, hydrothermal vent systems would have been present in perhaps even greater abundance than they are today. Residual heat of formation would have resulted in a volcanically and seismically more active planet, with longer mid-ocean ridges and more plate tectonic activity (Hargraves, 1986), resulting in a higher incidence of water-rock reactions at the bottom of the ocean. Both basalt-hosted and peridotite-hosted hydrothermal systems are likely to have been present in the Hadean Earth. Metal-sulfide minerals in basalt-hosted systems, including pyrite, have been implicated as key catalysts in several important prebiotic reactions in which

CO or CO_2 is fixed into simple organic compounds (Cody 2004; Wachtershauser 1990; Wächtershäuser 1988; Wächtershäuser 1988). Peridotite-hosted systems, formed off-axis and powered by serpentinization through the interaction of seawater with peridotite, are characterized by lower-temperature fluids with high pH and form large calcium carbonate structures (Kelley *et al.* 2001).These vents may have also acted as a source for key organic compounds generated in the process of serpentinization, including formate, acetate, methane, organic sulfur compounds, and larger hydrocarbons (Heinen & Lauwers 1996; Lang *et al.* 2010; Proskurowski *et al.* 2008). Moreover, the calcium carbonate porous structures formed in these systems may have acted as a concentrating mechanism for early prebiotic compounds (Baaske *et al.*, 2007).

One of the most appealing aspects of hydrothermal vents as a setting for the origin of life is the formation of geological, physical, and chemical gradients in these systems (Baross & Hoffman 1985). These gradients provide a wide range of environmental conditions within a relatively small physical space with fluid flow between them, facilitating the occurrence of multiple chemical processes across a multiplicity of environmental conditions in parallel. These gradients extend beyond hydrothermal vent fields themselves to other regions of the deep subsurface. The minerals catalyzing reactions in one region, such as at basalt-hosted hydrothermal vents, would have differed from those in other regions of the subsurface, such as at peridotite-hosted systems or in sedimented regions. As mentioned above, much of the ocean crust is linked by fluid flux, which moves at different flow rates and volumes depending on the depth and degree of porosity in the crust. Thus, compounds synthesized in one region of the ocean crust, whether at a hydrothermal system or more distal to a mid-ocean ridge, could be transferred from one region of the subsurface to the next.

In this sense, the deep subsurface may have acted as a natural laboratory for the origin of life, in which multiple “experiments” could have been carried out in tandem. Later, the products of these natural experiments could have been combined to form an autocatalytic network. Several studies have suggested that the chemiosmotic gradients at vent sites, combined with enclosed pore spaces and organic syntheses, could have resulted in the first autocatalytic networks (Koonin and Martin 2005; Lane *et al.* 2010; Martin and Russell, 2007; Martin *et al.* 2008). Martin, Russell and others describe a model in which a chemiosmotic potential is generated across the membrane of an iron-sulfide bubble, which they presume would form in an anoxic Hadean ocean. The potential could then have been harnessed to yield a protometabolism based on the reduction of CO_2 by H_2.

The question that arises is how the first self-replicating entities formed and evolved in these settings. Koonin and Martin (2005) suggest that self-replicating networks could have formed within the walls of these iron-sulfide compartments. In their scenario, each component of the network consisted of a selfish RNA molecule encoding one or a few proteins, with the original selection pressures favoring rapid self-replication. The authors refer to these replicating entities as "virus-like RNA molecules," which Koonin then elaborated upon in a later publication detailing the "Virus World" (Koonin *et al.* 2006). In this model, the authors describe a scenario in which viruses emerged early from the various replicating entities and networks that formed part of the RNA world. These scenarios suggest that viruses may have played a primary role at the earliest stages of life's evolution.

The viral role in the origin of life

Viruses have not always been considered to be primordial elements. Historically, three theories were put forward regarding the origin of viruses: first, that viruses were originally parasitic cells that evolved into a viral-like form (the "reduction hypothesis"); second, that viruses were rogue genetic elements from cells that developed a protein capsid to survive in an extracellular state (the "escape hypothesis"); and third, that viruses originated in parallel with cells (the "virus first hypothesis") (Prangishvili *et al.* 2006). The last hypothesis, however, has been gaining favor as scientists have found that viruses infecting different domains of life share certain "hallmark genes" that are missing from cellular genomes, perhaps pointing to an early origin that predates the divergence of the three domains of life (Koonin *et al.* 2006). Others have suggested that DNA as a genetic material first arose in a virus, which later spread to the cellular world (Forterre 2006).

The tremendous diversity of viruses, though, greatly complicates an elucidation of their origin. Viruses encompass one portion of a spectrum of mobile genetic elements, which range in size and complexity from simple introns and transposons, to GTAs, to

RNA viruses, viroids, and satellite viruses, to dsDNA viruses and the giant Mimivirus described above. These elements may not share a common origin, yet in many cases, many virus-like elements share genes that are not found in the cellular world (Koonin *et al.* 2006), or share structural features in their protein capsids (Bamford *et al.* 2005). Many of these shared attributes transcend domains, leading many to consider viruses to be ancient.

The attribute that all viruses share is their dependence on a host for the purposes of replication: in a word, these are parasites. Consideration of the role of parasites in the origin of life is not a new concept. In an RNA-protein world, or even a pre-RNA world, parasites could have undermined replication networks, as they could take resources from these replication networks (or "hosts," in a sense) without benefiting them, and thus destroy the cycle (Maynard Smith and Szathmáry 1997). In these networks, elements are linked such that each element replicates another element in the cycle. Parasites emerge when a mutant of one of the elements is preferentially replicated, but does not replicate another element in the cycle (Figure 0.4A). It has been suggested that containing replication cycles within a compartment, or at least confining them to a surface, may circumvent this problem by placing the selective pressure not on the individual elements within a replication cycle, but on the cycle as a whole (Maynard Smith and Szathmáry 1997). This effectively provides a basis for heredity and competition between individuals, which are required for natural selection to occur. Moreover, spatial structuring of the environment may reduce the spread of parasites from one hypercycle to the next (Boerlijst and Hogeweg, 1991). However, spatial structuring may prevent "sharing" of new functions through horizontal gene transmission (Poole, 2009).

Yet as discussed previously, parasites can at times improve the fitness of the host they depend upon. Just as modern viruses can express fitness factors to boost the fitness of their host, the same may have applied in life's early evolution: for example, if a selfish element were to contribute toward the overall fitness of a given replication cycle, such as through stabilizing another element in the cycle, this would improve the fitness of the whole cycle and therefore the fitness of the element as well (Figure 0.4B). Indeed, this may have been the means by which new functions were added to replication networks. Just as modern viruses can contribute novel genetic material through transduction or expression of fitness-boosting genes, ancient parasites may have increased the functionality of the networks they were part of by linking together disparate networks. Rather than (or in addition to) presenting a problem for early replication networks, early viruses may have provided a means by which to increase their functionality. Viral-like particles or selfish elements may have acted as a means to share genes between networks, and ultimately may have allowed early genomes to expand (Figure 0.4C). In this sense the meaning of words like “parasite” or “mutualist” become blurred, as selection at this level may favor varying degrees of parasitism or mutualism; in the RNA world, selection operated at the level of both the individual elements and at the level of whole networks.

This scenario is consistent with previously published ideas about the origin of life that may or may not have explicitly specified roles for viruses. For example, the idea that all life today evolved from primitive cells in an early biofilm-like community evolving “through prolific genetic exchange with other 'precells' in the community, perhaps involving structures resembling transposable genetic elements and viral-like particles” was inspired by the initial discovery of prolific subsurface life evident at hydrothermal vents (Baross and Hoffman, 1985). Woese has developed in detail the idea of a communal ancestor in which horizontal gene transfer is the primary driver of evolution (Woese 1998, 2002) and viral-like elements could very well have been one of the mediators of this gene transfer.

However, as with modern viruses, these parasitic elements likely would have had a wide host range, in which some could spread rapidly to other networks, whereas others were more restricted. Modern viruses also exhibit a range of virulence, in which some were almost entirely parasitic whereas others are almost entirely mutualistic, and prebiotic selfish elements may have had similar characteristics. In this sense, indiscriminate horizontal gene transfer in the communal ancestor may have been disruptive by facilitating the spread of the more virulent, entirely deleterious parasites (Poole, 2009). It is also unclear how the communal ancestor would have evolved “as a unit” (Woese 1998) without competition or selection with other units. In this light, the most likely scenario is one in which spatial structuring of the environment facilitates selection at the level of an entire network, rather than on individual elements. This would also serve to restrict the movement of wide-ranging, deleterious parasites. Iron-sulfur bubbles or pores in hydrothermal systems could have acted as a structuring mechanism prior to the emergence of lipid membranes (Koonin and Martin 2005; Martin and Russell 2007; Russell and Hall 1997) (Figure 0.4D).

On a larger scale in the prebiotic world, there would have been extensive fluid flux in the subsurface, and this could have served as a conduit for nutrients or products of prebiotic reactions from other environments. This would have connected these networks and facilitated some degree of gene sharing between them, as well as provided them with important prebiotic precursors (Figure 0.4E). Gradients in temperature, pH, chemical and mineralogical composition through the subsurface or in hydrothermal structures would have generated diversity in these replicating networks, creating variation in the population and facilitating selection among them. In this sense, the environment may have fostered the earliest stages of natural selection, and these early viral-like or “selfish” elements may have been important in facilitating gene transfer between these networks, allowing them to grow and change.

Regardless of the degree to which horizontal gene transfer occurred, subpopulations within the ancestral community became more resistant to genetic exchange with other subpopulations over time, which may have arisen as a defense against parasitic genetic elements: the first viruses. The crossing of this “Darwinian threshold” from one ancestral community to many independent cells marked the origin of speciation and the emergence of the life forms we know today (Woese 2002), and it is possible that viruses or viral-like elements were intimately involved in this critical stage in the evolution of life.

Conclusion

There is clearly much work that remains to be done to understand the nature of the viral impact on biogeochemical cycling, microbial community structure, and evolution in the deep subsurface. Yet the few available details provide tantalizing hints that the role of viruses in the deep subsurface could be profound on many levels. Viral infection may significantly impact biogeochemical cycling in the subsurface through lysis of cellular biomass, releasing nutrients and compounds that would otherwise be entrained in biomass. Viruses are also known to alter the structure of the microbial communities they infect, potentially increasing overall diversity through lysis of cells that become most abundant in a given region. Through the process of lysogeny and transduction, viruses may manipulate the genomes and expression of the hosts they infect throughout the subsurface, effectively resulting in a mutualistic, symbiotic relationship between host and virus that transcends traditional notions of viruses as parasites. Indeed, this role of virus as parasite, as mutualist, and as a sharer of information through gene transfer may be a fundamental underpinning of life in the deep subsurface that extends back in time to the dawn of life itself. Finally, the hot subsurface environments associated with hydrothermal systems harbor many of the most “deeply rooted” microorganisms on the universal phylogenetic tree of life; most are hyperthermophilic archaea. Very little is known about the viruses that are associated with these microorganisms. Given their antiquity including their primordial setting, it is possible they harbor viruses and virus-like particles that could lead to a better understanding of the origin of viruses and their role in the early evolution of life.

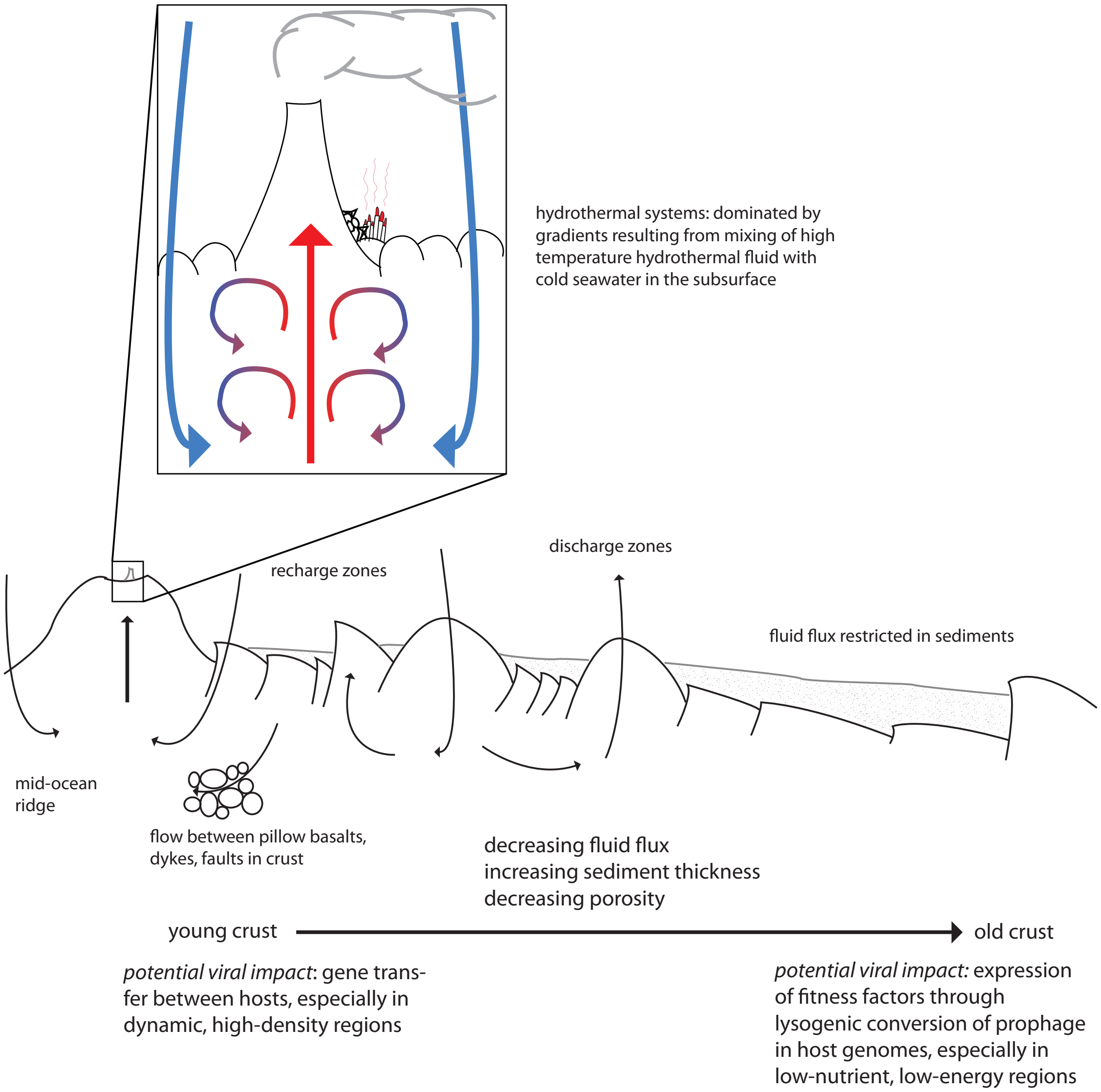

hydrothermal systems: dominated by gradients resulting from mixing of high temperature hydrothermal fluid with cold seawater in the subsurface
recharge zones
discharge zones
fluid flux restricted in sediments
mid-ocean ridge
flow between pillow basalts, dykes, faults in crust
decreasing fluid flux
increasing sediment thickness
decreasing porosity
young crust
old crust
potential viral impact: gene transfer between hosts, especially in dynamic, high-density regions
potential viral impact: expression of fitness factors through lysogenic conversion of prophage in host genomes, especially in low-nutrient, low-energy regions

Figure 0.1. Schematic depicting fluid flux and porosity in the marine subsurface. Arrows depict fluid flux across the seawater/subsurface interface, and throughout the subsurface. Inset shows detail of fluid flux and mixing of seawater and high-temperature hydrothermal fluid in mid-ocean ridge hydrothermal systems, where high-temperature hydrothermal fluid (red) rises from high-temperature water rock reactions deeper in the subsurface and mixes with colder seawater (blue). Crust ages as it moves away from the mid-ocean spreading ridge.

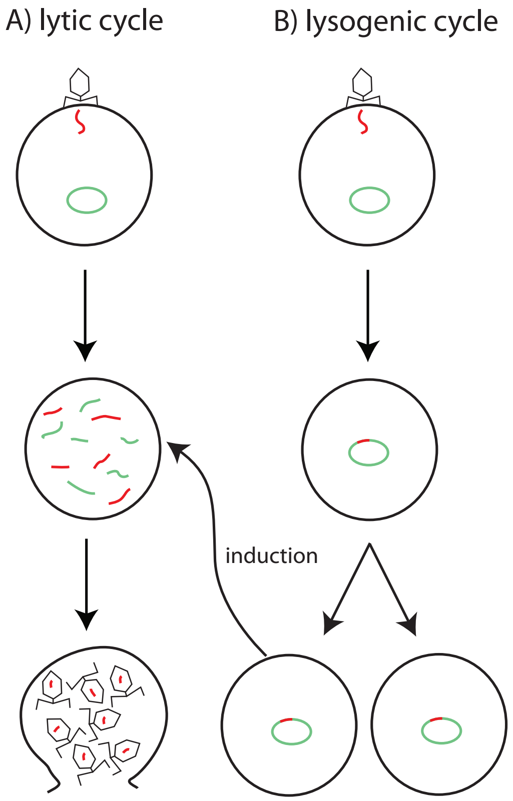


Figure 0.2. Schematic and generalized depiction of the lytic and lysogenic cycles of phage. A) Lytic cycle, in which a virus lands on the cellular membrane, injects genetic material into the cytoplasm, resulting in viral takeover of cellular machinery. New viral capsids are synthesized, packaged with viral genomes, and lyse the cell. B) Lysogenic cycle, in which a virus lands on the cellular membrane and injects its genetic material into the cytoplasm, and then integrates into the host genome. This "prophage" lies latent for several generations, then enters the lytic cycle in response to an induction event.

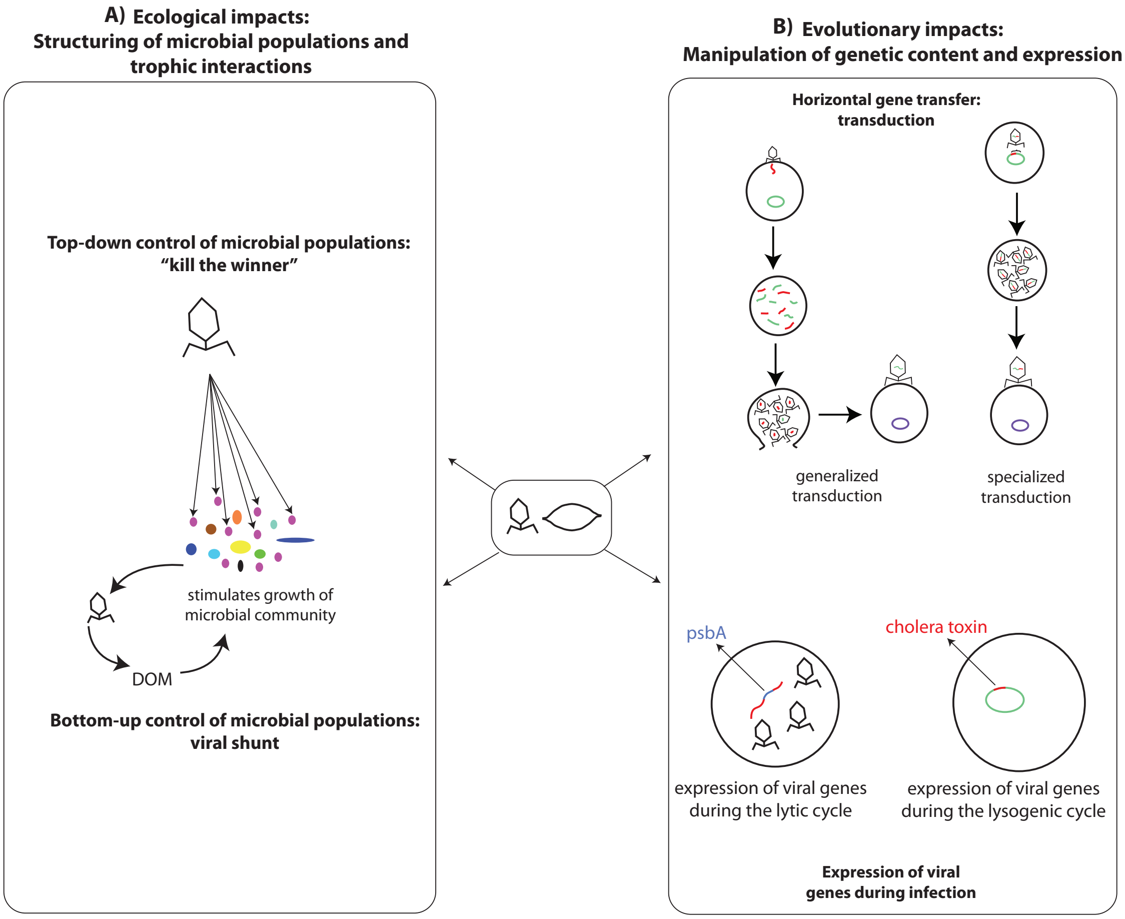
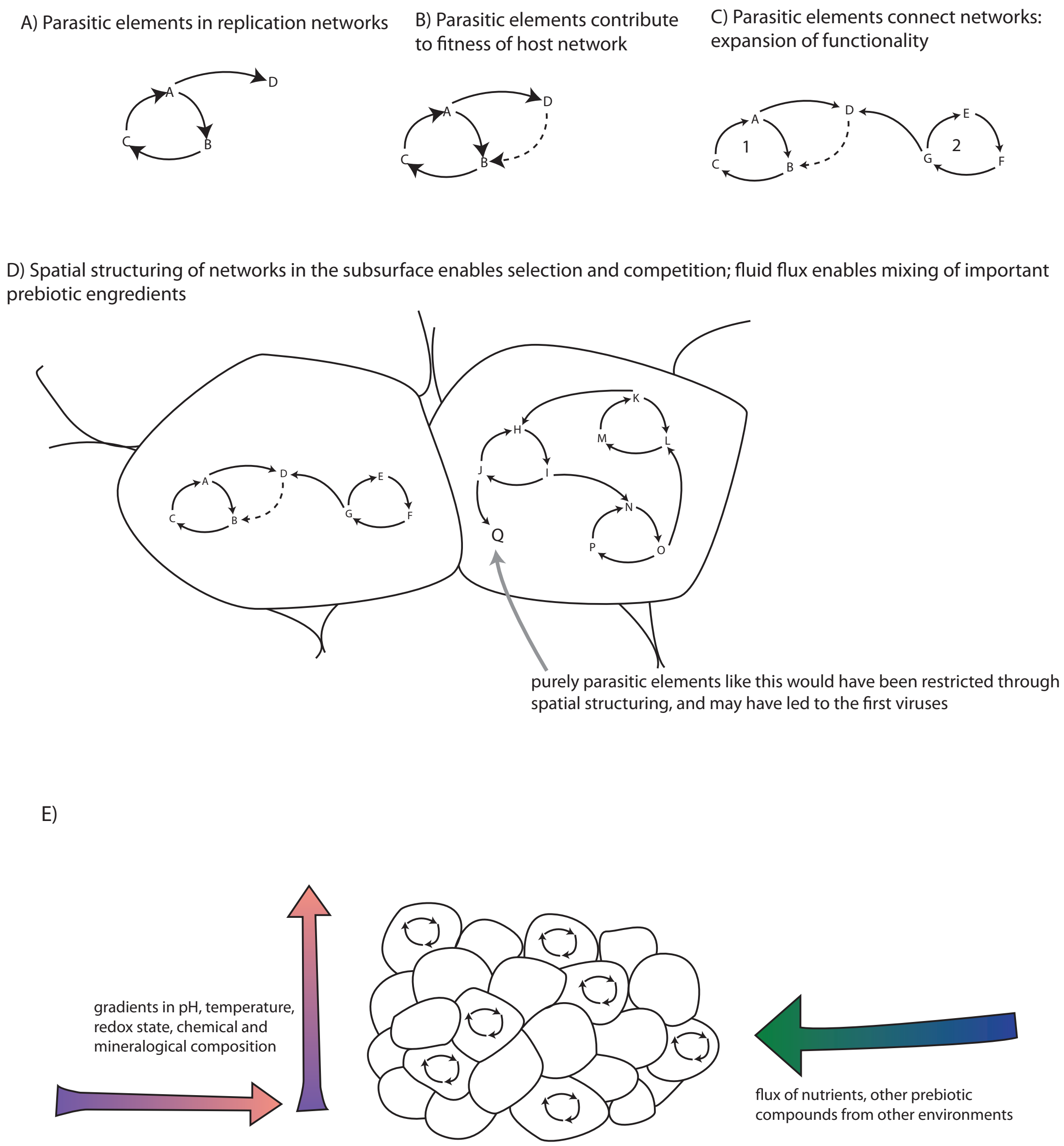

A) Ecological impacts:
Structuring of microbial populations and trophic interactions
Top-down control of microbial populations: "kill the winner"
stimulates growth of microbial community
DOM
Bottom-up control of microbial populations: viral shunt
B) Evolutionary impacts:
Manipulation of genetic content and expression
Horizontal gene transfer: transduction
generalized transduction
specialized transduction
psbA
cholera toxin
expression of viral genes during the lytic cycle
expression of viral genes during the lysogenic cycle
Expression of viral genes during infection

Figure 0.3. Schematic summary of the effects of archaeal and bacterial viruses on hosts. A) Ecological impacts in which viruses play a role in structuring microbial diversity and trophic interactions. Through "kill the winner" type dynamics (see text for details), viruses maintain greater evenness within a population by reducing the abundance of the most abundant strains. Through the viral shunt, viruses lyse cells and thus contribute to the pool of dissolve organic matter (DOM), which may stimulate growth of the microbial community. B) Evolutionary impacts in which viruses play a role in manipulating genetic content and expression of their hosts. Viruses can alter the genetic content of their hosts through horizontal gene transfer (called transduction). Lytic viruses do this through the process of generalized transduction, in which individual viral capsids are packaged with host genetic material instead of viral genetic material, which is then transferred to the next host. Lysogenic viruses can also do this through the process of specialized transduction, in which enzymes removing the viral genome from the host genome accidentally remove some host genetic material adjacent to the prophage. This genetic material is also packaged into a viral capsid and transmitted to the next host. Viruses can also express viral genes during the process of infection: during the lytic cycle, viruses can express genes to support cellular activity while viral capsids and genomes are synthesized (such as psbA); they can also do this during the lysogenic cycle, while integrated as prophage (such as cholera toxin).



A) Parasitic elements in replication networks

B) Parasitic elements contribute to fitness of host network

C) Parasitic elements connect networks: expansion of functionality

D) Spatial structuring of networks in the subsurface enables selection and competition; fluid flux enables mixing of important prebiotic engredients

purely parasitic elements like this would have been restricted through spatial structuring, and may have led to the first viruses

E)

gradients in pH, temperature, redox state, chemical and mineralogical composition

flux of nutrients, other prebiotic compounds from other environments

Figure 0.4. Role of parasitic elements in early replication cycles. A) Cooperative replication network. Element D is a parasite to the network because it uses network resources, but does not contribute to network fitness. B) Selection may favor elements that are able to contribute to the fitness of the host network. Here, element D contributes to the stability of element B, thus improving network fitness. This feedback improves the fitness of D as well. C) If element D is also replicated by element G in another network, this replication could link the two networks, thereby increasing network functionality. D) Spatial structuring through restriction to mineral surfaces, or enclosure in pore spaces, could restrict the degree to which parasitic elements (like element Q, considered "parasitic" here because it does not contribute to the parent network or any other networks) spread between networks. E) Scenario in which replicating networks operate in the subsurface, with diverse replicating networks defined by gradients in environmental conditions, and fed by an influx of prebiotic compounds through fluid flux in the subsurface.

References

Anderson, R.E., Brazelton, W.J., and Baross, J.A. (2011) Is the genetic landscape of the deep subsurface biosphere affected by viruses? Front Extreme Microbiol **2**:

Anderson, R.E., Brazelton, W.J., and Baross, J.A. (2011) Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. FEMS Microbiol Ecol **77**: 120–133.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., *et al.* (2006) The marine viromes of four oceanic regions. PLoS Biol **4**: e368.

Baaske, P., Weinert, F.M., Duhr, S., Lemke, K.H., Russell, M.J., and Braun, D. (2007) Extreme accumulation of nucleotides in simulated hydrothermal pore systems. Proc Natl Acad Sci U S A **104**: 9346–51.

Bamford, D.H., Grimes, J.M., and Stuart, D.I. (2005) What does structure tell us about virus evolution? Curr Opin Struct Biol **15**: 655–63.

Baross, J.A. and Hoffman, S.E. (1985) Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. Origins Life Evol B **15**: 327–345.

Biddle, J.F., Lipp, J.S., Lever, M.A., Lloyd, K.G., Sørensen, K.B., Anderson, R.E., *et al.* (2006) Heterotrophic Archaea dominate sedimentary subsurface ecosystems off Peru. Proc Natl Acad Sci U S A **103**: 3846–51.

Bize, A., Karlsson, E.A., Ekefjärd, K., Quax, T.E.F., Pina, M., Prevost, M.-C., *et al.* (2009) A unique virus release mechanism in the Archaea. Proc Natl Acad Sci U S A **106**: 11306–11.

Boerlijst, M.C. and Hogeweg, P. (1991) Spiral wave structure in pre-biotic evolution: Hypercycles stable against parasites. Physica D **48**: 17–28.

Bouvier, T. and del Giorgio, P.A. (2007) Key role of selective viral-induced mortality in determining marine bacterial community composition. Environ Microbiol **9**: 287–97.

Bratbak, G., Egge, J., and Heldal, M. (1993) Viral mortality of the marine alga *Emiliania huxleyi* (Haptophyceae) and termination of algal blooms. Mar Ecol Prog Ser **93**: 39–48.

Brazelton, W.J., Ludwig, K.A., Sogin, M.L., Andreishcheva, E.N., Kelley, D.S., Shen, C.-C., *et al.* (2010) Archaea and bacteria with surprising microdiversity show shifts in dominance over 1,000-year time scales in hydrothermal chimneys. Proc Natl Acad Sci U S A **107**: 1612–1617.

Breitbart, M., Miyake, J.H., and Rohwer, F. (2004) Global distribution of nearly identical phage-encoded DNA sequences. FEMS Microbiol Lett **236**: 249–256.

Breitbart, M. and Rohwer, F. (2005) Here a virus, there a virus, everywhere the same virus? Trends Microbiol **13**: 278–284.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., *et al.* (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A **99**: 14250 –14255.

Cody, G.D. (2004) Transition metal sulfides and the origins of metabolism. Annu Rev Earth Planet Sci **32**: 569–599.

Cohan, F.M. (2002) What are bacterial species? Annu Rev Microbiol **56**: 457–87.

Corinaldesi, C., Dell'Anno, A., and Danovaro, R. (2011) Viral infections stimulate the metabolism and shape prokaryotic assemblages in submarine mud volcanoes. ISME J.

Culley, A.I., Lang, A.S., and Suttle, C.A. (2003) High diversity of unknown picorna-like viruses in the sea. Nature **424**: 1054–7.

Culley, A.I., Lang, A.S., and Suttle, C.A. (2006) Metagenomic analysis of coastal RNA virus communities. Science **312**: 1795–8.

D'Hondt, S., Spivack, A.J., Pockalny, R., Ferdelman, T.G., Fischer, J.P., Kallmeyer, J., *et al.* (2009) Subseafloor sedimentary life in the South Pacific Gyre. Proc Natl Acad Sci U S A **106**: 11651–6.

Von Damm, K.L. (1990) Seafloor hydrothermal activity: black smoker chemistry and chimneys. Annu Rev Earth Planet Sci **18**: 173–204.

Danovaro, R., Dell'Anno, A., Corinaldesi, C., Magagnini, M., Noble, R., Tamburini, C., and Weinbauer, M. (2008) Major viral impact on the functioning of benthic deep-sea ecosystems. Nature **454**: 1084–1087.

Deming, J. and Baross, J. (1993) Deep-sea smokers: Windows to a subsurface biosphere? Geochim Cosmochim Acta **57**: 3219–3230.

Desnues, C., Rodriguez-Brito, B., Rayhawk, S., Kelley, S., Tran, T., Haynes, M., *et al.* (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature **452**: 340–343.

Diemer, G.S. and Stedman, K.M. (2012) A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biology Direct **7**: 13.

Dobzhansky, T. (1964) Biology, Molecular and Organismic. American Zoologist **4**: 443–452.

Edwards, K.J., Bach, W., and McCollom, T.M. (2005) Geomicrobiology in oceanography: microbe-mineral interactions at and below the seafloor. Trends Microbiol **13**: 449–456.

Edwards, K.J., Wheat, C.G., and Sylvan, J.B. (2011) Under the sea: microbial life in volcanic oceanic crust. Nat Rev Microbiol **9**: 703–12.

Engelhardt, T., Sahlberg, M., Cypionka, H., and Engelen, B. (2011) Induction of prophages from deep-subseafloor bacteria. Environ Microbiol Rep **3**: 459–465.

Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R., *et al.* (2007) Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. Appl Environ Microbiol **73**: 7059–66.

Filippini, M., Buesing, N., Bettarel, Y., Sime-Ngando, T., and Gessner, M.O. (2006) Infection paradox: high abundance but low impact of freshwater benthic viruses. Appl Environ Microbiol **72**: 4893–8.

Fischer, M.G., Allen, M.J., Wilson, W.H., and Suttle, C.A. (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. Proc Natl Acad Sci U S A **107**: 19508.

Fisher, A. and Becker, K. (2000) Channelized fluid flow in oceanic crust reconciles heat-flow and permeability data. Nature **403**: 71–4.

Forterre, P. (2006) The origin of viruses and their possible roles in major evolutionary transitions. Virus Res **117**: 5–16.

Gibbons, S.M., Caporaso, J.G., Pirrung, M., Field, D., Knight, R., and Gilbert, J.A. (2013) Evidence for a persistent microbial seed bank throughout the global ocean. Proc Natl Acad Sci U S A **110**: 4651–4655.

Hargraves, R.B. (1986) Faster spreading or greater ridge length in the Archean? Geology **14**: 750.

Häring, M., Vestergaard, G., Rachel, R., Chen, L., Garrett, R.A., and Prangishvili, D. (2005) Virology: independent virus development outside a host. Nature **436**: 1101–2.

Heinen, W. and Lauwers, A.M. (1996) Organic sulfur compounds resulting from the interaction of iron sulfide, hydrogen sulfide and carbon dioxide in an anaerobic aqueous environment. Origins Life Evol B **26**: 131–150.

Held, N.L. and Whitaker, R.J. (2009) Viral biogeography revealed by signatures in Sulfolobus islandicus genomes. Environ Microbiol **11**: 457–466.

Holden, J.F., Summit, M., and Baross, J.A. (1998) Thermophilic and hyperthermophilic microorganisms in 3-30 ^{o}C hydrothermal fluids following a deep-sea volcanic eruption. FEMS Microbiol Ecol **25**: 33–41.

Huber, J.A., Butterfield, D.A., and Baross, J.A. (2003) Bacterial diversity in a subseafloor habitat following a deep-sea volcanic eruption. FEMS Microbiol Ecol **43**: 393–409.

Huber, J.A., Butterfield, D.A., and Baross, J.A. (2002) Temporal changes in archaeal diversity and chemistry in a mid-ocean ridge subseafloor habitat. Appl Environ Microbiol **68**: 1585–1594.

Huber, J.A., Mark Welch, D.B., Morrison, H.G., Huse, S.M., Neal, P.R., Butterfield, D.A., and Sogin, M.L. (2007) Microbial population structures in the deep marine biosphere. Science **318**: 97–100.

Jiang, S.C. and Paul, J.H. (1998) Gene transfer by transduction in the marine environment. Appl Environ Microbiol **64**: 2780–2787.

Jiao, N., Herndl, G.J., Hansell, D.A., Benner, R., Kattner, G., Wilhelm, S.W., *et al.* (2010) Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. Nat Rev Microbiol **8**: 593–9.

Johnson, H.P. and Pruis, M.J. (2003) Fluxes of fluid and heat from the oceanic crustal reservoir. Earth Planet Sci Lett **216**: 565–574.

Jones, S.E. and Lennon, J.T. (2010) Dormancy contributes to the maintenance of microbial diversity. Proc Natl Acad Sci U S A **107**: 5881–6.

Jørgensen, B.B. and D'Hondt, S. (2006) Ecology. A starving majority deep beneath the seafloor. Science **314**: 932–4.

Kelley, D.S., Baross, J.A., and Delaney, J.R. (2002) Volcanoes, fluids, and life at mid-ocean ridge spreading centers. Annu Rev Earth Planet Sci **30**: 385–491.

Kelley, D.S., Karson, J.A., Blackman, D.K., Früh-Green, G.L., Butterfield, D.A., Lilley, M.D., *et al.* (2001) An off-axis hydrothermal vent field near the Mid-Atlantic Ridge at 30 degrees N. Nature **412**: 145–9.

Koonin, E. V and Martin, W. (2005) On the origin of genomes and cells within inorganic compartments. Trends Genet **21**: 647–54.

Koonin, E. V, Senkevich, T.G., and Dolja, V. V (2006) The ancient Virus World and evolution of cells. Biology Direct **1**: 29.

Kristensen, D.M., Mushegian, A.R., Dolja, V. V, and Koonin, E. V (2009) New dimensions of the virus world discovered through metagenomics. Trends Microbiol **18**: 11–19.

Lane, N., Allen, J.F., and Martin, W. (2010) How did LUCA make a living? Chemiosmosis in the origin of life. BioEssays **32**: 271–80.

Lang, A.S. and Beatty, J.T. (2000) Genetic analysis of a bacterial genetic exchange element: the gene transfer agent of *Rhodobacter capsulatus*. Proc Natl Acad Sci U S A **97**: 859–64.

Lang, S.Q., Butterfield, D.A., Schulte, M., Kelley, D.S., and Lilley, M.D. (2010) Elevated concentrations of formate, acetate and dissolved organic carbon found at the Lost City hydrothermal field. Geochim Cosmochim Acta **74**: 941–952.

Lennon, J.T. and Jones, S.E. (2011) Microbial seed banks: the ecological and evolutionary implications of dormancy. Nat Rev Microbiol **9**: 119–30.

Lindell, D., Jaffe, J.D., Johnson, Z.I., Church, G.M., and Chisholm, S.W. (2005) Photosynthesis genes in marine viruses yield proteins during host infection. Nature **438**: 86–89.

Mann, N.H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (1993) Bacterial photosynthesis genes in a virus. Environ Microbiol **59**: 3736–3743.

Martin, W., Baross, J., Kelley, D., and Russell, M.J. (2008) Hydrothermal vents and the origin of life. Nat Rev Microbiol **6**: 805–14.

Martin, W. and Russell, M.J. (2007) On the origin of biochemistry at an alkaline hydrothermal vent. Philos T Roy Soc B **362**: 1887–925.

Matson, E.G., Thompson, M.G., Humphrey, S.B., Zuerner, R.L., and Stanton, T.B. (2005) Identification of genes of VSH-1, a prophage-like gene transfer agent of *Brachyspira hyodysenteriae*. J Bacteriol **187**: 5885–5892.

McDaniel, L.D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K.B., and Paul, J.H. (2010) High frequency of horizontal gene transfer in the oceans. Science **330**: 50.

Middelboe, M. (2000) Bacterial growth rate and marine virus-host dynamics. Microb Ecol **40**: 114–124.

Middelboe, M., Glud, R.N., Wenzhöfer, F., Oguri, K., and Kitazato, H. (2006) Spatial distribution and activity of viruses in the deep-sea sediments of Sagami Bay, Japan. Deep-Sea Res Pt I **53**: 1–13.

Molin, S. and Tolker-Nielsen, T. (2003) Gene transfer occurs with enhanced efficiency in biofilms and induces enhanced stabilisation of the biofilm structure. Curr Opin Biotechnol **14**: 255–261.

O'Malley, M.A. (2007) The nineteenth century roots of "everything is everywhere". Nat Rev Microbiol **5**: 647–51.

Ortmann, A.C. and Suttle, C.A. (2005) High abundances of viruses in a deep-sea hydrothermal vent system indicates viral mediated microbial mortality. Deep-Sea Res Pt I **52**: 1515–1527.

Paul, J.H. (2008) Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? ISME J **2**: 579–589.

Pedrós-Alió, C. (2012) The Rare Bacterial Biosphere. Annu Rev Mar Sci **4**: 449–466.

Pietilä, M.K., Roine, E., Paulin, L., Kalkkinen, N., and Bamford, D.H. (2009) An ssDNA virus infecting archaea: a new lineage of viruses with a membrane envelope. Mol Microbiol **72**: 307–19.

Poole, A.M. (2009) Horizontal gene transfer and the earliest stages of the evolution of life. Res Microbiol **160**: 473–80.

Prangishvili, D., Forterre, P., and Garrett, R.A. (2006) Viruses of the Archaea: a unifying view. Nat Rev Microbiol **4**: 837–848.

Proskurowski, G., Lilley, M.D., Seewald, J.S., Früh-Green, G.L., Olson, E.J., Lupton, J.E., *et al.* (2008) Abiogenic hydrocarbon production at lost city hydrothermal field. Science **319**: 604–7.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., *et al.* (2004) The 1.2-megabase genome sequence of Mimivirus. Science **306**: 1344–50.

Resch, A., Fehrenbacher, B., Eisele, K., Schaller, M., and Götz, F. (2005) Phage release from biofilm and planktonic *Staphylococcus aureus* cells. FEMS Microbiol Lett **252**: 89–96.

Reysenbach, A.-L. and Cady, S.L. (2001) Microbiology of ancient and modern hydrothermal systems. Trends Microbiol **9**: 79–86.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B.B., Rodriguez-Brito, B., Pasic, L., Thingstad, T.F., Rohwer, F., *et al.* (2009) Explaining microbial population genomics through phage predation. Nat Rev Microbiol **7**: 828–36.

Rosario, K. and Breitbart, M. (2011) Exploring the viral world through metagenomics. Curr Opin Virol **1**: 289–297.

Russell, M.J. and Hall, A.J. (1997) The emergence of life from iron monosulphide bubbles at a submarine hydrothermal redox and pH front. J Geol Soc London **154**: 377–402.

Schrenk, M.O., Kelley, D.S., Delaney, J.R., and Baross, J.A. (2003) Incidence and diversity of microorganisms within the walls of an active deep-sea sulfide chimney. Appl Environ Microbiol **69**: 3580–3592.

La Scola, B., Audic, S., Robert, C., Jungang, L., de Lamballerie, X., Drancourt, M., *et al.* (2003) A giant virus in amoebae. Science **299**: 2033.

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., *et al.* (2008) The virophage as a unique parasite of the giant mimivirus. Nature **455**: 100–4.

Skraber, S., Schiven, J., Gantzer, C., and de Roda Husman, A.M. (2005) Pathogenic viruses in drinking-water biofilms: A public health risk? Biofilms **2**: 105–117.

Slobodkin, A., Campbell, B., Cary, S.C., Bonch-Osmolovskaya, E., and Jeanthon, C. (2001) Evidence for the presence of thermophilic Fe(III)-reducing microorganisms in deep-sea hydrothermal vents at 13 degrees N (East Pacific Rise). FEMS Microbiol Ecol **36**: 235–243.

Smith, J.M. and Szathmáry, E. (1997) The major transitions in evolution. Oxford University Press, Oxford.

Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R., *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A **103**: 12115–20.

Stanton, T.B. (2007) Prophage-like gene transfer agents–Novel mechanisms of gene exchange for Methanococcus, Desulfovibrio, Brachyspira, and Rhodobacter species. Anaerobe **13**: 43–49.

Steward, G.F., Montiel, J.L., and Azam, F. (2000) Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. Limnol Oceanogr **45**: 1697–1706.

Stoodley, P., Sauer, K., Davies, D.G., and Costerton, J.W. (2002) Biofilms as complex differentiated communities. Annu Rev Microbiol **56**: 187–209.

Summit, M. and Baross, J.A. (2001) A novel microbial habitat in the mid-ocean ridge subseafloor. Proc Natl Acad Sci U S A **98**: 2158–63.

Suttle, C.A. (2007) Marine viruses--major players in the global ecosystem. Nat Rev Microbiol **5**: 801–12.

Suttle, C.A. (2005) Viruses in the sea. Nature **437**: 356–361.

Takai, K. and Horikoshi, K. (1999) Genetic diversity of archaea in deep-sea hydrothermal vent environments. Genetics **152**: 1285–1297.

Takai, K., Komatsu, T., Inagaki, F., and Horikoshi, K. (2001) Distribution of archaea in a black smoker chimney structure. Appl Environ Microbiol **67**: 3618–29.

Thingstad, T. and Lignell, R. (1997) Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. Aquat Microb Ecol **13**: 19–27.

Thoulouze, M.-I. and Alcover, A. (2011) Can viruses form biofilms? Trends Microbiol **19**: 257–62.

Thurber, R.V. (2009) Current insights into phage biodiversity and biogeography. Curr Opin Microbiol **12**: 582–7.

Vidgen, M., Carson, J., Higgins, M., and Owens, L. (2006) Changes to the phenotypic profile of Vibrio harveyi when infected with the *Vibrio harveyi* myovirus-like (VHML) bacteriophage. J Appl Microbiol **100**: 481–7.

Wachtershauser, G. (1990) Evolution of the first metabolic cycles. Proc Natl Acad Sci U S A **87**: 200–204.

Wächtershäuser, G. (1988) Before enzymes and templates: theory of surface metabolism. Microbiol Rev **52**: 452–84.

Wächtershäuser, G. (1988) Pyrite Formation, the first energy source for life: a hypothesis. Sys Appl Microbiol **10**: 207–210.

Waldor, M.K. and Mekalanos, J.J. (1996) Lysogenic conversion by a filamentous phage encoding cholera toxin. Science **272**: 1910 –1914.

Webb, J.S., Givskov, M., and Kjelleberg, S. (2003) Bacterial biofilms: prokaryotic adventures in multicellularity. Curr Opin Microbiol **6**: 578–585.

Webb, J.S., Lau, M., and Kjelleberg, S. (2004) Bacteriophage and phenotypic variation in *Pseudomonas aeruginosa* biofilm development. J Bacteriol **186**: 8066–73.

Weinbauer, M.G. (2004) Ecology of prokaryotic viruses. FEMS Microbiol Rev **28**: 127–81.

Whiteley, M., Bangera, M.G., Bumgarner, R.E., Parsek, M.R., Teitzel, G.M., Lory, S., and Greenberg, E.P. (2001) Gene expression in *Pseudomonas aeruginosa* biofilms. Nature **413**: 860–864.

Williamson, S.J., Cary, S.C., Williamson, K.E., Helton, R.R., Bench, S.R., Winget, D., and Wommack, K.E. (2008) Lysogenic virus–host interactions predominate at deep-sea diffuse-flow hydrothermal vents. ISME J **2**: 1112–1121.

Williamson, S.J., Rusch, D.B., Yooseph, S., Halpern, A.L., Heidelberg, K.B., Glass, J.I., *et al.* (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS ONE **3**: e1456.

Woese, C. (1998) The universal ancestor. Proc Natl Acad Sci U S A **95**: 6854–6859.

Woese, C.R. (2002) On the evolution of cells. Proc Natl Acad Sci U S A **99**: 8742–7.

Ziebuhr, W., Krimmer, V., Rachid, S., Lossner, I., Gotz, F., and Hacker, J. (1999) A novel mechanism of phase variation of virulence in *Staphylococcus epidermidis*: evidence for control of the polysaccharide intercellular adhesin synthesis by alternating insertion and excision of the insertion sequence element IS256. Mol Microbiol **32**: 345–356.

CHAPTER ONE

Microbial community structure across fluid gradients in the Juan de Fuca Ridge hydrothermal system[2]

Summary

Physical and chemical gradients are dominant factors in shaping hydrothermal vent microbial ecology, where archaeal and bacterial habitats encompass a range between hot, reduced hydrothermal fluid and cold, oxidized seawater. To determine the impact of these fluid gradients on microbial communities inhabiting these systems, we surveyed bacterial and archaeal community structure among and between hydrothermal plumes, diffuse flow fluids, and background seawater in several hydrothermal vent sites on the Juan de Fuca Ridge using 16S rRNA gene diversity screening (clone libraries and terminal restriction length polymorphisms) and quantitative polymerase chain reaction methods. Community structure was similar between hydrothermal plumes and background seawater, where a number of taxa usually associated with low-oxygen zones were observed, whereas high-temperature diffuse fluids exhibited a distinct phylogenetic profile. SUP05 and Arctic96BD-19 sulfur-oxidizing bacteria were prevalent in all three mixing regimes where they exhibited overlapping but not identical abundance patterns. Taken together, these results indicate conserved patterns of redox-driven niche partitioning between hydrothermal mixing regimes and microbial communities associated with sinking particles and oxygen-deficient waters. Moreover, the prevalence of SUP05 and Arctic96BD-19 in plume and diffuse flow fluids indicates a more cosmopolitan role for these groups in the ecology and biogeochemistry of the dark ocean.

Introduction

Hydrothermal vents are dynamic, gradient-dominated ecosystems supporting high levels of microbial production and consumption (Orcutt *et al.*, 2011; McCollom, 1997). Water-rock reactions deep in the subsurface generate high-temperature fluids that emerge at the crust-water interface, causing the precipitation of minerals to form sulfide structures. These reduced hydrothermal fluids mix with background seawater, creating

[2] Previously published in FEMS Microbiology Ecology, Vol. 83, p. 324-339, 2013.

gradients of temperature, pH and chemical composition. In basalt-hosted vent ecosystems, such as those found along mid-ocean ridges, diffuse flows often emerge from the seafloor at porous regions alongside sulfide chimneys, and comprise a mixture of high-temperature fluid and deep seawater.

The microbial community structure of diffuse flow fluids is thought to comprise a diverse mixture of cells originating in the deep hot biosphere and the surrounding seawater interface. Though temperatures are usually below 50˚C, some portion of the archaeal and bacterial groups sampled in diffuse flow fluids are thought to represent the subsurface microbial community that has been flushed up by the vent fluid (Huber *et al.*, 2002, 2003, 2006; Opatkiewicz *et al.*, 2009). As a result, diffuse flow fluids generally harbor thermophilic and hyperthermophilic organisms, with particularly high diversity in the *Thermococcales* (Huber *et al.*, 2006) and *Epsilonproteobacteria* (Nakagawa *et al.*, 2005). In the deep subsurface, it is thought that a high-temperature environment dominated by hydrothermal fluid, hosting largely thermophilic organisms, gives way to a more seawater-dominated regime populated by more mesophilic organisms (Huber *et al.*, 2003.) Variations in temperature as well as the availability of different carbon sources and electron acceptors and donors result in variations in the composition of the microbial community, depending on its situation within the gradient. These mixing gradients result in high microbial diversity in diffuse fluids. Huber and colleagues (2007) used high resolution pyrotag sequencing targeting the small subunit ribosomal RNA (SSU or 16S rRNA) gene to identify more than 3,000 archaeal and 37,000 bacterial operational taxonomic units from diffuse flow fluids at Axial Seamount, located on the Juan de Fuca Ridge off the coast of Washington and Oregon, with different population structures at each vent reflecting different geochemical regimes.

Geochemical gradients extend beyond the subsurface seawater interface to encompass hydrothermal plumes, whose native microbial communities are not as well-characterized. Hydrothermal plumes, which rise tens to hundreds of meters above the seafloor, are formed when high-temperature hydrothermal fluid is emitted from vent structures into background seawater. The reduced hydrothermal fluid entrains deeper seawater as it rises, emerging as a neutrally buoyant plume above the vent field. An estimated 70% of plume waters are entrained from deep seawater, 30% donated from seawater at the depth of the plume and ≤0.01% retained from the hydrothermal source (Lupton *et al.*, 1985). Despite the minor contribution of hydrothermal vent fluid to the rising plume, the presence of trace metals (iron, manganese), gases (hydrogen, hydrogen sulfide, methane), and other reduced substrates create new niche spaces that are distinct from the surrounding background seawater. These substrates are removed from the plume sequentially, with manganese remaining as one of the longest-lasting plume signatures in the water column (Kadko *et al.*, 1990).

Several studies have attempted to assess the metabolism of the microbial communities based on the removal of specific compounds from hydrothermal plumes. High rates of methane oxidation measured in plumes above the Main Endeavour Segment on the Juan de Fuca Ridge were linked to methanotrophic bacteria that appeared to be entrained in the plume from deep seawater (DeAngelis *et al.*, 1993). Similarly, high rates of ammonia oxidation in plumes above the Main Endeavour Segment were influenced by the presence of organic particles and abundant particle-associated ammonia-oxidizing bacteria (Lam *et al.*, 2008). High temperature cultivation of manganese-oxidizing bacteria from a plume in the Guaymas Basin indicated hydrothermal fluid entrainment of subsurface bacteria into the water column (Dick & Tebo, 2010), and more recent time-series analysis of hydrothermal vent plumes in the East Pacific Rise based on automated rRNA intergenic spacer analysis (ARISA) identified contrasting microbial community structures associated with changes in the particle composition of the plume (Sylvan *et al.*, 2012).

Nevertheless, no study has yet assessed hydrothermal plume and diffuse flow community structure within the context of mixing regimes and fluid gradients. Here, we charted the microbial community structure of three different mixing regimes within the hydrothermal fluid-background seawater gradient: diffuse flow fluids (with a high input of hydrothermal fluid), hydrothermal plumes (with a minimal input of hydrothermal fluid), and background seawater (with little to no hydrothermal input) using a combination of small subunit ribosomal RNA (SSU rRNA or 16S rRNA) gene clone library sequencing, terminal restriction fragment length polymorphism (T-RFLP) analysis and quantitative polymerase chain reaction (qPCR). Our results show that while some groups are confined to very specific regimes within these mixing gradients, the sulfur oxidizing Gammaproteobacteria of both the SUP05 and Arctic96BD-19 groups are predominant across all three sample types, suggesting that these groups are more cosmopolitan than most other phylotypes in these habitats.

Materials and Methods

Sample collection

Samples were collected aboard the R/V Atlantis in June 2009 at both the Main Endeavour Field and at Axial Seamount on the Juan de Fuca Ridge. Four samples each of diffuse flow fluid and hydrothermal vent plume water as well as one background seawater sample were used in this study. Maps of these regions are shown in Figure 1.1. Diffuse flow samples were collected with a hydrothermal fluid and particle sampler (HFPS) (Butterfield *et al.*, 2004) aboard DSV Alvin. At each sample site, 2.4 to 3.8L of fluid were pumped through 0.22 μm Sterivex filters (Millipore, USA) mounted on the HPFS on the submarine. Upon shipboard recovery, filters were placed in sterile 50-mL Falcon tubes (BD Sciences Labware), and frozen at -80˚C. Diffuse flow temperatures were measured on the HFS as the fluids from Grotto, Easter Island, and Lobo vents in the Main Endeavour Field were sampled; in all cases, the average temperature from the entire sampling process is reported. The Hulk diffuse flow sample was collected using a 200L barrel sampler that was lowered to the seafloor on an elevator. The barrel sampler setup and sample processing is described in Anderson *et al.* (2011). Hulk diffuse flow fluids were filtered through four Steripaks, also with a 0.22 μm pore size, and frozen at -80˚C.

Hydrothermal vent plume samples were detected on the basis of temperature or transmissivity anomalies and collected using a Niskin bottle rosette mounted on a conductivity-temperature-depth profiler (CTD) (Seabird). In a 4˚C cold room on board the ship, 2L of fluid from each Niskin bottle containing plume fluids were filtered through a 0.22 μm Sterivex filter, then placed in a sterile 50-mL Falcon tube and frozen at -80˚C until further processing onshore. Plume samples were collected in this manner above Needle vent in the Main Endeavour Field, and above Castle vent and the CASM field at Axial Seamount. For the plume sample taken above Hulk vent, 50 L of plume fluid were filtered through a 0.22 μm Steripak (Millipore, USA), then placed in a sterile 50 mL Falcon tube and frozen at -80˚C for further analysis. A background sample (no detectable plume) was taken south of the Main Endeavour Field at 47˚ 56.00'N, 129˚ 04.30' W. Sterivex filter samples were collected with a Niskin bottle rosette as described above.

18ml fluid subsamples were taken from each sample site for cell counting. Formaldehyde (3.7% final concentration) was added to each fluid sample and placed in a 20mL scintillation vial, which was placed at 4˚C while on shipboard. Onshore, cells were enumerated on an epifluorescence microscope (Zeiss) using DAPI (4P,6-diamidino-2-phenylindole) (Sigma). At least 200 cells and 20 fields of view were counted for cell quantification.

DNA extraction and purification

DNA was extracted from Sterivex filters using a modified procedure from Huber *et al.* (2002). Briefly, DNA extraction buffer (0.1M Tris-HCl, 0.2M Na-EDTA, 0.1M NaH_2PO_4, 1.5M NaCl, and 1% cetyltrimethylammonium bromide) was added to each filter. Filters were capped with Medex caps (MedEx Supply) and sealed with parafilm. Sterivexes were freeze-thawed five times by alternating between a slurry of ethanol and dry ice and a 65˚C water bath. 36µL of 50mg/ml lysozyme was then added and the filter incubated at 37˚C for 30 minutes. 45µL proteinase K (1%) and 90µl SDS solution (20%) were then added and the filter incubated at 65˚C on a shaker for 1.5 hours. Lysate was removed from Sterivex filters and centrifuged for 5 minutes at 6000g. DNA was extracted from the supernatant using phenol:chloroform:isoamyl alcohol and chloroform:isoamyl alcohol as described in Huber *et al.* (2002). For DNA extraction from Steripaks, the filter units were freeze-thawed three times by alternating between a -80˚C freezer and a 60˚C oven. DNA was extracted as described above, but scaled up to accommodate for larger volumes.

Clone library construction

Clone libraries were constructed from the Hulk diffuse flow, Hulk plume, and background samples. Bacterial and archaeal 16S rRNA genes were amplified for clone library construction with GoTaq DNA polymerase (Promega) using universal bacterial primers 8Fb (Edwards *et al.*, 1989) (5'-AGAGTTTGATCCTGGCTCAG-3') and

BAC1492R (Stackebrandt & Liesack, 1993) (5'-RGYTACCTTGTTACGACTT-3') and universal archaeal primers ARC21F (DeLong, 1992) (5-TTCYGGTTGATCCYGCCRGA-3') and ARC922R (Opatkiewicz *et al.*, 2009) (5'-YCCGGCGTTGANTCCAATT-3'). For PCR amplification, an initial denaturing step of 94˚C for 5 min was followed by 30 cycles of 94˚C for 30 s, 45˚C for 30 s, and 72˚C for 2 min for bacteria, followed by a 72˚C extension step for 10 min. Attempts at amplification of the target region at annealing temperatures above 45˚C were unsuccessful, and so this temperature was used for all samples. For archaea, the annealing temperature was 55˚C, and only 24 cycles were used. To minimize the formation of heteroduplex molecules, PCR products were reconditioned prior to cloning by using PCR product as template in a new PCR cocktail and repeating the thermocycling protocol for 5-10 cycles (Thompson *et al.*, 2002).

Bacterial PCR products were cloned using the TOPO TA cloning kit (Invitrogen), and archaeal PCR products with the StrataClone PCR cloning kit (Agilent Technologies) according to manufacturer's instructions. Clones were amplified with the M13F (5'-GTAAAACGACGGCCAG-3') and M13R (5'-CAGGAAACAGCTATGAC-3') primers. Sequencing was conducted through the University of Washington High Throughput Genomics Center on an ABI 3730xl sequencing unit (Applied Biosystems). Primers used for sequencing included T3 (5'-ATTAACCCTCACTAAAGGGA-3') and T7 (5'-TAATACGACTCACTATAGGG-3'), and either 515Fb (5'-GTGCCAAGCMGCCGCGGTAA-3'), 907Rb (5'-CCGTCAATTCMTTTRAGTTT-3'), or 110R (5'-GGGTTGCGCTCGTTG-3') for bacterial clones, and 515Fa (5'-GTGGCASCMGCCGCGGTAA-3') for archaeal clones. Contigs were assembled using Sequencher 4.9 (Gene Codes Corporation). Sequences were aligned and checked for chimeras using Greengenes (DeSantis *et al.*, 2006). Taxon assignment was performed based on blastn queries against the Greengenes (DeSantis *et al.*, 2006) and ARB (Ludwig *et al.*, 2004) 16S rRNA databases. Resulting outputs were summarized in table format and visualized as a dot plot using the custom perl script, bubble.prl (http://www.cmde.science.ubc.ca/Hallam/bubble.php). Clone library diversity indices were calculated using mothur (Schloss *et al.*, 2009) at a clustering distance of 0.03.

Phylogenetic Analysis

Trees were constructed by aligning clone sequences and representative sequences from the NCBI database using the Greengenes pipeline (DeSantis *et al.*, 2006), then importing into ARB for comparison with reference sequences. Evolutionary history was inferred using the maximum likelihood method based on the Tamura-Nei model (Tamura & Nei, 1993). 1000 replicates were used in the bootstrap test (Felsenstein 1985). Initial trees for the heuristic search were obtained automatically as follows: when the number of common sites was <100 or less than one fourth of the total number of sites, the maximum parsimony was used; otherwise the BIONJ method with MCL distance matrix was used. All phylogenetic analyses were implemented in MEGA5 (Tamura *et al.*, 2011).

T-RFLP community profiling

Bacterial and archaeal 16S rRNA genes were amplified for T-RFLP using universal primers ARC21F (5'- [6-FAM]TTCYGGTTGATCCYGCCRGA-3'), ARC922R (5'-YCCGGCGTTGANTCCAATT-3'), BAC68F (5'-[6-FAM]TNANACATGCAAGTCGRRCG-3') and BAC1492R (5'-RGYTACCTTGTTACGACTT-3'). Each PCR reaction (25µL) contained 1X GoTaq buffer (Promega), 1U GoTaq polymerase (Promega) 2mM $MgCl_2$, 2mM dNTPs, and 0.4µM each primer. An initial denaturation step of 94˚C for 5 min was followed by 34 cycles of 94˚C for 30 s, 55˚C for 45 s, and 72˚C for 2 min for bacteria, with a final extension of 72˚C for 10 min. For archaea, the annealing step was 55˚C for 30 s, and only 23 cycles were used. To minimize PCR drift (Polz & Cavanaugh, 1998), between 5-10 replicate reactions were pooled and then cleaned and concentrated with QiaQuick PCR purification columns (Qiagen) according to the manufacturer's instructions.

Cleaned PCR products were digested with restriction endonucleases HaeIII or BstUI for all samples, plus MspI for bacterial PCR products and RsaI for archaeal PCR products. All digests were incubated overnight at 37˚C, except for BstUI, which was incubated at 60˚C. Digests were inactivated by freezing the solution at -20˚C. Samples were ethanol precipitated, dried, and resuspended in 0.25 µL ET900-R MegaBACE size standard, 4.75µL 70% formamide/1mM EDTA loading buffer, and 5µL water. T-RFLP profiling runs were conducted on a MegaBACE 1000 (GE LifeSciences).

T-RFLP profiles were processed using DAx (2006 Van Mierlo Software Consultancy, the Netherlands). Peaks were standardized using the variable percentage threshold method (Osborne *et al.*, 2006), then normalized between each sample according to peak height. Peaks were binned into 8 different bin shifts according to the method outlined by Hewson and Fuhrman (2006). Bins were 4bp wide, and shifted by 0.5bp for each bin shift. For diversity indices, the average of the 8 bin shifts was used, and indices were calculated using EstimateS (Version 8.2, R.K. Colwell). To create resemblance matrices, the maximum similarity of each of the 8 bin shifts was calculated using EstimateS and hierarchical clustering dendrograms were created using the group average method in PRIMER-E v.6.1.6 (Clarke and Gorley, 2006). Cophenetic correlation coefficients were determined by calculating the correlation coefficient between the resemblance matrix and the cophenetic matrix created by clustering. To identify T-RFLP peaks, clone sequences were trimmed in BioEdit 7.0.9 (Ibis Biosciences) and digested *in silico* using the program REPK (Collins and Rocap, 2007). Resulting fragments were compared to fragments obtained from restriction digests; peaks were positively identified if they fell within 2 bp of an *in silico* clone fragment.

Quantitative polymerase chain reaction

Relative percentages of SUP05 and Arctic96BD-19 compared to total bacteria were determined using qPCR. Total bacteria were quantified using a bacteria-specific forward primer (27F, 5′-AGAGTTTGATCCTGGCTCAG-3’) and a universal reverse primer (DW519R, 5′-GNTTTACCGCGGCKGCTG-3’) (Zaikova *et al.*, 2010). SUP05 was quantified using a bacteria-specific forward primer (Ba519F, 5′-CAGCMGCCGCGGTAANWC-3’) and a group-specific reverse primer (1048R_SUP05, 5′-CCATCTCTGGAAAGTTCCGTCT-3’) (Zaikova *et al.*, 2010). Arctic96BD-19 was quantified using Ba519F and a group-specific primer, 1048R_Arctic (5'-CTATTTCTAGAAAGTTCGCAGG-3’) (Walsh & Hallam, 2011). Each 20μl reaction contained 2μl sterile DNAse free water, 2μl each of 5 μM forward and reverse primers, 4μl template, and 10μl SsoFast EvaGreen Supermix (Bio-Rad Laboratories, California, USA). Reactions were carried out in 48 well white plates with optical caps (Bio-Rad). Reactions were run on a MiniOpticon Real-Time PCR System (Bio-Rad). Universal bacterial primers were run with the following protocol: initial denaturation at 95˚C for 3 min, followed by 45 cycles of 95˚C for 20 sec, primer annealing for 30 sec at 55˚C for total bacteria, 63˚C for SUP05, and 59˚C for Arctic96BD-19, and a plate read. The melt curve extended from 55-95˚C, increasing by 0.5˚C per sec. Data was analyzed with the CFX Manager for the MiniOpticon system (BioRad). A standard curve was created for each of the primer sets and run in parallel with each of the samples. A 10-fold dilution series of standards ranging from 4.3 x 10^2 to 4.3 x 10^5 (Arctic96BD-19) or 8.5 x 10^2 to 8.5 x 10^5 (SUP05) was prepared for each run. These standards were also used for quantification of total bacteria. To mitigate the impact of inhibitors (Lloyd *et al.*, 2010), samples were run at either 1/10 or 1/100 dilutions, and the dilution level was kept consistent for each sample across each of the three primer sets. All samples were run in duplicate, and ratios of SUP05 or Arctic96BD-19 to total bacteria were carried out by averaging all four ratio combinations from each set of duplicates. Standard error of the percent abundances were calculated from the standard error of all four ratio combinations.

Nucleotide sequence accession numbers.

The GenBank nucleotide sequence accession numbers for the sequences in this study are JQ678046 through JQ678591.

Results

A total of 4 diffuse flow, 4 plume and 1 background seawater samples were collected from the Main Endeavour Field and at Axial Seamount on the Juan de Fuca Ridge in June 2009. Sample number, location, temperature, and cell count data are listed in Table 1.1. The concentration of Mg (33.3 mmol/kg) and dissolved silica (6.38 mmol/kg) of the diffuse flow samples from Hulk vent indicate an average temperature of 125˚C (Anderson *et al.*, 2011). This fluid sample also had the highest cell counts (Table 1.1). The unique characteristics of this sample can be partially attributed to the nature of the sampling method: a funnel was attached to the sample hose on a barrel sampler, which was placed atop a region of diffuse flow covered in tube worms. The strong suction of the barrel sampler may have sealed the funnel onto the surface of the vent and drawn out higher temperature water from within the sulfide structure. The nozzle on the HFS, in contrast, collected samples consistent with the temperatures measured within the animal communities on the surface of the vents.

Clone libraries

16S rRNA gene clone libraries were constructed to compare microbial community composition in diffuse flow fluid (with a high input of hydrothermal fluid), hydrothermal plume (with a minimal input of hydrothermal fluid), and background seawater (with little to no hydrothermal input). Both the diffuse flow and plume clone libraries were constructed from samples taken at Hulk vent to compare fluids from the same vent structure; it should be noted that the diffuse flow sample from Hulk vent was at an extremely high temperature and therefore represents the extreme end of the spectrum within the mixing regime of hydrothermal fluid and seawater.

Proteobacteria dominated 16S rRNA gene clone libraries across all three mixing regimes (Figure 1.2). However, various subdivisions within the Proteobacteria exhibited distinct distribution patterns between samples. In both vent plume and background seawater *Alpha*, *Gamma*, and *Deltaproteobacteria*, including the SAR11 cluster, Agg47, Hyd24-01, SUP05, Arctic96BD-19 and ZD0417, *Myxococcales* and SAR324 respectively, were prevalent. SUP05 bacteria were most abundant in the vent plume sample, contributing 39% of total bacterial clones (Figure 1.3). ZA3420c, Arctic96B-1, *Geobacter* and NB1-I were recovered solely from background seawater. Within the diffuse flow sample *Alpha*, *Beta*, *Gamma*, and *Epsilonproteobacteria*, including the SAR11 cluster, *Sphingomonas, Comamonas, Alteromonas, Pseudomonas,* and *Caminibacter* were prevalent. In addition to Proteobacteria, *Microthrix*, *Chloroflexi* and Marine Group A were also recovered from vent plume and background seawater (Figure 1.2). Groups affiliated with the *Bacteroidetes* were recovered from background seawater and diffuse flow samples and groups affiliated with *Planctomycetes* and *Verrucomicrobia* were recovered from vent plume, diffuse flow and background seawater samples (Figure 1.2).

In the case of the archaeal domain, two major lineages affiliated with Marine Group I and II archaea were recovered from vent plume and background seawater samples. The proportion of Marine Group I and Marine Group II archaea was similar in both samples, with Marine Group I contributing 69.5% and 66.7% and Marine Group II 18.9% and 20.8% of total archaeal clones to vent plume and background seawater, respectively (Figure 1.2). In contrast, *Methanococci* and *Thermococci* were exclusively identified in diffuse flow fluids, where they comprised 8.5% and 85.1% of total archaeal clones respectively (Figure 1.2). Groups affiliated with Marine Group I, Marine Group II and *Thermoprotei* were also recovered from the diffuse flow sample, ranging between 1 to 2% of total archaeal clones. Raw values used for tabulating Figure 1.2 are listed in Table 1.2.

T-RFLP community profiles

To determine whether community composition patterns recovered in clone libraries from the Hulk vent were representative for the Juan de Fuca Ridge system, T-RFLP profiles for archaeal and bacterial 16S rRNA genes were obtained across multiple diffuse flow and plume samples from different vent locations (Table 1.1). Peaks were identified based on *in silico* digestion of clone library sequences (see methods).

Archaeal T-RFLP profiles indicated that the majority of plume and diffuse flow samples were strikingly similar, with Marine Groups I and II dominating the community structure (Figure 1.4A). Plume T-RFLP traces, such as the Castle trace shown in Figure 1.4A, were characteristic of all archaeal samples digested with RsaI, with the exception of the Hulk diffuse flow sample. This extremely high-temperature sample, shown in Figure 1.4B, exhibited a unique community profile, with *Thermococcus* and *Methanocaldococcus* groups dominating the community. T-RFLP community profiling with other restriction enzymes indicated similar patterns. Clustering of T-RFLP profiles based on the Chao abundance-based Jaccard Index (Figure 1.4C) indicate that the Hulk diffuse flow sample was less than 20% similar to all other samples, with most of the other plume and cooler diffuse flow samples clustering together at over 70% similarity, though some heterogeneity is evident in the Hulk plume and background seawater samples.

Bacterial profiles exhibited a much higher degree of variation based primarily on differences in the presence or height of minor peaks. In samples digested with HaeIII, a 369-bp peak corresponding to sulfur-oxidizing Gammaproteobacteria groups SUP05 and

Arctic96BD-19 was visible in all but two of the T-RFLP traces. Examples of this peak can be seen in the plume sample from Needle, (Figure 1.5A) and in the diffuse flow sample from Easter Island (Figure 1.5B). Diffuse flow samples isolated from Hulk and Grotto vents, the two highest-temperature samples, were the only samples lacking this peak. Samples digested with restriction enzymes MspI and BstUI did not resolve a peak unique to SUP05 or Arctic96BD-19, but community similarity analyses from these samples did indicate trends corresponding to the type of environment from which samples were taken. A community similarity cluster dendrogram based on samples digested with BstUI, (Figure 1.5C) and shows that the very high-temperature Hulk diffuse flow sample clustered separately from all other samples at a very low level of similarity. Other samples clustered roughly according to the temperature at which they were sampled: diffuse flow samples Lobo and Grotto clustered together at about 70% similarity, while the plume samples and the Easter Island diffuse flow sample clustered together at about 60% similarity. Easter Island was one of the two lower-temperature diffuse flow samples taken for this study. Plume samples from CASM and Castle vents, with the lowest temperature anomalies of the plumes sampled here, clustered together as well, while the background seawater sample clustered at a low level of similarity with the cooler diffuse flow and plume samples.

Diversity indices calculated for both clone libraries and T-RFLP profiles indicated that the hydrothermal plume and diffuse flow samples, in general, had higher diversity than background seawater for archaea across all diversity indices calculated (S_{obs}, Chao1, and Jackknife) (Table 1.3). Within the hydrothermal plume samples, the two plume samples with a higher temperature anomaly (Hulk and Needle) had higher diversity than the plume samples with a lower temperature anomaly (Castle and CASM), in both the archaeal and the bacterial domains, across all diversity indices reported. However, on the whole, no clear trend in terms of relative diversity emerged between plume and diffuse flow samples in either the bacterial or archaeal domains. Finally, while the diversity of the bacterial community in background seawater appeared to be lower according to T-RFLP analyses, clustering of clone libraries at 97% resulted in a higher number of observed OTUs in the background sample than in the other samples. This discrepancy could be the result of diversity in the 16S gene that could not be detected with the restriction enzymes used.

SUP05 and Arctic96BD-19 diversity and abundance

Given the prevalence of SUP05 and Arctic96BD-19 among and between mixing regimes, we conducted a more in-depth phylogenetic analysis based on 16S rRNA gene sequences recovered from Hulk samples and other marine ecosystems to better constrain biogeographic or ecological type (ecotype) relationships. SUP05 and Arctic96BD-19 16S rRNA gene sequences recovered from Hulk plume and background seawater samples partitioned into previously defined clades (Figure 1.6). Specifically, most of the plume clones in this cluster grouped with other SUP05 samples obtained from vent environments, such as the Suiyo Seamount (Sunamura *et al.*, 2004), or with vent endosymbionts. In contrast, the majority of background 16S rRNA gene sequences fell into the Arctic96BD-19 group, along with clones recovered from the northeastern subarctic Pacific (Walsh *et al.* 2009) and the San Pedro Channel, CA (Brown *et al.*, 2005), as well as from the Saanich Inlet (Walsh *et al.*, 2009) and the Namibian shelf (Lavik *et al.*, 2008).

We next used quantitative PCR to determine 16S rRNA gene copy numbers for SUP05 and Arctic96BD-19 in relation to total bacteria across the diffuse flow and plume samples used in T-RFLP analysis. The SUP05 group was abundant in the plume samples, reaching up to 27% of the total bacteria (Table 1.4). SUP05 16S rRNA gene copy number decreased in samples with a weaker plume signature such as CASM, at 4.1%. This was consistent with reduced recovery of SUP05 in the background sample, at 3.2%. SUP05 16S rRNA gene copy number was also high in most diffuse flow samples, reaching up to 18.7% in the Easter Island sample. However, the relative abundance decreased dramatically in the high temperature Hulk diffuse flow sample, dropping to 0.4% of total bacteria. Arctic96BD-19 16S rRNA gene copy number was high across all sample types, reaching up to 64.7% of the bacterial community in the CASM plume sample (Table 1.4). Moreover Arctic96BD-19 16S rRNA gene copy number tended to decrease as plume signatures became stronger. However, this trend did not continue for the diffuse flow samples or the background sample, where Arctic96BD-19 16S rRNA gene copy number reached up to 25.5% and 22.7%, respectively. Similar to SUP05, Arctic96BD-19 16S rRNA gene copy number decreased in the high temperature Hulk diffuse flow sample, but was still greater than that of SUP05.

Discussion

Geochemical gradients resulting from mixing between reduced, high temperature hydrothermal fluid and cooler, oxidized seawater are a dominant feature of hydrothermal vent ecosystems. Temperature, considered a proxy for chemistry in these systems, is positively correlated with sulfide and negatively correlated with oxygen (Corliss *et al.*, 1979; Johnson *et al.*, 1986). The availability of different electron acceptors and donors changes with the degree of mixing between fluid types. As a result, diffuse flow samples are enriched in reduced compounds including hydrogen, sulfide, ammonia, and iron relative to plume or background seawater; and plume samples, in which only 0.01% of the fluid is derived from a hydrothermal source, still manifest elevated concentrations of reduced compounds and metals relative to background seawater (Kadko *et al.*, 1990). By studying microbial community structure within these gradients, we can better understand patterns of redox-driven niche partitioning and adaptive radiation among and between microbial groups in the dark ocean. Samples in the current study ranged between 9-125˚C among diffuse flow samples, with temperature anomalies between 0.00-0.011 ˚C in plume samples indicating multiple different geochemical conditions. While more exhaustive geochemical analyses were not available for this particular study, some patterns were revealed in these analyses that are worthy of note, and can act as a starting point for future analyses of community partitioning across geochemical gradients in these systems.

Archaeal community composition was relatively homogeneous between samples, with similar communities found across background seawater, plume, and diffuse flow samples. Marine Group I and II archaea from diffuse flow samples collected between 9-18˚C were similar to plume and background samples, suggesting that members of these groups are adapted to a wide range of temperatures and geochemical conditions. In contrast, in the high-temperature Hulk diffuse flow sample, at 125˚C, Marine Groups I and II were almost entirely replaced by *Thermococcales* and *Methanococcales*.

Consistent with this observation, the optimal growth temperatures of *Thermococcus* and *Methanocaldococcus* strains range between 80-100˚C, confining them to a narrow range within the hydrothermal fluid-seawater gradient where the increased proportion of high-temperature hydrothermal fluids enrich for different taxa.

As with the archaeal communities, the high-temperature diffuse flow sample from Hulk was found to be unique among the bacterial communities. This sample was dominated by *Epsilonproteobacteria* in the *Caminibacter* and *Nautiliales* groups, which has been observed previously in diffuse flow fluids (Huber *et al.*, 2003; 2007). These groups appear to flourish in the warmer, more reduced fluids characteristic of higher temperatures in these vent systems. The bacterial communities in general were quite heterogeneous between samples, a trait also observed previously in vent systems (Opatkiewicz *et al.*, 2009), and may be the result of differences in vent chemistry between sites, as well as microbial endemism from one vent site to the next. Also interestingly, clustering of communities based on the relative abundance of different T-RFLP peaks indicated that samples from similar mixing regimes clustered together, providing evidence of partitioning across fluid gradients. Castle and CASM, the plume samples with the lowest temperature anomaly, tended to cluster together and also had the lowest species richness of the samples taken. The high-temperature Hulk sample, in contrast, did not necessarily exhibit higher species richness, yet clustered separately in cluster dendrograms for both the bacterial and archaeal domains. This was likely due to a compositional shift in community membership, from dominance of mesophiles such as *Gammaproteobacteria* and Marine Groups I and II in cooler plume and diffuse flow samples, to dominance of thermophiles in the *Epsilonproteobacteria*, *Thermococcales* and *Methanococcales* groups in the high temperature sample.

Despite the abundance of many different species, however, two groups of sulfur-oxidizing *Gammaproteobacteria*, SUP05 and Arctic96BD-19, were particularly abundant across all three sample types. The SUP05 and Arctic96BD-19 groups are related to sulfur-oxidizing gill symbionts of deep-sea clams and mussels (Cavanaugh, 1983; Newton *et al.*, 2007). The SUP05 lineage, initially identified in a hydrothermal vent plume originating from the Suiyo Seamount (Sunamura *et al.*, 2004), encompasses the clam and mussel symbionts, while Arctic96BD-19 forms a closely related sister clade to

SUP05. Within marine oxygen minimum zones SUP05 and Arctic96BD-19 exhibit overlapping but not identical distribution patterns consistent with redox-driven niche partitioning. Indeed, SUP05 appears to thrive in regions of sulfide and nitrate depletion, deriving energy from the oxidation of reduced sulfur compounds and using nitrate as terminal electron acceptor (Lavik *et al.*, 2008; Walsh *et al.*, 2009). In contrast, Arctic96BD-19 appears to thrive under more oxic water column conditions, deriving energy from reduced sulfur compounds and using oxygen as a terminal electron acceptor (Walsh & Hallam, 2011; Swan *et al.*, 2011). Both SUP05 and Arctic96BD-19 have the potential to harness the energy obtained from sulfur-oxidation to fix inorganic carbon via 1,5-bisphosphate carboxylase/oxygenase (RubisCO) (Walsh & Hallam, 2011; Swan *et al.*, 2011), implicating them as primary producers in the dark ocean. The extent to which they contribute to food web structures in hydrothermal vent ecosystems remains to be determined.

Phylogenetic placement of SSU rRNA gene sequences recovered from Hulk plume and diffuse flow fluids resolved into the previously recognized SUP05 and Arctic96BD-19 groups, and partitioned roughly between plume and background seawater. The SUP05 group contained the majority of hydrothermal plume-derived sequences and some background seawater sequences that grouped most closely with sequences recovered from Saanich Inlet, a seasonally anoxic basin on the coast of Vancouver Island, British Columbia and with sequences recovered from the Suiyo Seamount plume (Figure 5). In the case of Arctic96BD-19, the majority of hydrothermal plume- and background-derived sequences grouped most closely with sequences recovered from the Line-P transect in the northeastern subarctic Pacific water column (Walsh *et al.*, 2009) and clones recovered from the San Pedro Channel (Brown *et al.*, 2005). The partitioning of Juan de Fuca sequences with sequences recovered from Saanich Inlet, San Pedro, and the northeastern subarctic Pacific is consistent with gene flow between the vent ecosystem and northeastern Pacific waters as a whole.

Although no SUP05 or Arctic96BD-19 SSU rRNA gene sequences were recovered in the clone libraries from the high temperature Hulk diffuse flow fluid sample, both groups were indicated in other diffuse flow samples using T-RFLP and qPCR. Given the limited environmental parameter data and taxonomic resolution of the current study we can only begin to speculate on the forces driving ecotype selection along geochemical gradients in plume and diffuse flow fluids. While sequences matching the SUP05 group have been recovered in samples from several hydrothermal systems (*i.e.* Sunamura *et al.*, 2004, (Bourbonnais *et al.*, 2012; Dick and Tebo, 2010; German *et al.*, 2010), the relatively high abundance of Arctic96BD-19 in diffuse flow fluids (with a relatively high input of hydrothermal fluid) and hydrothermal plumes in the Juan de Fuca system had not been observed in previous studies. In some plume samples, Arctic96BD-19 dominated the bacterial community, reaching up to 64% of total bacterial SSU rRNA gene copies in the CASM plume. As previously identified members of this group appear to thrive under more oxic water column conditions than SUP05, the Arctic96BD-19 group may take advantage of elevated concentrations of reduced sulfur compounds present in attenuating plume fluids. Arctic96BD-19 was also prevalent in background seawater, where sulfide levels would be undetectable, reaching up to 22% of total bacterial SSU rRNA gene copies. It remains possible that even in such background waters, geochemical traces of the plume continue to fuel microzones of chemolithoautotrophic growth. Similar observations have been made for particles in the dark ocean with the potential to support anaerobic processes such as sulfate reduction and methanogenesis that in turn fuel chemolithoautotrophic growth in the surrounding water column (Shanks & Reeder, 1993; Karl *et al.*, 1984; Allredge & Cohen, 1987; Woebken *et al.*, 2007; Karl & Tilbrook, 1994). In diffuse flow samples where the contribution of reduced hydrothermal fluid is greater, Arctic96BD-19 reached up to 25% of total bacterial SSU rRNA gene copies despite elevated temperature and lower oxygen conditions. Indeed, in most of these samples including the high temperature Hulk sample, Arctic96BD-19 was more prevalent than SUP05, suggesting a potentially more versatile energy metabolism and temperature tolerance than previously recognized. Given the emerging role for Arctic96BD-19 in carbon and sulfur cycling in the dark ocean, this versatility warrants more in-depth exploration of ecotypes using cultivation and single-cell genomic approaches.

The extent to which SUP05 and Arctic96BD-19 are biogeochemically active members of hydrothermal plumes, diffuse flow fluids and background seawater cannot be determined from the present study given our inability to distinguish between active and dormant cells. Thus while we can place these groups in specific mixing regimes and comment on their potential ecological and biogeochemical roles, we are unable to link these groups with specific processes in the environment. A recent study by Bourbonnais and colleagues working at some of the same vent sites surveyed in this study implicated SUP05 in nitrogen loss processes in diffuse flow fluids (Bourbonnais *et al.*, 2012). Curiously, SSU rRNA gene sequences affiliated with Arctic96BD-19 were not detected in a clone library recovered from 24.8 °C diffuse flow fluids at the Hulk site one year earlier. These observations highlight the challenges associated with dynamic hydrothermal vent ecosystems and point to the need for more statistical sampling approaches to more accurately identify ecological patterns under changing geochemical conditions.

Geochemical models of hydrothermal vent plumes have suggested that oxidation of elemental sulfur and metal sulfides represents one of the largest potential sources of metabolic energy in these fluids (McCollom, 2000). However, similar models suggest that methanogenesis and reduction of sulfate or elemental sulfur are favored thermodynamically at temperatures above 38˚C (McCollom & Shock, 1997). While these models are based on sulfur and sulfide concentrations from hydrothermal vent fluids on the East Pacific Rise, the relative proportions of compounds at the Juan de Fuca Ridge are similar (Butterfield *et al.*, 1997). Moreover, these geochemical models indicate that the activities of sulfur-oxidizer bacteria are likely to diminish in the later stages of the plume, as elemental sulfur and sulfide are depleted (McCollom, 2000). Our results are broadly consistent with temperature effects predicted in these models, with *Methanococcales* and *Thermococcales* dominating in the high temperature samples. However, the prevalence of presumptive SUP05 and Arctic96BD-19 in all three mixing regimes deviates from the expectation that attenuated plumes are less hospitable to sulfur-oxidizing bacteria and points to the presence of cryptic elemental cycling in these fluids (Canfield *et al.*, 2010). Moreover, the identification of Arctic96BD-19 in high temperature diffuse flow fluids adds a new perspective to the microbial ecology and biogeochemistry of hydrothermal vent ecosystems. Under more stratified water column conditions associated with increasing sulfide concentrations, both SUP05 and Arctic96BD-19 are replaced by sulfur oxidizers affiliated with the *Epsilonproteobacteria*

(Labrenz *et al.*, 2007; Grote *et al.*, 2008; Grote *et al.*, 2012; Lin *et al.*, 2008). It will be of interest to determine the effect of sulfide concentration on SUP05 and Arctic96BD-19 in relation to physiological succession and taxon replacement under the more dynamic geochemical conditions found in plume and diffuse flow fluids.

In conclusion, our results point to common and unique microbial community structures associated with geochemical gradients within three hydrothermal vent mixing regimes within the Juan de Fuca system. While compositional changes reflected known temperature constraints on microbial community structure, the prevalence of SUP05 and Arctic96BD-19 in most samples posits a conserved role for these groups in carbon, sulfur and nitrogen cycling at different ecological scales throughout the dark ocean. It is possible that in marine ecosystems SUP05 and Arctic96BD-19 are important constituents of the "rare biosphere" (Sogin *et al.*, 2006) that exist in low abundances under most environmental conditions, but "bloom" in response to specific geochemical conditions. Under these circumstances, different SUP05 and Arctic96BD-19 ecotypes associated with geochemical gradients as diverse as mussel gills, sinking particles, stratified water columns, and hydrothermal fluids have the potential to serve as ecological indicators for a changing global ocean.

Table 1.1. Summary of sample locations, depth, temperature/temperature anomaly, oxygen, and cell counts. For plume samples, temperature is reported as the temperature anomaly, which is calculated as the difference in temperature between the plume temperature spike and the background seawater temperature.

Sample	Latitude/ Longitude	Depth (m)	Temperature (˚C)	Temperature anomaly (˚C)	Oxygen (uM)	Cell counts (cells/mL)
Background	47˚ 56.00'N 129˚ 04.30' W	2300	1.83	--	85.65	4.94E+04
Needle Plume	47˚ 56.875'N 129˚ 5.940'W	2135	1.87	0.06862	79.99	3.93E+04
Castle Plume	45˚ 55.5690'N 129˚ 55.8242'W	1474	2.42	0.01383	41.72	5.09E+04
CASM Plume	45˚ 59.326'N 130˚ 1.634'W	1525	2.38	0.001	44.15	1.21E+05
Hulk Plume	47˚ 57.024'N 129˚5.762'W	2060	1.90	0.1129	82.45	7.37E+04
Hulk DF	47˚ 57.00' N, 129˚ 5.81' W	2198	125[1]	--		1.69E+07
Lobo DF	47˚ 56.952'N, 129˚ 5.910'W	2188	7.1	--		5.71E+04
Grotto DF	47˚ 56.95'N, 129˚ 5.898'W	2187	18.1	--		3.62E+04
Easter Island DF	47˚ 56.880'N, 129˚ 5.940'W	2197	9.3	--		2.82E+05

[1] Temperature of the Hulk diffuse flow sampled was calculated from magnesium and silica concentrations in the fluids. This calculation is explained in greater detail in Anderson et al, 2011.

Table 1.2. Raw taxonomic assignments of clone library sequences for all bacterial and archaeal clones in all three environmental regimes, as assigned by Greengenes (DeSantis *et al.* 2006). These taxonomic assignments were used as the basis for the bubble plot shown in Figure 2.

GreenGenes Taxonomic Category	Deep seawater	Diffuse flow	Vent plume
Archaea;Methanococci_Eury;Methanocaldococcaceae;OTU	0	8	0
Archaea;Thaumarchaeota;Cenarchaeales;Cenarchaeum;pIVWA5;Unclassified;OTU	64	1	66
Archaea;Thaumarchaeota;Cenarchaeales;Cenarchaeum;Unclassified;OTU	6	1	8
Archaea;Thermococci_Eury;OTU	0	80	0
Archaea;Thermoplasmata_Eury;marine_group_II;CTD005-13A;OTU	4	0	1
Archaea;Thermoplasmata_Eury;marine_group_II;SB95-72;OTU	20	1	18
Archaea;Thermoplasmata_Eury;marine_group_III;DH148-W24;OTU	2	0	2
Archaea;Thermoprotei_Cren;Pyrodictiaceae;OTU	0	1	0
Archaea;Thermoprotei_Cren;Thermoproteaceae;Vulcanisaeta;OTU	0	2	0
Bacteria;Acidobacteria;iii1-15;Unclassified;OTU	0	0	1
Bacteria;Actinobacteria;Acidimicrobidae;Microthrixineae;Unclassified;OTU	1	0	7
Bacteria;Aquificae;Desulfurobacteria;OTU	0	1	0
Bacteria;Bacteroidetes;Bacteroidales;Unclassified;OTU	3	0	0
Bacteria;Bacteroidetes;Flavobacteriales;Cytophaga;Psychroserpens_burtonensis;AEGEAN_179;OTU	0	3	0
Bacteria;Bacteroidetes;Flavobacteriales;Arctic97A-17;OTU	1	0	0
Bacteria;Bacteroidetes;Flavobacteriales;Cytophaga;Psychroserpens_burtonensis;Unclassified;OTU	0	1	0
Bacteria;Bacteroidetes;Flavobacteriales;F1CA7;Unclassified;OTU	1	1	0
Bacteria;Bacteroidetes;Flavobacteriales;Sporocytophaga;OTU	0	2	0
Bacteria;Bacteroidetes;Flavobacteriales;Unclassified;OTU	2	2	0
Bacteria;Chloroflexi;Chloroflexi-4;SAR307;OTU	1	0	1
Bacteria;Cyanobacteria;Chloroplasts;Unclassified;OTU	0	1	0
Bacteria;Marine_group_A;Arctic95A-2;OTU	2	0	0
Bacteria;Marine_group_A;Arctic96B-7;OTU	0	0	2
Bacteria;Marine_group_A;SAR406;OTU	0	0	1
Bacteria;Marine_group_A;Unclassified;OTU	1	0	2
Bacteria;Marine_group_A;ZA3312c;OTU	0	0	1
Bacteria;Marine_group_A;ZA3648c;OTU	2	0	2
Bacteria;Planctomycetes;Planctomycetacia;P._marina;Unclassified;OTU	1	0	0
Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Unclassified;OTU	1	0	0
Bacteria;Planctomycetes;agg27;OM190;ARKCH2Br2-76;OTU	1	0	1
Bacteria;Planctomycetes;Unclassified;OTU	0	0	1
Bacteria;Planctomycetes;WPS-1;CL500-3;CL120-56;DE613;OTU	1	0	1
Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Brevundimonas;Caulobacter;Caulobacter_henricii;Unclassified;OTU	0	1	0
Bacteria;Proteobacteria;Alphaproteobacteria;Consistiales;AEGEAN_187;OTU	0	4	0
Bacteria;Proteobacteria;Alphaproteobacteria;Consistiales;Pelagibacter;SAR11;Candidatus_Pelagibacter;Candidatus_Pelagibacter_ubique;OTU	4	5	4
Bacteria;Proteobacteria;Alphaproteobacteria;Consistiales;Pelagibacter;SAR11;Candidatus_Pelagibacter;Unclassified;OTU	5	0	3
Bacteria;Proteobacteria;Alphaproteobacteria;Oleomonas;ctg_NISA150;OTU	0	0	1
	0	4	0

Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingobium;Sphingomonas_xenophaga;OTU			
Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingobium;Unclassified;OTU	0	2	0
Bacteria;Proteobacteria;Alphaproteobacteria;Consistiales;Unclassified;OTU	2	0	0
Bacteria;Proteobacteria;Alphaproteobacteria;Unclassified;OTU	1	0	0
Bacteria;Proteobacteria;Alphaproteobacteria;ZA3420c;OTU	1	0	0
Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonas_parapaucimobilis;Unclassified;OTU	0	1	0
Bacteria;Proteobacteria;Deltaproteobacteria;Desulfobacterium_catecholicum;Desulfobulbus_rhabdoformis;Unclassified;OTU	0	1	0
Bacteria;Proteobacteria;Deltaproteobacteria;Geobacter;Pelobacter_propionicus;Unclassified;OTU	1	0	0
Bacteria;Proteobacteria;Deltaproteobacteria;Geobacter;Unclassified;OTU	1	0	0
Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;OM27;Unclassified;OTU	0	0	4
Bacteria;Proteobacteria;Deltaproteobacteria;Myxococcales;Unclassified;OTU	3	0	1
Bacteria;Proteobacteria;Deltaproteobacteria;NB1-j;NB1-i;JTB38;OTU	2	0	0
Bacteria;Proteobacteria;Deltaproteobacteria;Nitrospina;OTU	0	0	2
Bacteria;Proteobacteria;Deltaproteobacteria;PB19;OTU	0	0	1
Bacteria;Proteobacteria;Deltaproteobacteria;Sva0853;SAR324;OTU	1	0	2
Bacteria;Proteobacteria;Epsilonproteobacteria;Nautilliales;Nautillaceae;OTU	0	7	0
Bacteria;Proteobacteria;Epsilonproteobacteria;Nautilliales;Thioreductaceae;OTU	0	10	0
Bacteria;Proteobacteria;Epsilonproteobacteria;Ppalm_CA39;SF_C23-A7_shell;OTU	0	0	1
Bacteria;Proteobacteria;Gammaproteobacteria;agg47;OTU	2	0	2
Bacteria;Proteobacteria;Gammaproteobacteria;Aeromonadaceae;Aeromonas;OTU	1	0	0
Bacteria;Proteobacteria;Gammaproteobacteria;Alteromonadales;Alteromonas;OTU	0	7	0
Bacteria;Proteobacteria;Gammaproteobacteria;Arctic96B-1;Gammaproteobacteria;OTU	4	0	0
Bacteria;Proteobacteria;Gammaproteobacteria;Alteromonadales;Pseudoalteromonadaceae;Pseudoalteromonas;Unclassified;OTU	0	7	1
Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteria;Comamonadaceae;Comamonas;Comamonas_testosteroni;Unclassified;OTU	0	1	0
Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteria;Comamonadaceae;Comamonas;Unclassified;OTU	0	2	0
Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteria;Comamonadaceae;Unclassified;OTU	0	3	0
Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteria;Ralstoniaceae;Unclassified;OTU	0	1	0
Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacteriales;Unclassified;OTU	0	1	2
Bacteria;Proteobacteria;Gammaproteobacteria;HTCC2207;OTU	0	1	0
Bacteria;Proteobacteria;Gammaproteobacteria;Hyd24-01;OTU	1	0	3
Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;Unclassified;OTU	0	1	0
	0	9	0

Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadaceae;Unclassified;OTU			
Bacteria;Proteobacteria;Gammaproteobacteria;SAR86;environmental_sequence;ZA3913c;OTU	0	1	0
Bacteria;Proteobacteria;Gammaproteobacteria;SAR86;ZD0433;OTU	0	0	1
Bacteria;Proteobacteria;Gammaproteobacteria;SUP05;mussel_thioautotrophic_gill_symbiont_MAR1;OTU	5	0	42
Bacteria;Proteobacteria;Gammaproteobacteria;SUP05;Unclassified;OTU	11	0	3
Bacteria;Proteobacteria;Gammaproteobacteria;Unclassified;OTU	5	0	6
Bacteria;Proteobacteria;Gammaproteobacteria;ZD0417;Unclassified;OTU	3	0	1
Bacteria;Proteobacteria;Gammaproteobacteria;ZD0417;ZA3605c;OTU	0	0	1
Bacteria;Verrucomicrobia;Opitutae;MB11C04;Unclassified;OTU	6	1	0
Bacteria;Verrucomicrobia;Verrucomicrobia_subdivision_3;Verruco-3;CTD005-1B-02;Unclassified;OTU	0	0	1
Bacteria;Verrucomicrobia;Verrucomicrobiae;Unclassified;OTU	1	0	0

Table 1.3. Diversity indices calculated using various methods for plume, diffuse flow, and background samples used in this study. T-RFLP diversity indices are listed as the average across all restriction enzymes used ± standard deviation of the average. Bacteria estimates are averaged across T-RFLP profiles for BstUI, HaeIII, and MspI; archaeal estimates averaged across T-RFLP profiles for BstUI and HaeIII. T-RFLP diversity indices were calculated using ESTIMATES (Version 8.2; R.K. Colwell); clone library diversity estimates were calculated using mothur (Schloss et al., 2009). See Methods for details.

Domain	Sample location	Sample type	Method	S_{obs}*	Chao1†	Jackknife‡
Bacteria	Deep seawater	Deep seawater	Clone library	51	174.5	226.94
	Deep seawater	Deep seawater	T-RFLP	14.68 +/- 4.06	14.6 +/- 4.01	16.52 +/- 7.81
	Hulk	Plume	Clone library	45	107	110.96
	Hulk	Plume	T-RFLP	46.375 +/- 1.73	46.31 +/- 1.83	65.91 +/- 4.97
	Needle	Plume	T-RFLP	52.14 +/- 3.08	52.09 +/- 3.08	73.30 +/- 6.31
	CASM	Plume	T-RFLP	23.02 +/- 3.10	22.98 +/- 3.02	32.52 +/- 4.33
	Castle	Plume	T-RFLP	29.72 +/- 2.85	29.43 +/- 2.87	43.52 +/- 4.61
	Hulk	Diffuse flow	Clone library	42	117.6	117.63
	Hulk	Diffuse flow	T-RFLP	43.52 +/- 1.21	43.45 +/- 1.23	62.58 +/- 3.92
	Easter Island	Diffuse flow	T-RFLP	34.54 +/- 3.38	34.3 +/- 3.03	49.96 +/- 4.78
	Lobo	Diffuse flow	T-RFLP	48.91 +/- 2.21	48.71 +/- 2.25	68.71 +/- 5.62
	Grotto	Diffuse flow	T-RFLP	39.03 +/- 2.53	38.92 +/- 2.31	56.59 +/- 4.04
Archaea	Deep seawater	Deep seawater	Clone library	12	17	17
	Deep seawater	Deep seawater	T-RFLP	10.08 +/- 6.71	10.39 +/- 6.55	10.39 +/- 12.28
	Hulk	Plume	Clone library	14	19	20
	Hulk	Plume	T-RFLP	88.81 +/- 5.91	88.96 +/- 6.04	124.425 +/- 9.31

	Needle	Plume	T-RFLP	63.26 +/- 5.73	63.27 +/- 5.79	86.74 +/- 9.47
	CASM	Plume	T-RFLP	32.40 +/- 2.98	32.80 +/- 2.93	47.47 +/- 4.27
	Castle	Plume	T-RFLP	41.12 +/- 3.48	41.39 +/- 3.52	59.75 +/- 4.44
	Hulk	Diffuse flow	Clone library	18	51	48.00
	Hulk	Diffuse flow	T-RFLP	55.26 +/- 5.56	55.41 +/- 5.72	77.89 +/- 8.56
	Easter Island	Diffuse flow	T-RFLP	45.81 +/- 3.24	46.02 +/- 3.16	65.98 +/- 4.18
	Lobo	Diffuse flow	T-RFLP	60.14 +/- 6.29	60.19 +/- 4.67	83.33 +/- 10.32
	Grotto	Diffuse flow	T-RFLP	52.56 +/- 3.69	52.71 +/- 3.73	74.74 +/- 5.37

*Number of OTUs observed. For clone libraries, this is the number of clusters at a distance of 0.04. For T-RFLP, this is the Mau Tau expected richness (Colwell et al., 2004).

†Chao1 richness estimator (Chao, 1987).

‡First-order Jackknife richness estimator (Burnham & Overton, 1978, 1979; Smith & van Belle, 1984; Heltshe & Forrester, 1983).

Table 1.4. Relative abundances of SUP05 and ARCTIC96BD-19 across all sample types. Abundances quantified through quantitative PCR. Quantities are expressed as percentages of each group relative to total bacteria.

Sample	**Temperature/ temp anomaly (˚C)**	**SUP05 % Abundance** (relative to total bacteria)	**Std. Error of SUP05 % abundance**	**Arctic96BD-19 % Abundance** (relative to total bacteria)	**Std. Error of Arctic96BD-19 % abundance**
Plume					
Hulk	*0.11*	14.1	0.29	12.9	0.69
Needle	*0.068*	27.3	1.52	22.7	1.67
Castle	*0.014*	14.7	0.56	43	2.17
CASM	*0.001*	4.1	0.034	64.7	1.20
Diffuse Flow					
Hulk	125	0.4	.034	6.1	0.40
Grotto	18.1	11.2	0.27	25.5	0.77
Easter Island	9.3	18.7	0.95	11.9	0.92
Lobo	7.1	17.7	0.87	21.5	2.49
Background	1.8	3.2	0.16	22.7	0.56

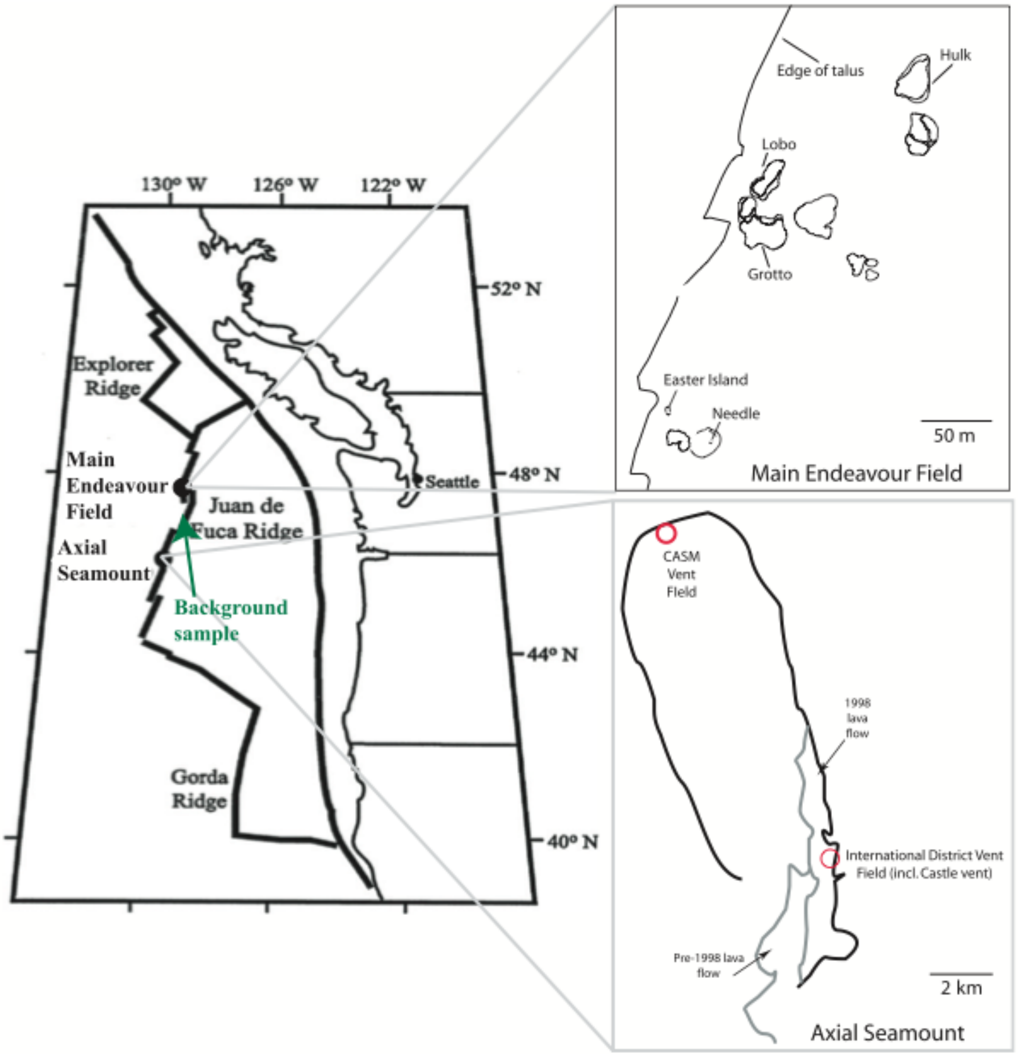


Figure 1.1. Schematic map of the Juan de Fuca plate, Main Endeavour Field, and Axial Seamount. Adapted from figures in Huber et al. (2002) and V. Robigou (unpublished data).

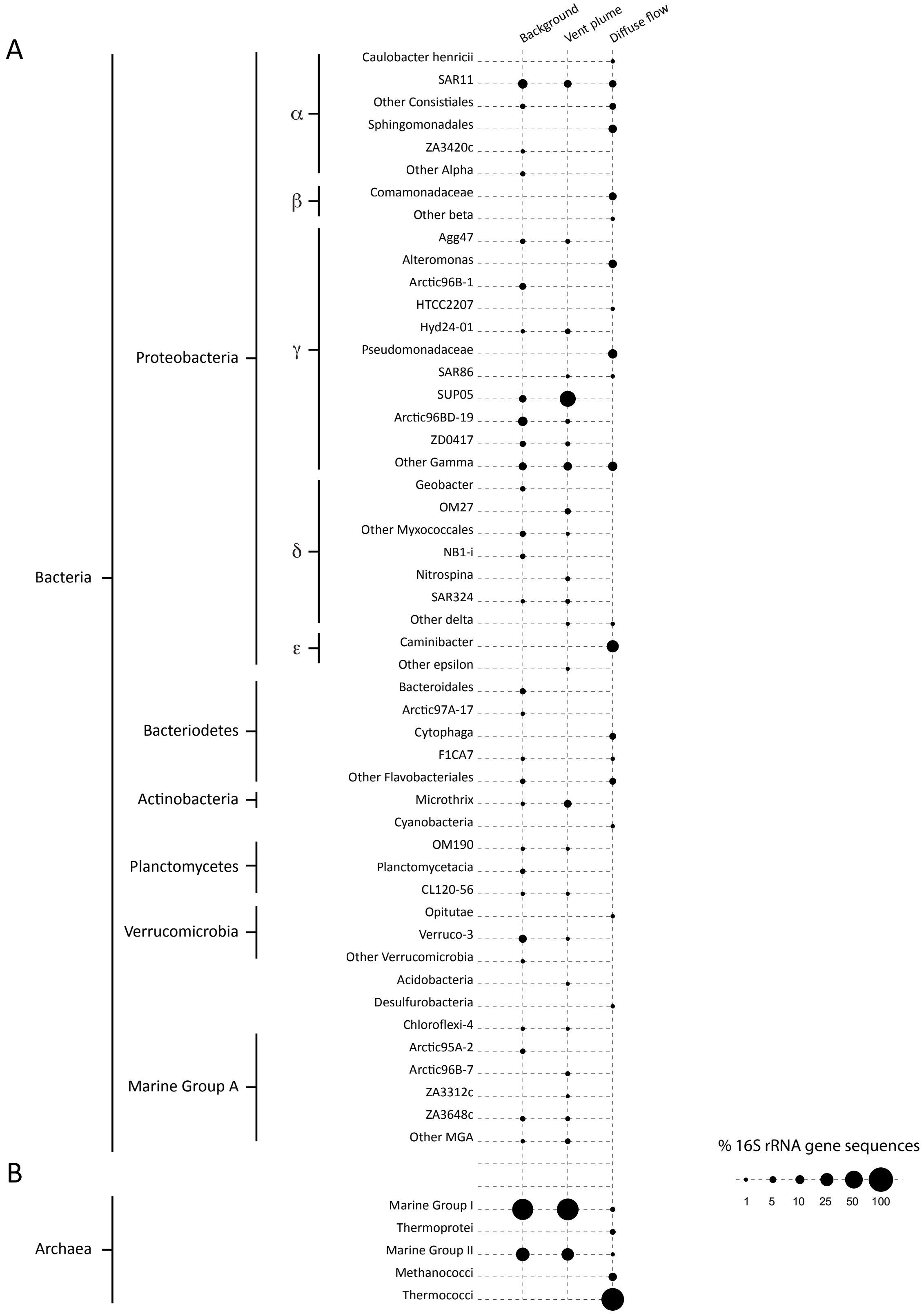


Figure 1.2. Dot plot of (a) bacterial and (b) archaeal diversity from diffuse flow fluid and hydrothermal plume associated with Hulk vent, as well as background seawater, based on 16S rRNA gene sequence profiles (see methods). The size of each dot indicates the percentage identified 16S rRNA gene sequences falling within a particular taxonomic group. The number of bacterial clones sequenced per sample is background seawater = 78, vent plume = 102 and diffuse flow = 81. The number of archaeal clones sequenced per sample is background seawater = 96, vent plume = 95 and diffuse flow = 94.

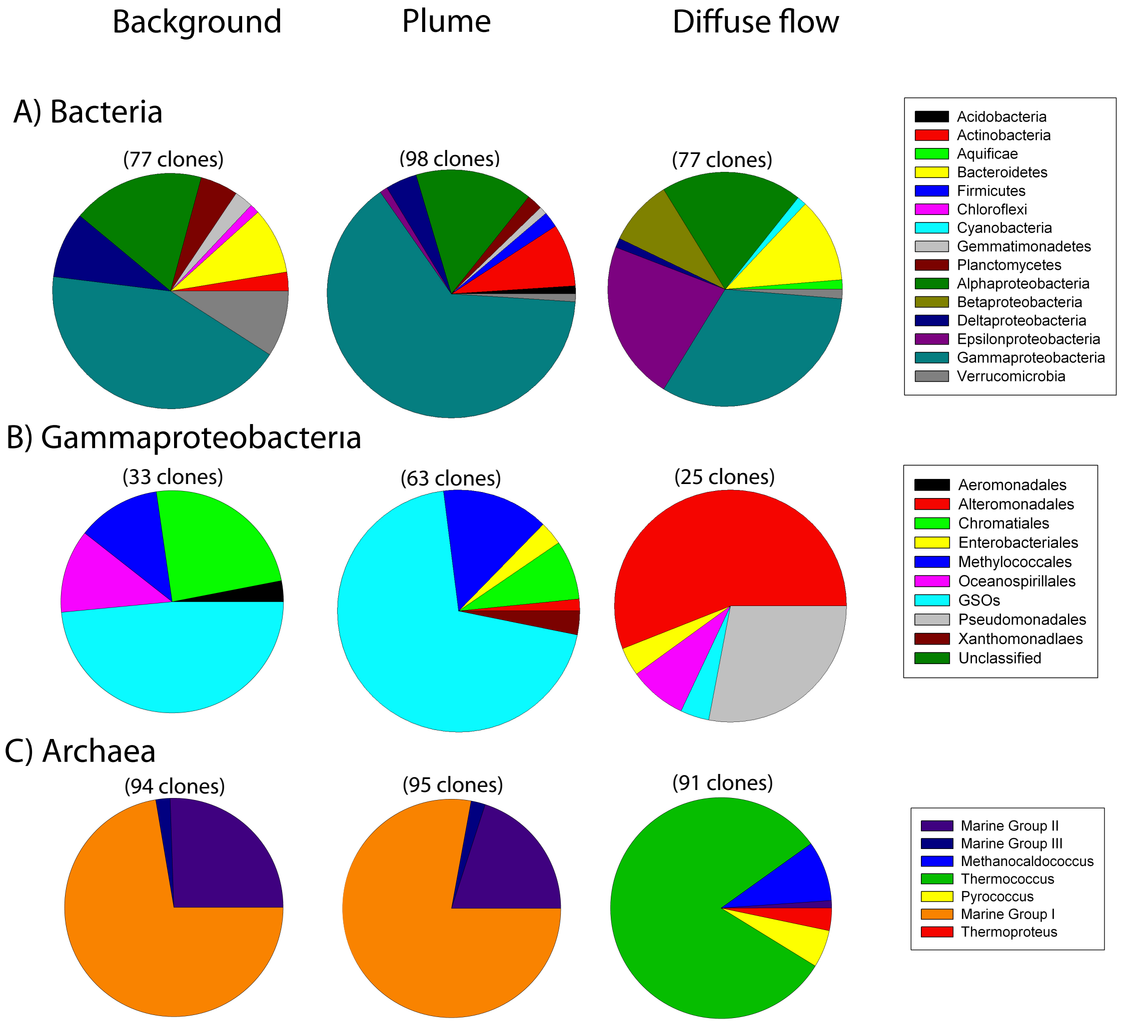


Figure 1.3. Pie charts of archaeal and bacterial clone libraries from Hulk diffuse flow, plume, and background samples. A) Breakdown of bacterial clone taxonomic assignments, B) breakdown of Gammaproteobacteria groups only from the bacterial clone library, C) breakdown of archaeal clone library taxonomic assignments. SUP05 and Arctic96BD-19 groups are here grouped together and abbreviated as GSOs (Gammaproteobacteria sulfur oxidizers). Chimeras were removed, sequences aligned, and taxonomic assignments made using GreenGenes (deSantis et al. 2006).

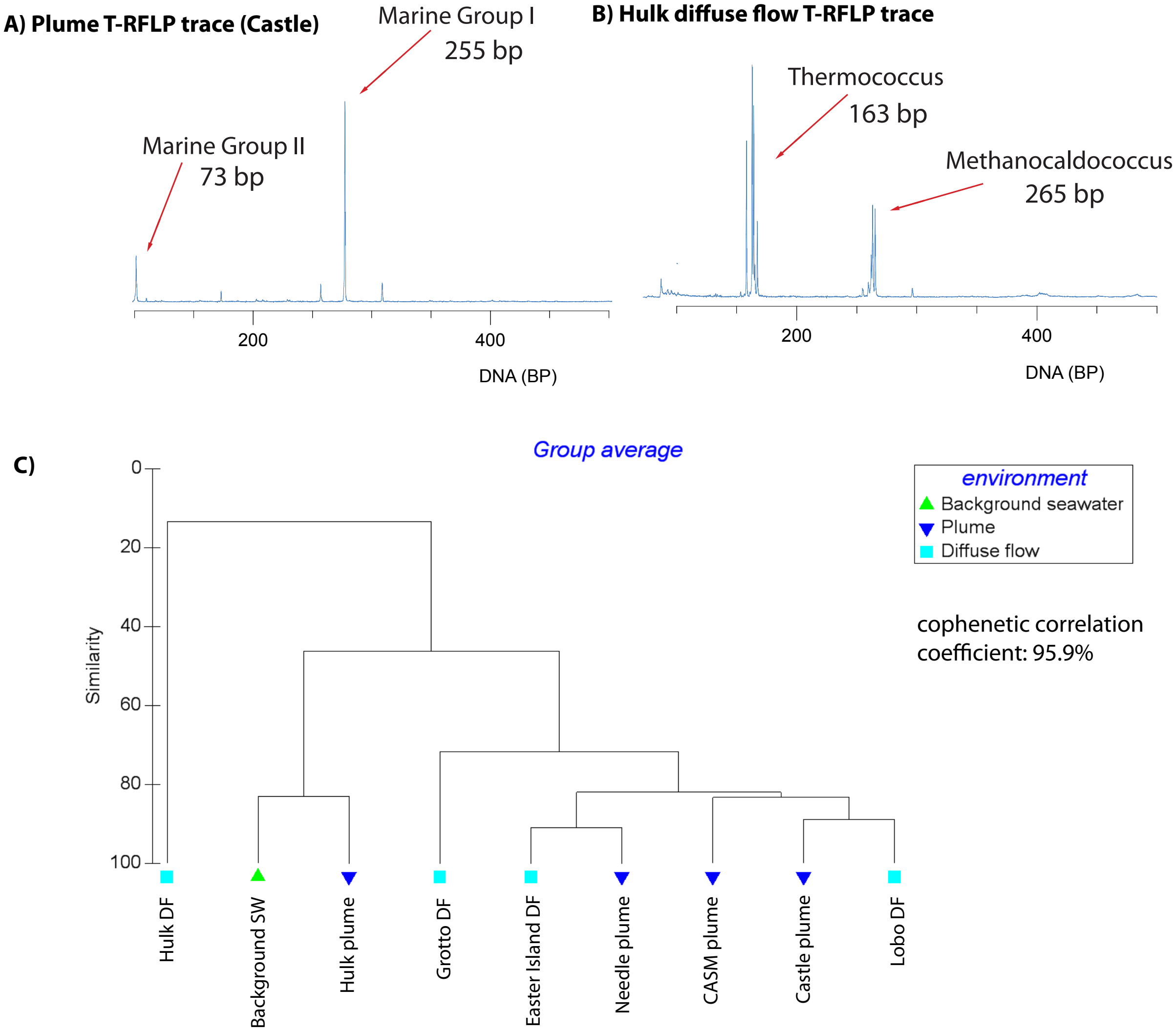


Figure 1.4. T-RFLP community profiling of samples amplified with universal archaeal primers. (a) Representative T-RFLP trace of plume samples digested with restriction enzyme RsaI. (b) T-RFLP trace of Hulk diffuse flow sample. Y-axes are relative fluorescence units (RFUs) and are not to scale. Peaks were identified through in silico digestion of clone library sequences using the online tool REPK (Collins & Rocap, 2007). (c) Cluster diagrams of sample similarity based on archaeal sample T-RFLP traces digested with BstUI. Distance matrices were produced using Chao's abundance-based Jaccard Index (Chao et al., 2005) and were calculated from the maximum similarity of eight different bin shifts. Samples were clustered using the group average method in PRIMER-E. Cophenetic correlation coefficient for the dendrogram is shown.

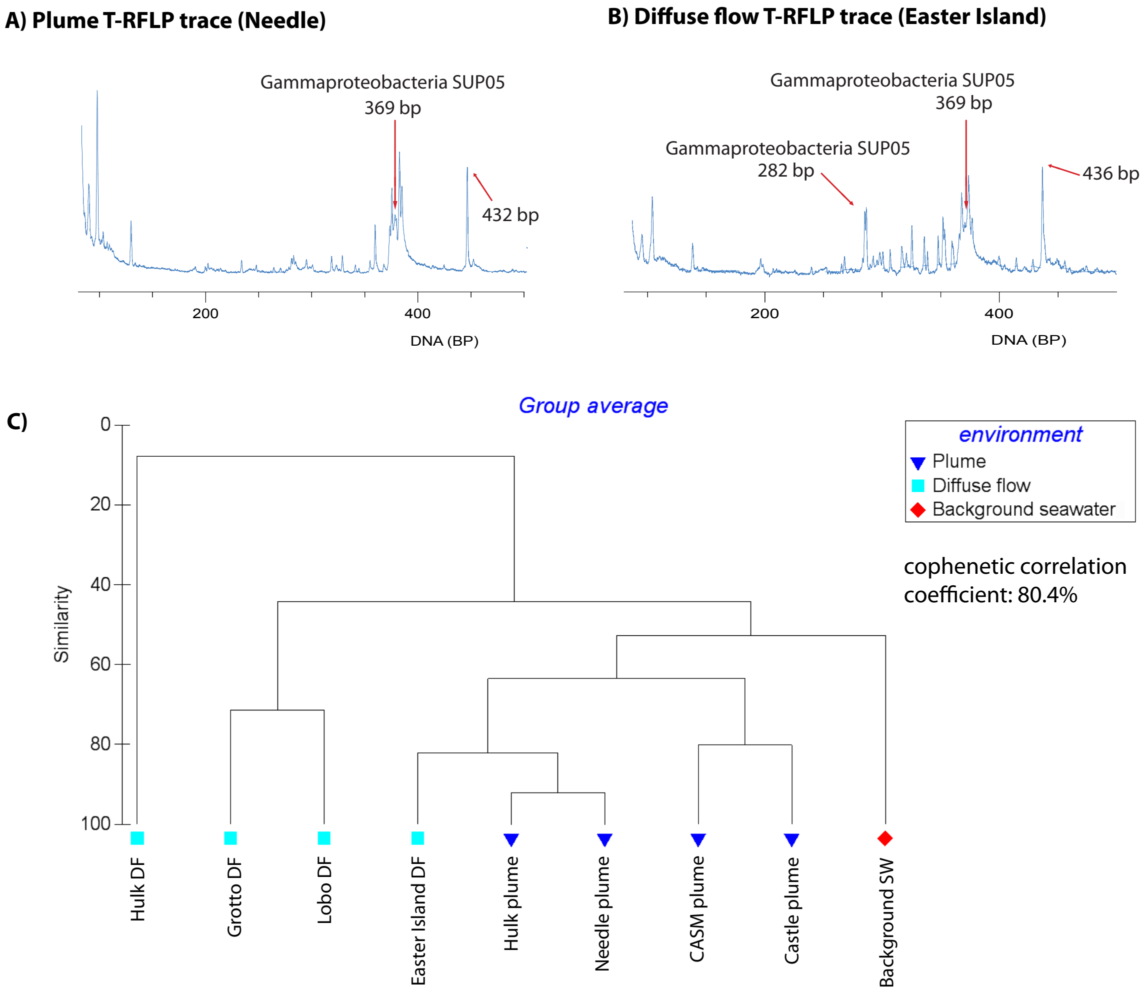


Figure 1.5. T-RFLP community profiling of samples amplified with universal bacterial primers. (a) Representative T-RFLP trace of plume samples digested with restriction enzyme HaeIII. Needle is shown here. (b) Representative T-RFLP trace of diffuse flow samples digested with restriction enzyme HaeIII, and Easter Island is shown here. Y-axes are relative fluorescence units (RFUs) and are not to scale. Peaks were identified through in silico digestion of clone library sequences using the online tool REPK (Collins & Rocap, 2007). (c) Cluster diagrams of sample similarity based on archaeal sample T-RFLP traces digested with BstUI. Distance matrices were produced using Chao's abundance-based Jaccard Index (Chao et al., 2005) and were calculated from the maximum similarity of eight different bin shifts. Samples were clustered using the group average method in PRIMER-E. Cophenetic correlation coefficient for the dendrogram is shown.

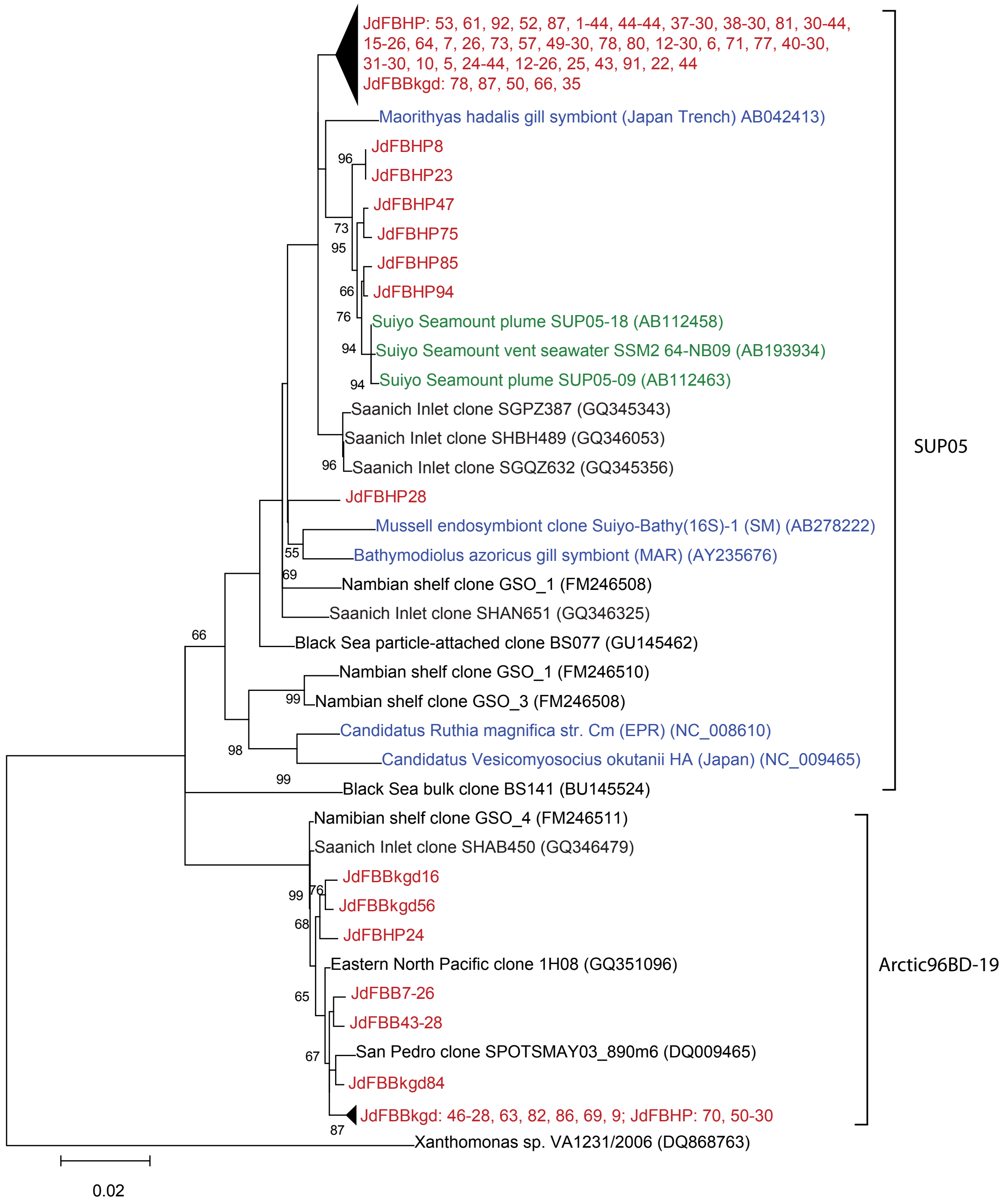

JdFBHP: 53, 61, 92, 52, 87, 1-44, 44-44, 37-30, 38-30, 81, 30-44, 15-26, 64, 7, 26, 73, 57, 49-30, 78, 80, 12-30, 6, 71, 77, 40-30, 31-30, 10, 5, 24-44, 12-26, 25, 43, 91, 22, 44
JdFBBkgd: 78, 87, 50, 66, 35
Maorithyas hadalis gill symbiont (Japan Trench) AB042413)
96
JdFBHP8
JdFBHP23
JdFBHP47
73
JdFBHP75
95
JdFBHP85
66
JdFBHP94
76
Suiyo Seamount plume SUP05-18 (AB112458)
94
Suiyo Seamount vent seawater SSM2 64-NB09 (AB193934)
94
Suiyo Seamount plume SUP05-09 (AB112463)
Saanich Inlet clone SGPZ387 (GQ345343)
Saanich Inlet clone SHBH489 (GQ346053)
96
Saanich Inlet clone SGQZ632 (GQ345356)
JdFBHP28
Mussell endosymbiont clone Suiyo-Bathy(16S)-1 (SM) (AB278222)
55
Bathymodiolus azoricus gill symbiont (MAR) (AY235676)
69
Nambian shelf clone GSO_1 (FM246508)
Saanich Inlet clone SHAN651 (GQ346325)
66
Black Sea particle-attached clone BS077 (GU145462)
Nambian shelf clone GSO_1 (FM246510)
99
Nambian shelf clone GSO_3 (FM246508)
Candidatus Ruthia magnifica str. Cm (EPR) (NC_008610)
98
Candidatus Vesicomyosocius okutanii HA (Japan) (NC_009465)
99
Black Sea bulk clone BS141 (BU145524)
SUP05
Namibian shelf clone GSO_4 (FM246511)
Saanich Inlet clone SHAB450 (GQ346479)
JdFBBkgd16
99
76
JdFBBkgd56
68
JdFBHP24
Eastern North Pacific clone 1H08 (GQ351096)
65
JdFBB7-26
JdFBB43-28
San Pedro clone SPOTSMAY03_890m6 (DQ009465)
67
JdFBBkgd84
JdFBBkgd: 46-28, 63, 82, 86, 69, 9; JdFBHP: 70, 50-30
87
Xanthomonas sp. VA1231/2006 (DQ868763)
Arctic96BD-19
0.02

Figure 1.6. Evolutionary relationships of clones and reference sequences within the sulfur-oxidizing Gammaproteobacteria group. See Methods for techniques in tree construction. The evolutionary history was inferred using the maximum likelihood method; the bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analyzed. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Clones from this study are shown in red, clones from other hydrothermal systems are shown in green, and symbionts are shown in blue. GenBank accession numbers are provided for clones not from this study.

References

Alldredge, A.L. and Cohen, Y. (1987) Can microscale chemical patches persist in the sea? Microelectrode study of marine snow, fecal pellets. Science **235**: 689–691.

Anderson, R.E., Brazelton, W.J., and Baross, J.A. (2011) Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. FEMS Microbiol Ecol **77**: 120–133.

De Angelis, M.A., Lilley, M.D., Olson, E.J., and Baross, J.A. (1993) Methane oxidation in deep-sea hydrothermal plumes of the Endeavour Segment of the Juan de Fuca Ridge. Deep-Sea Res Pt I **40**: 1169–1186.

Bourbonnais, A., Juniper, S.K., Butterfield, D.A., Devol., A.H., Kuypers, M.M.M., Lavik, G., *et al.* (2012) Activity and abundance of denitrifying bacteria in the subsurface biosphere of diffuse hydrothermal vents of the Juan de Fuca Ridge. Biogeosciences Discuss **9**: 4177–4223.

Brown, M. V, Schwalbach, M.S., Hewson, I., and Fuhrman, J.A. (2005) Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: development and application to a time series. Environ Microbiol **7**: 1466–79.

Butterfield, D.A., Jonasson, I.R., Massoth, G.J., Feely, R.A., Roe, K.K., Embley, R.E., *et al.* (1997) Seafloor eruptions and evolution of hydrothermal fluid chemistry. Philos T Roy Soc A **355**: 369–386.

Butterfield, D.A., Roe, K.K., Lilley, M.D., Huber, J.A., Baross, J.A., Embley, R.W., and Massoth, G.J. (2004) Mixing, reaction and microbial activity in the sub-seafloor revealed by temporal and spatial variation in diffuse flow vents at Axial Volcano. In Wilcock, W.S.D. *et al.* (eds), *The Subseafloor Biosphere at Mid-Ocean Ridges*. American Geophysical Union, Washington, D.C., pp. 269–290.

Canfield, D.E., Stewart, F.J., Thamdrup, B., De Brabandere, L., Dalsgaard, T., Delong, E.F., *et al.* (2010) A cryptic sulfur cycle in oxygen-minimum-zone waters off the Chilean coast. Science **330**: 1375–8.

Cavanaugh, C.M. (1983) Symbiotic chemoautotrophic bacteria in marine invertebrates from sulphide-rich habitats. Nature **302**: 58–61.

Clarke, K.R. and Gorley, R.N. (2006) PRIMER v6: User Manual/tutorial. Primer-E Ltd Plymouth. the text.

Collins, R.E. and Rocap, G. (2007) REPK: an analytical web server to select restriction endonucleases for terminal restriction fragment length polymorphism analysis. Nucleic Acids Res **35**: W58–W62.

Corliss, J.B., Dymond, J., Gordon, L.I., Edmond, J.M., von Herzen, R.P., Ballard, R.D., *et al.* (1979) Submarine thermal springs on the Galapagos Rift. Science **203**: 1073–83.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol **72**: 5069.

Dick, G.J. and Tebo, B.M. (2010) Microbial diversity and biogeochemistry of the Guaymas Basin deep‐sea hydrothermal plume. Environ Microbiol **12**: 1334–1347.

Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**: 783–791.

German, C.R., Bowen, A., Coleman, M.L., Honig, D.L., Huber, J.A., Jakuba, M. V, *et al.* (2010) Diverse styles of submarine venting on the ultraslow spreading Mid-Cayman Rise. Proc Natl Acad Sci U S A **107**: 14020.

Grote, J., Jost, G., Labrenz, M., Herndl, G.J., and Jurgens, K. (2008) *Epsilonproteobacteria* represent the major portion of chemoautotrophic bacteria in sulfidic waters of pelagic redoxclines of the Baltic and Black Seas. Appl Environ Microbiol **74**: 7546–7551.

Grote, J., Schott, T., Bruckner, C.G., Glöckner, F.O., Jost, G., Teeling, H., *et al.* (2012) Genome and physiology of a model *Epsilonproteobacterium* responsible for sulfide detoxification in marine oxygen depletion zones. Proc Natl Acad Sci U S A **109**: 506–10.

Hewson, I. and Fuhrman, J.A. (2006) Improved strategy for comparing microbial assemblage fingerprints. Microb Ecol **51**: 147–153.

Huber, J.A., Butterfield, D.A., and Baross, J.A. (2003) Bacterial diversity in a subseafloor habitat following a deep-sea volcanic eruption. FEMS Microbiol Ecol **43**: 393–409.

Huber, J.A., Butterfield, D.A., and Baross, J.A. (2006) Diversity and distribution of subseafloor *Thermococcales* populations in diffuse hydrothermal vents at an active deep-sea volcano in the northeast Pacific Ocean.

Huber, J.A., Butterfield, D.A., and Baross, J.A. (2002) Temporal changes in archaeal diversity and chemistry in a mid-ocean ridge subseafloor habitat. Appl Environ Microbiol **68**: 1585–1594.

Huber, J.A., Mark Welch, D.B., Morrison, H.G., Huse, S.M., Neal, P.R., Butterfield, D.A., and Sogin, M.L. (2007) Microbial population structures in the deep marine biosphere. Science **318**: 97–100.

Johnson, K.S., Beehler, C.L., Sakamoto-Arnold, C.M., and Childress, J.J. (1986) In situ measurements of chemical distributions in a deep-sea hydrothermal vent field. Science **231**: 1139–41.

Kadko, D.C.C., Rosenberg, N.D.D., Lupton, J.E.E., Collier, R.W.W., and Lilley, M.D.D. (1990) Chemical reaction rates and entrainment within the Endeavour Ridge hydrothermal plume. Earth Planet Sci Lett **99**: 315–335.

Karl, D.M., Knauer, G.A., Martin, J.H., and Ward, B.B. (1984) Bacterial chemolithotrophy in the ocean is associated with sinking particles. Nature **309**: 54–56.

Karl, D.M. and Tilbrook, B.D. (1994) Production and transport of methane in oceanic particulate organic matter. Nature **368**: 732–734.

Labrenz, M., Jost, G., and Jurgens, K. (2007) Distribution of abundant prokaryotic organisms in the water column of the central Baltic Sea with an oxic-anoxic interface. Aquat Microb Ecol **46**: 177–190.

Lam, P., Cowen, J.P., Popp, B.N., and Jones, R.D. (2008) Microbial ammonia oxidation and enhanced nitrogen cycling in the Endeavour hydrothermal plume. Geochim Cosmochim Acta **72**: 2268–2286.

Lavik, G., Stührmann, T., Brüchert, V., Van der Plas, A., Mohrholz, V., Lam, P., *et al.* (2008) Detoxification of sulphidic African shelf waters by blooming chemolithotrophs. Nature **457**: 581–584.

Lin, X.J., Scranton, M.I., Chistoserdov, A.Y., Varela, R., and Taylor, G.T. (2008) Spatiotemporal dynamics of bacterial populations in the anoxic Cariaco Basin. Limnol Oceanogr **53**: 37–51.

Lloyd, K.G., Macgregor, B.J., and Teske, A. (2010) Quantitative PCR methods for RNA and DNA in marine sediments: maximizing yield while overcoming inhibition. FEMS Microbiol Ecol **72**: 143–51.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, *et al.* (2004) ARB: a software environment for sequence data. Nucleic Acids Res **32**: 1363 –1371.

Lupton, J.E., Delaney, J.R., Johnson, H.P., and Tivey, M.K. (1985) Entrainment and vertical transport of deep-ocean water by buoyant hydrothermal plumes. Nature **316**: 621–623.

McCollom, T.M. (2000) Geochemical constraints on primary productivity in submarine hydrothermal vent plumes. Deep-Sea Res Pt I **47**: 85–101.

McCollom, T.M. and Shock, E.L. (1997) Geochemical constraints on chemolithoautotrophic metabolism by microorganisms in seafloor hydrothermal systems. Geochim Cosmochim Acta **61**: 4375–4391.

Nakagawa, S., Takai, K., Inagaki, F., Hirayama, H., Nunoura, T., Horikoshi, K., and Sako, Y. (2005) Distribution, phylogenetic diversity and physiological characteristics of *Epsilonroteobacteria* in a deep-sea hydrothermal field. Environ Microbiol **7**: 1619–1632.

Newton, I.L.G., Woyke, T., Auchtung, T.A., Dilly, G.F., Dutton, R.J., Fisher, M.C., *et al.* (2007) The Calyptogena magnifica chemoautotrophic symbiont genome. Science (New York, NY) **315**: 998–1000.

Opatkiewicz, A.D., Butterfield, D.A., and Baross, J.A. (2009) Individual hydrothermal vents at Axial Seamount harbor distinct subseafloor microbial communities. FEMS Microbiol Ecol **70**: 81–92.

Orcutt, B.N., Sylvan, J.B., Knab, N.J., and Edwards, K.J. (2011) Microbial ecology of the dark ocean above, at, and below the seafloor. Microbiol Mol Biol R **75**: 361–422.

Osborne, C.A., Rees, G.N., Bernstein, Y., and Janssen, P.H. (2006) New threshold and confidence estimates for terminal restriction fragment length polymorphism analysis of complex bacterial communities. Appl Environ Microbiol **72**: 1270.

Rzhetsky, A. and Nei, M. A simple method for estimating and testing minimum-evolution trees. Mol Biol Evol **9**: 945–967.

Shanks, A.L. and Reeder, M.L. (1993) Reducing microzones and sulfide production in marine snow. Mar Ecol Prog Ser **96**: 43–47.

Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R., *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A **103**: 12115–20.

Sunamura, M., Higashi, Y., Miyako, C., Ishibashi, J-I., and Maruyama, A. (2004) Two Bacteria phylotypes are predominant in the Suiyo Seamount hydrothermal plume. Appl Environ Microbiol **70**: 1190–1198.

Swan, B.K., Martinez-Garcia, M., Preston, C.M., Sczyrba, A., Woyke, T., Lamy, D., *et al.* (2011) Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. Science **333**: 1296–1300.

Sylvan, J.B., Pyenson, B.C., Rouxel, O., German, C.R., and Edwards, K.J. (2012) Time-series analysis of two hydrothermal plumes at 9°50'N East Pacific Rise reveals distinct, heterogeneous bacterial populations. Geobiology **10**: 178–92.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol **24**: 1596–9.

Walsh, D.A. and Hallam, S.J. (2011) Bacterial community structure and dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia. In, de Bruijn, F.J. (ed), *Handbook of Molecular Microbial Ecology II: Metagenomics in Different Habitats*. Wiley-Blackwell, Hoboken, NJ, USA, pp. 253–267.

Walsh, D.A., Zaikova, E., Howes, C.G., Song, Y.C., Wright, J.J., Tringe, S.G., *et al.* (2009) Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. Science **326**: 578.

Woebken, D., Fuchs, B.M., Kuypers, M.M.M., and Amann, R. (2007) Potential interactions of particle-associated anammox bacteria with bacterial and archaeal partners in the Namibian upwelling system. Appl Environ Microbiol **73**: 4648–57.

Zaikova, E., Walsh, D.A., Stilwell, C.P., Mohn, W.W., Tortell, P.D., and Hallam, S.J. (2010) Microbial community dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia. Environ Microbiol **12**: 172–91.

CHAPTER TWO
Biogeography and ecology of the rare and abundant microbial lineages in deep-sea hydrothermal vents

Summary

Environmental gradients generate countless ecological niches in deep-sea hydrothermal vent systems, which foster diverse microbial communities. The majority of the microbial lineages in these communities are observed in very low (rare) abundance. However, the ecological role of rare lineages in hydrothermal vent microbial communities is not yet clear, nor is it known how the community structures of rare abundant lineages compare from one vent system to the next. Here, we use 16S rRNA gene pyrotag sequencing to describe biogeographic patterning and microbial community structure of both archaea and bacteria in hydrothermal vent systems. The abundant lineages of archaeal communities tended to dominate and were more widely dispersed than the abundant lineages of bacterial communities. Rare lineages in both the archaeal and bacterial domains were generally restricted biogeographically. Additionally, analysis of a single (but high-volume) high-temperature fluid sample thought to represent the deep hot biosphere revealed a unique microbial community that was distinct from microbial communities found in diffuse flow fluid or sulfide samples, although similarities were observed between rare thermophilic archaeal groups and those found in sulfides. These results dispel the picture of the rare biosphere as based on the concept of "everything is everywhere" by indicating that the rare biosphere in hydrothermal vent systems displays a high degree of endemism, while a small percentage of lineages has a widespread distribution.

Introduction

In almost all ecosystems investigated to date, a minority of microbial lineages dominates the community, while the bulk of the diversity is comprised by many species that occur in very low abundance. These rare lineages are collectively known as the "rare biosphere," a term first coined by Sogin et al. (2006). The rare biosphere applies to both the archaeal and bacterial domains, though it remains unclear what role the rare biosphere plays in the ecology or evolution of a microbial community in a given niche. It has been suggested that the rare biosphere provides a source of genes for horizontal gene transfer (Sogin *et al.*, 2006), that it acts as a seed bank of dormant cells that bloom when more favorable conditions arise (Jones and Lennon, 2010; Lennon and Jones, 2011; Gibbons *et al.*, 2013), and that it acts as a repository of "genetic memory" retained from past conditions that may arise again (Brazelton *et al.*, 2010).

An important tenet of microbial biogeography that is used to partially explain the existence of a rare biosphere is the hypothesis put forward by Baas Becking and Beijerinck that "everything is everywhere, but the environment selects" (Baas Becking, 1934). This hypothesis, if true, would imply that there is a globally distributed seed bank of rare microbial lineages that bloom when conditions are optimal for those lineages. According to this scenario, most species are rare at some point and become abundant only under ideal conditions. Supporting this hypothesis, Gibbons et al. (2013) found that most microbial lineages identified in the International Census of Marine Microbes dataset could be found at a single deeply-sequenced site in the Western English Channel, which the authors interpreted as evidence for the existence of a microbial seed bank throughout the oceans. However, a similar global study found that bacteria exhibit a bipolar distribution such that microbial lineages were more geographically confined by hemisphere than would be expected from a null model in which lineages had an even global distribution (Sul *et al.*, 2013). Thus, while the rare biosphere of a given sample can consist of taxa that are found at various sites globally, the distribution of those taxa throughout the globe is not even.

The structure of microbial communities across sites suggests that the ecological role of the rare biosphere is more complex than simply acting as a seed bank. A study of rare and abundant operational taxonomic units (OTUs) in the Arctic Ocean found that rather than being widely dispersed, as would be expected if "everything were everywhere," rare OTUs exhibited similar geographic patterns to those of the abundant OTUs (Galand *et al.*, 2009). Therefore, the rare lineages must be subject to the same ecological processes or limits affecting the more abundant lineages. One study found that in the open ocean, over half the bacterial taxa cycled between being abundant and rare over the course of the seasons, but up to 12% of bacterial lineages were always rare (Campbell *et al.*, 2011). Whether most rare taxa are active is still unclear. One possible explanation for the existence of persistently rare strains is that certain taxa are kept at low abundances due to high susceptibility to viral infection (Bouvier and del Giorgio, 2007). Thus, some proportion of the rare taxa may be active, but high mortality rates or slow growth rates keep their abundances low.

The existence of a rare biosphere may simply reflect limitations in our sampling strategies. Many of the lineages that appear to be rare may in fact be abundant in microzones or over short time periods. Copiotrophic lineages can be associated with detrital particles, small animals, or mineral strata and are therefore highly abundant in concentrated regions, which would not be reflected in a homogenized fluid sample. For example, *Thermococcales* lineages are often rare in diffuse flow hydrothermal vent samples, yet are easily cultured from diffuse fluids, suggesting that these lineages are abundant in microzones where organic material is readily available, or have been transported from other regions where they dominate (Huber *et al.*, 2002, 2006; Summit and Baross, 2001). Thus, it is important to distinguish between organisms that may be rare but active, and others that are dominant elsewhere but rare in a particular sample.

An important caveat is that the 16S rRNA gene sequence does not necessarily reflect physiological diversity encoded on the rest of the genome. While a certain 16S sequence may be ubiquitous, individual phenotypes may be confined to particular regions. Marine Group II, for example, is known to encode proteorhodopsins in the photic zone, but not deeper in the water column (Frigaard *et al.*, 2006). Thus, the study of 16S sequences may not adequately address the “everything is everywhere” concept, because while a particular 16S sequence may be “everywhere,” the physiology of that lineage may not be universal.

Hydrothermal vent systems present a compelling test case for the rare biosphere and the “everything is everywhere” concept. Their diverse microhabitats and global distribution allow us to investigate the biogeography and dynamics of the rare biosphere on both the macro- and micro-scale. Hydrothermal vent systems are host to a wide variety of ecological niches that are produced by mixing of high-temperature, reduced, metal-enriched hydrothermal fluid with cooler oxidized seawater both above and below the seafloor. Diffuse flow vents, with fluid created by the mixing of cool seawater with hydrothermal fluid, host richly diverse microbial communities with members that range from deep subsurface hyperthermophiles to mesophiles and psychrophiles entrained from deep seawater (Huber *et al.*, 2003, 2002, 2007; Deming and Baross, 1993). Previous studies indicate that a portion of the microbial community found in diffuse fluids from hydrothermal vents draws from a deep subsurface habitat hosting thermophilic, anaerobic archaeal and bacterial communities; effectively, hydrothermal systems provide a "window" to the deep biosphere (Summit and Baross, 2001; Deming and Baross, 1993). Above the seafloor, focused hydrothermal flow travels through iron sulfide structures at high temperatures. These structures are inhabited by microbial communities that tend to be much more archaeal-dominated than the diffuse flow communities (Schrenk *et al.*, 2003; Takai and Horikoshi, 1999; Takai *et al.*, 2001; Slobodkin *et al.*, 2001; Kelley *et al.*, 2002). Archaeal communities in the hot interior of sulfide structures are dominated by uncultured crenarchaeal hyperthermophiles, which then give way to Marine Group I crenarchaea and uncultured euryarchaea in the cooler, more oxidized exterior of the chimney (Schrenk *et al.*, 2003).

With vastly different environmental conditions positioned in close proximity to each other, these habitats provide a natural laboratory from which to observe the influence of environmental parameters on microbial community structure. A comparison of *Thermococcus* isolates sampled from sulfide structures and diffuse flow fluids revealed that, while no physical barrier separates microbial communities in these two habitat types, there was a phylogenetic distinction between sulfide and diffuse flow *Thermococcus* isolates (Summit and Baross, 2001). Niche partitioning according to geochemical conditions and physical parameters creates differentially structured microbial communities across gradients in the vent system. For example, sulfur-oxidizing *Gammaproteobacteria* in the SUP05 clade dominate the microbial communities in the cooler, neutrally buoyant hydrothermal plumes above the vents, but were rarely observed in high temperature fluids (Anderson *et al.*, 2013). Deep sequencing of archaea and bacteria from seamount diffuse fluids revealed richly diverse communities, with over 680,000 bacterial lineages and over 216,000 archaeal lineages (Huber *et al.*, 2007). As in many other environments, the majority of the lineages in microbial communities from hydrothermal vent fluids occur in very low abundances (Huber *et al.*, 2007; Sogin *et al.*, 2006). However, the dynamics of the rare biosphere have not been investigated across the many niches of hydrothermal vents, nor have they been studied from a global perspective.

Basalt-hosted hydrothermal systems occur at mid-ocean ridges throughout the oceans, but can be separated by thousands of miles of deep seawater. They serve as excellent testing grounds for global biogeography because, while they are isolated from each other by hundreds to thousands of miles, they provide similar local selection pressures and are connected by ocean currents. While similar taxa are found at vent systems globally, the extent of organism dispersal and gene flow between these systems is not clear, nor is it known whether certain lineages in geologically separated hydrothermal vents have evolved into "ecotypes" to match local geochemistry. In the surface ocean, fine-scale differences in phylogenetic structure have revealed the formation of ecotypes in taxa such as *Prochlorococcus* (Rocap *et al.*, 2003) and SAR11 (Vergin *et al.*, 2013). If gene flow and dispersal among geographically separated vents occurs frequently, then we would expect to observe shared lineages at vents worldwide and limited phylogenetic differentiation. In contrast, restricted dispersal would most likely lead to the formation of distinct phylogenetic ecotypes between vents. The long-distance spread between systems, combined with the dominance of gradients within them, allows us to test the relative influence of geographic proximity and ecological niche partitioning on microbial community structure and the rare biosphere. Outstanding questions include: are rare OTUs always rare across all niches of hydrothermal systems, or do they dominate in certain conditions? Do rare and abundant lineages behave similarly in the archaeal and bacterial domains? How widely dispersed are archaeal and bacterial OTUs at vent sites globally? Is ecological niche or geographic location the stronger driver of community similarity? Does the deep subsurface act as a seeding reservoir for habitats connected by fluid flux, and is this reflected in the rare lineages present in these communities?

Here, we use basalt-hosted hydrothermal systems as a test case for closely examining the biogeographic paradigm of "everything is everywhere, but the environment selects" with respect to both rare and more abundant OTUs, across ecological niches within hydrothermal systems and across semi-isolated hydrothermal systems worldwide. We investigated all publicly available DNA sequences for the v6 region of 16S rRNA gene (pyrotags) from hydrothermal vent samples on the VAMPS database, which includes samples from diffuse flow fluids and vent sulfides worldwide, and combined this evaluation with analysis of a single, large-volume high-temperature sample that provides the best representation currently available of the deep subsurface biosphere in vent systems. We show that while both rare and abundant bacterial OTUs exhibit similar biogeographic patterns, abundant lineages of archaeal OTUs are widely dispersed. We also present an analysis of the high-temperature fluid sample that shows this representation of the deep hot biosphere to be distinct from the microbial communities found in the available diffuse flow and sulfide samples.

Materials and Methods

Sample site description and sampling procedures

All 16S v6 pyrotag datasets used in this study, with the exception of those newly acquired from the high-temperature sample, were obtained from the publicly available VAMPS database (www.vamps.mbl.edu). Table 2.1 presents a list of all samples and associated metadata; Figure 2.1 depicts a map of the sample sites.

The high-temperature fluid sample used in this study was collected at Hulk vent in the Main Endeavour Field on the Juan de Fuca Ridge, a spreading center located about 200 miles from the coast of Washington and Oregon (Figure 2.1). Hulk is a large sulfide chimney located at 47˚ 57.00' N, 129˚ 5.81' W. The sample was collected in August 2009 aboard the *R/V Atlantis*. A custom-built barrel sampler was deployed using *DSV Alvin* to collect 170 L of high-temperature diffuse flow fluid from the base of the sulfide structure. The average temperature of the sample was calculated from its silica and magnesium concentrations to have been about 125˚C (Anderson *et al.*, 2011). This sample most likely represents a wide range of niches, since it was placed on top of a colony of tube worms at approximately 20˚C, and was close to a fluid conduit measured to be about 300˚C. On deck, samples were put on ice prior to filtering through three 0.02 μm Steripaks (Millipore, USA). DNA extraction procedures are described in detail in Anderson et al. (2013). Briefly, one of the Steripaks was freeze-thawed three times, then DNA extraction buffer (0.1M Tris-HCl, 0.2M Na-EDTA, 0.1M NaH_2PO_4, 1.5M NaCl, and 1% cetyltrimethylammonium bromide), 50 mg ml^{-1} lysozyme, 1% proteinase K, and

20% SDS solution were added to the filter. DNA was extracted using phenol:chloroform:isoamyl alcohol and chloroform:isoamyl alcohol.

Other diffuse flow samples for this study were collected by J. Huber from eight different seamounts at Axial Seamount, the Mariana Arc, and Loihi Seamount, depicted in Figure 2.1. Axial Seamount is an active volcano located on the Juan de Fuca Ridge about 300 miles off the coast of Oregon. The caldera is about 700 m above the level of the ridge, and is bordered on three sides by a boundary fault. Several areas of active venting are located within the caldera. Other diffuse flow fluid samples were taken from several seamounts along the Mariana Arc, located in the Western Pacific Ocean from about 12 to 24˚N. All seamounts sampled (NW Eifuku, Daikoku, Nikko, NW Rota, and E Diamante) were located along the active front of the Mariana Arc, with the exception of Forecast, which may have greater influence from the backarc spreading axis (Huber *et al.*, 2010; Embley *et al.*, 2004). These samples were taken directly from the seafloor, rather than from sulfide structures (Huber *et al.*, 2010). Loihi Seamount is an active submarine volcano located above the Hawaiian hotspot; it differs from vent systems on plate boundaries in that the fluids tend to be enriched in carbon dioxide, iron, and methane and to contain low levels of sulfide (Moyer *et al.*, 1994). Diffuse flow samples, usually ranging between 10 and 50˚C, were collected with a sampling apparatus mounted aboard a remotely operated vehicle (ROV) that filtered the fluids through Sterivex samples *in situ*. Sampling methods for diffuse flow samples are the same as those discussed in (Huber *et al.*, 2010). DNA was extracted from Sterivex units according to the procedure described in Sogin et al. (2006).

All sulfide samples for this study were collected by A.-L. Reysenbach from the Lau Basin, which is a back-arc basin formed by the subduction of the Pacific plate below the Australian plate. Sulfide samples were collected with an ROV and placed in biobioxes after collection (Flores *et al.*, 2012, 2011). DNA was extracted from sulfide samples using the Ultra Clean Soil DNA Isolation Kit (MoBio Laboratories). Samples were sequenced as part of the International Census of Marine Microbes (ICoMM) initiative.

Sequencing

V4–v6 and v6 amplicon libraries for all samples were constructed and sequenced at the Josephine Bay Paul Center at the Marine Biological Laboratory on a Roche 454 GSFLX Titanium platform using the techniques described in (Huber *et al.*, 2007; Sogin *et al.*, 2006). All sequences are publicly available on the VAMPS website (http://vamps.mbl.edu) under dataset names REA_HDF_Av6v4, REA_HDF_Bv6v4 and REA_HDF_Bv6 for the Hulk sample and under project names ICM_ALR for all sulfide samples and KCK_SMT for all diffuse flow fluid samples used in this study.

Bioinformatic analysis

All reads used in this study were trimmed and quality filtered through the VAMPS pipeline using the quality control parameters outlined in Huse et al. (2007). We analyzed samples in the projects ICM_ALR_Av6, ICM_ALR_Bv6, KCK_SMT_Av6, and KCK_SMT_Bv6. Taxonomic analyses of each sample were performed using the GAST process in VAMPS (Huse *et al.*, 2008). Sequences from each sample were further screened, filtered, and trimmed as a batch set with all samples included in this analysis using mothur (Schloss *et al.*, 2009). The trimming of sequences removed the v4–v5 region of the Hulk sequences, leaving behind only the v6 region to facilitate direct comparison. Both archaeal and bacterial sequences were aligned against the SILVA database (Quast *et al.*, 2013) through the mothur pipeline. Sequences were clustered into operational taxonomic units (OTUs) using average-neighbor hierarchical clustering to the 0.03 level using mothur. Diversity indices (rarefaction curves, Shannon and Simpson evenness) were calculated in the mothur pipeline. For comparison between samples, distance matrices were constructed in mothur using the Bray-Curtis calculator of community membership and structure. Cluster dendrograms were generated from these distance matrices using PRIMERv6 (Clarke and Gorley, 2006).

For the rare vs. abundant OTU analysis, we separated sequences from OTUs that were considered to be abundant in each sample (representing equal to or greater than 1% of all sequences in the sample) from those considered to be rare (representing equal to or less than 0.1% of all sequences in the sample). Analysis of similarity (ANOSIM) tests were carried out using PRIMERv6 to determine whether there were assemblage differences between groups of samples specified according to geographic location. Nine hundred ninety nine permutations of the test were conducted for each ANOSIM analysis, using a resemblance matrix of Bray-Curtis dissimilarity as determined in mothur. The statistical software package R (R Core Team, 2013) was used to generate heatmaps with data on OTU relative abundance generated in mothur.

To create phylogenetic trees of sequences from the *Thermococcales* and *Methanococcales*, we identified all OTUs belonging to either of these groups according to the SILVA taxonomic classification conducted in mothur, and selected a reference sequence from each OTU. We created reference data sets with full-length 16S sequences from the SILVA database (Quast *et al.*, 2013); both the reference sequences and sample sequences were aligned in the SILVA aligner (Pruesse *et al.*, 2012). We created a base tree from the reference sequences in RAxML (Stamatakis, 2006) using a rapid bootstrap analysis to search for the best maximum likelihood tree with 100 alternative runs on distinct starting trees. We used EPA (Berger *et al.*, 2011) within the RAxML package to insert the short v6 sample sequences into the base tree. For tree construction, we used the GTR+ optimization of substitution rates and the GAMMA model of rate heterogeneity.

Comparisons between v4–v6 sequences in the Hulk sample and uncultured crenarchaeal sequences were conducted with USEARCH v6 (Edgar, 2010) using the usearch_global command.

Results

Comparative community structure of diffuse flow and sulfide structures

Taxonomic classification of bacterial v6 sequences for all samples revealed high abundances of both the *Epsilonproteobacteria* and *Gammaproteobacteria* groups in most samples for both diffuse flow and sulfide samples (Figure 2.2A). *Alphaproteobacteria* and *Betaproteobacteria* were also commonly observed groups. No clear distinctions emerged between sulfide and diffuse flow samples at this taxonomic resolution, except for a slightly higher abundance of *Epsilonproteobacteria* in sulfide samples compared to diffuse flow samples, which had slightly higher abundances of *Gammaproteobacteria* and *Alphaproteobacteria*. Taxonomic classification of archaeal v6 samples revealed a much greater distinction between sulfide and diffuse flow samples: sulfide samples exhibited high abundances of *Archaeoglobi* and *Halobacteria*, whereas Marine Group I and *Thermoplasmata* dominated most diffuse flow samples (Figure 2.2B). The high-temperature Hulk sample, in contrast, was unique in its high abundance of *Thermococcales*.

All sequences were clustered into OTUs at a 3% distance, for a total of 3711 OTUs in the archaea and 22029 OTUs in the bacteria. In almost all samples, bacterial communities had higher richness than the archaeal communities, as depicted in the rarefaction curves for samples from both domains (Figure 2.3). None of these rarefaction curves reached an asymptote, indicating that none of the datasets captured the total diversity of the sample. No clear patterns emerged regarding the relative richness of the different types of samples, though several diffuse flow samples had higher richness than any of the sulfide samples for both domains. The rarefaction curve for the high-temperature Hulk sample fell in the middle of the distribution for both archaeal and bacterial sequences. Simpson and Shannon evenness indices showed that bacterial communities had higher evenness than archaeal communities, indicating that archaeal communities had a greater tendency to be dominated by a small number of abundant OTUs (Table 2.2). This pattern was consistent across sample types.

Clustering of samples according to community similarity

Cluster dendrograms indicated the degree of similarity between samples based on the relative abundance of OTUs. For the bacteria, some clustering according to seamount was apparent among the diffuse flow samples (Figure 2.4A). Even background seawater samples grouped according to geographic location, indicating that the bacterial community associated with the vent habitat disperses into background water fairly easily. For the archaea, there was much higher similarity between diffuse flow samples and between sulfide samples than observed for the bacteria. Archaea in diffuse flow samples also showed fewer tendencies to cluster by location (Figure 2.5A). While these dendrograms were constructed based on OTUs at a 3% distance, cluster dendrograms of OTUs grouped at a 4% or 2% distance or for unique sequences showed few qualitative differences from the 3% distance OTUs (Figure 2.6 for bacteria and 2.7 for archaea) (OTU dendrograms at the 4% distance were not created for the bacterial domain because there was no difference between 3% and 4% OTUs). In most cases, the sulfide structures clustered separately from the diffuse flow samples. The high-temperature Hulk sample grouped with the other diffuse flow samples, though at very low similarity.

We used ANOSIM analyses to test the hypothesis that archaeal and bacterial samples clustered according to geographic location. Diffuse flow samples were grouped according to seamount, while sulfide samples, which were all collected in the Lau Basin, were grouped together. The high temperature Hulk sample, from the Main Endeavor Field, was grouped separately from other samples. Location designations are shown in the legends of Figures 2.4 and 2.5. ANOSIM analysis indicated that clustering according to geographic location was significant for the bacteria ($p \leq 0.1\%$), but not for the archaea (Table 2.3). However, while there was significant clustering of samples according to seamount, there did not appear to be a tendency to group according to general region: for example, samples from the Philippine Sea did not necessarily cluster separately from Axial samples.

Biogeography and distribution of rare and abundant OTUs in hydrothermal systems

Analyses of all OTUs together cannot identify differences in ecological patterning between the rare and abundant OTUs. Therefore, we separated the rare and abundant OTUs within each sample to determine whether they exhibit different biogeographic and community structuring patterns. For this analysis, rare OTUs were considered to be those OTUs representing less than or equal to 0.1% of the sequences in the sample; abundant OTUs were considered to be those OTUs representing greater than or equal to 1% of all the sequences in the sample. This scoring follows definitions of rare and abundant groups previously established by Pedros-Alió (2006) and Fuhrman (2009). The taxonomic identification of rare and abundant OTUs did not differ drastically from each other, though certain OTUs had a greater tendency to be either rare or abundant. There was a slightly higher percentage of *Thermoplasmata* and Marine Group I among abundant archaeal OTUs than among rare OTUs, though the difference is not large (Figure 2.8). Similarly, *Gammaproteobacteria* and *Epsilonproteobacteria* comprised a higher percentage of the abundant bacterial OTUs than the rare OTUs (Figure 2.9).

We identified the rare and abundant OTUs in each sample and clustered the samples according to community similarity as before to determine whether the rare and abundant OTUs exhibited similar biogeographic patterns. In the bacteria, biogeographic patterning was similar, though not identical, between the rare and the abundant OTUs (Figure 2.4 B,C). General groupings can be seen according to seamount within the diffuse flow samples. Overall, rare OTUs showed less similarity from sample to sample (averaging about 80% dissimilarity) compared to abundant OTUs (averaging 60–70% dissimilarity). ANOSIM analyses for bacteria indicated that clustering according to seamount was significant in all cases (Table 2.3).

For archaeal lineages, different patterns appeared. Abundant OTUs showed much higher similarity between samples than rare OTUs. For abundant OTUs the samples were 20–30% dissimilar on average, but for rare OTUs the samples averaged approximately 80% dissimilarity. While samples with all OTUs did not cluster according to geographic location, separating the rare and abundant OTUs revealed that this lack of biogeographic patterning was driven almost entirely by the abundant OTUs (Figure 2.5 B, C). In contrast to the bacterial case, the only archaeal OTUs to cluster significantly by geographic location were the rare OTUs (Table 2.3). An extremely low R statistic and high p-value indicated that abundant OTUs showed almost no tendency to group according to geographic location, especially when considering only diffuse flow samples; this result for abundant OTUs was most likely responsible for the lack of significant clustering by location for all OTUs analyzed together. The high-temperature Hulk fluid sample grouped with the sulfide samples only when examining the rare OTUs.

Persistence of OTUs across samples

The distinctive differences in biogeographic distribution between the rare and abundant archaeal OTUs raises the question of how widely these abundant OTUs are distributed, and to what extent rare OTUs are unique to single samples. Figure 2.10 depicts the percent of OTUs found in different numbers of groups; 66% of the archaeal OTUs and 69% of the bacterial OTUs were found only in one sample. The asymptote to the right of the graph indicates that a low percentage of OTUs was found in multiple groups. OTUs that were abundant in at least one sample (depicted graphically in the inset of Figure 2.10) were similarly confined to a single sample, in general. However, in both graphs, the line for the archaea did not drop off as quickly as the bacterial line in the figure inset, indicating that the archaeal OTUs tended to be found in a greater percentage of samples. This result suggests that a higher percentage of archaeal OTUs tended to be more widely dispersed among samples than bacterial OTUs.

As suggested in Figure 2.5, the widely dispersed archaeal OTUs tended to be those that were more abundant within samples. This pattern is illustrated visually in the heatmap in Figure 2.11, depicting the relative abundance of archaeal OTUs from the high-temperature Hulk sample across all other samples analyzed here. Generally, abundant OTUs in Hulk were more likely to be abundant or at least present in other samples, while rare OTUs were more likely to be rare or undetected across samples. OTU 3147, for example, a member of the Marine Group I crenarchaea, was abundant in almost all diffuse flow samples. An exception to this trend was the most abundant archaeal OTU in the Hulk sample, OTU 3645, a *Thermococcus* sequence that comprised 63% of the sample. While found in other samples, this OTU did not reach such a high abundance in any other sample we examined, reaching a maximum of only 9% in one sulfide sample (sulf_20) while being rare or absent in most diffuse flow samples. Similar patterns were observed for the bacterial domain (Figure 2.12). As with the archaeal domain, abundant bacterial OTUs were more consistently present across multiple samples than rare OTUs.

Phylogenetics of the Thermococcales and Methanococcales

Clustering samples into OTUs does not give an indication of phylogenetic relatedness between sequences and across samples, yet understanding phylogenetic relatedness can provide another layer of insight into the similarity and gene flow between samples. *Thermococcales* are a general indicator of high-temperature fluids and were dominant in the Hulk sample. Thus we created a phylogenetic tree of sequences falling within the order *Thermococcales* to investigate relationships between samples that might be based on temperature, especially the high-temperature Hulk sample and the sulfide samples (Figure 2.13). Both the sulfide samples and the Hulk sample had high diversity within the *Thermococcales* order, with several OTUs falling into many different clades in the tree. It was much more common for sulfide and Hulk sequences to group together into the same OTU or branch (at 3% distance) than it was for diffuse flow sequences to group with sequences from sulfides or Hulk. Fewer *Thermococcales* OTUs were found in diffuse flow samples overall; those that were present tended to cluster into a few clades on the tree, particularly within the *Palaeococcus* genus, or to group on branches with no cultured representatives. OTU 3645, the OTU that dominated the Hulk sample, was found in all three sample types.

Similar results were found in a phylogenetic tree of OTUs falling in the *Methanococcales* order, though these OTUs were found with greater frequency in diffuse flow samples (Figure 2.14). The separation between sulfide and diffuse flow OTUs was more distinct in this tree. The OTU from the Hulk sample fell within a clade shared with other sulfide OTUs, despite being geographically closer to Axial Volcano, where most of the *Methanococcales* OTUs were found. Both samples FS317 and FS521 are from the Marker 113 vent at Axial, a vent known historically to host high abundances of methanogens (Huber *et al.*, 2009). The two Marker 113 samples were dominated by a single methanogen OTU, labeled here as OTU 2977.

Given the phylogenetic similarities between sulfide sample and the high-temperature sample, we also sought to determine whether abundant sequences in sulfides matched rare sequences in the Hulk high-temperature sample. Because previous work has indicated that uncultured crenarchaea dominate the interior of sulfide structures (Schrenk *et al.*, 2003), we conducted global sequence comparisons of the Hulk high-temperature v4–v6 region against a database of uncultured crenarchaea identified from sulfides. Two sequences were found that matched previously identified crenarchaea in sulfides at 99–100% identity: a *Desulfurococcales* lineage from a white smoker spire on the East Pacific Rise (Kormas *et al.*, 2006), and a *Pyrodictium* lineage identified in an in-situ growth chamber deployed within a sulfide structure (Nercessian *et al.*, 2003).

Discussion

Domain differences in the ecology of rare and abundant lineages of the bacteria and archaea

The advent of pyrosequencing has allowed us to probe more deeply into the structuring of microbial communities across ecological niches in both the archaeal and bacterial domains. Most studies in hydrothermal systems until now have focused on diversity and community structuring in hydrothermal systems, but without contrasting biogeographic structuring patterns between rare and abundant OTUs across niches or vent sites. Our results show that rare and abundant OTUs have different biogeographic patterns across sample types, a distinction particularly strong in the archaeal domain.

For bacteria, microbial community structuring according to geographic location applied to both the rare and abundant lineages for the bacteria. Cluster dendrograms showed similar patterns for both rare and abundant bacterial lineages, with a tendency for samples from the same seamounts to cluster together. Community structure patterns of rare lineages were slightly more distinct between samples, as evidenced by the higher dissimilarity between samples. However, most of the rare lineages were either rare or absent across samples, leaving unclear whether these lineages were persistently rare or geographically restricted, such that a more comprehensive sampling effort might have detected them in higher abundances or in other locations.

A contrasting picture was observed for the archaea. Among archaeal OTUs, two trends appeared. First, the most successful archaeal lineages tended to dominate a community to a greater degree than was the case for bacteria, as evidenced by the high unevenness of archaeal communities relative to bacterial communities. Other studies of hydrothermal archaeal communities, such as the case of *Methanosarcinales*-dominated biofilms found at Lost City (Brazelton *et al.*, 2006), are consistent with this trend. Second, abundant archaeal OTUs were widely dispersed with little evidence of distinct structuring between samples. In contrast, rare OTUs were subject to greater biogeographical constraint; samples clustered according to geographic location, and community structure was not highly similar between samples for rare OTUs. The results suggest an ecological pattern in which a few abundant archaeal OTUs dominate and are widespread, whereas the majority of archaeal OTUs are rare and biogeographically restricted. Thus abundant archaeal OTUs seem to follow the paradigm “everything is everywhere,” but the rare OTUs do not.

These data indicate that some archaeal OTUs are well-adapted to the vent environment and are widespread. The most abundant and widespread archaeal OTUs in diffuse fluids belonged to Marine Groups I and II, which are native to deep seawater and therefore can more easily travel in ocean currents from one vent system to the next. These particular lineages of Marine Groups I and II may have been ecotypes that gained a fitness advantage through some means, such as horizontal gene transfer, that allowed them to proliferate rapidly.

However, it is unclear why certain abundant archaeal OTUs are so widely dispersed, while abundant bacterial OTUs are more biogeographically restricted. Archaea generally exhibit lower richness compared to bacteria in various environments globally (Aller and Kemp, 2008), potentially because they are less physiologically flexible or tend to specialize in low-energy environments (Valentine, 2007). We speculate that specific archaeal OTUs gaining fitness advantages through horizontal gene transfer or mutation may gain advantages through physiological flexibility, thus allowing them to dominate over other archaeal lineages. Intra-species genetic diversity through gene transfer and recombination has been observed in natural archaeal populations (Allen *et al.*, 2007), a fact that combined with the low diversity of natural archaeal populations suggests that the archaeal pangenome is quite extensive. Similarity in the 16S sequence may not necessarily indicate similarity in genome sequence or physiology. This distinction may pertain especially for thermophilic archaea, given that high rates of horizontal gene transfer have been observed among thermophiles (Koonin *et al.*, 2001; Beiko *et al.*, 2005). Thus, while the same OTUs were observed across multiple vent sites, it is possible that there was a range of physiological variation within those OTUs from one site to the next that was not discernible by examining only the 16S v6 sequence. Further research involving comparisons of full genome sequences will provide insight into this possibility.

Aside from the cosmopolitanism of these abundant strains, however, the overall trend observed for all lineages in the bacteria and for rare lineages in the archaea indicates biogeographic dispersal limitation. These results run counter to a simple interpretation of the “everything is everywhere, but the environment selects” model because they indicate a strong degree of biogeographic restriction, even for the rare strains.

Community structuring within hydrothermal niches and the deep subsurface

The gradients within deep-sea hydrothermal systems, created by the mixing of hydrothermal fluid and deep seawater, establish multiple ecological niches that foster high microbial diversity. The results of our study indicate that niche partitioning among diffuse flow, sulfide, and high-temperature fluid settings occurs for both the rare biosphere and the more abundant microbial lineages. Specific examination of the Hulk high-temperature sample provides insight into dispersal and niche colonization in vent environments. While the community structure of the Hulk high-temperature sample was generally more similar to diffuse fluid samples than to sulfide samples, the rare archaeal lineages of the high-temperature sample were slightly more similar to those of the sulfides than to diffuse flow. Closer examination of the thermophilic archaeal groups *Thermococcales* and *Methanococcales,* as well as uncultured crenarchaea*,* indicated that these lineages in the high-temperature Hulk fluid sample were similar to those found in sulfide samples. Microbial communities in sulfides are exposed to more focused hydrothermal fluid flow, and thus higher temperatures, than microbial communities in diffuse flow. Therefore, while a general overall community similarity appears between fluid samples, certain rare archaeal sequences in the high-temperature sample were more similar to those found in sulfides than in diffuse flow fluids. This finding most likely reflects niche selection according to temperature, and may also point to seeding from the hot deep subsurface.

In deep-sea hydrothermal systems, a deep subsurface reservoir of hydrothermal fluid flowing through fissures and porous crust is thought to support a deep biosphere microbial community that is occasionally flushed from porous structures and mineral surfaces by the rising fluids, which can then can be sampled from hydrothermal fluids emerging from the seafloor. Thus, hydrothermal vents are "windows" to the deep subsurface biosphere (Deming and Baross, 1993). The high temperature and high abundance of thermophilic *Thermococcales* in the Hulk sample suggest that it represents the microbial community found in high-temperature niches, including the deep hot subsurface. Overall, clustering dendrograms for both archaeal and bacterial OTUs (Figures 2.4 and 2.5) indicate that the high-temperature Hulk sample was strikingly different from both sulfide and diffuse flow samples from all biogeographic provinces. However, similarities between the high-temperature *Thermococcales* in the Hulk sample and rare archaea in the sulfide samples suggest that many of the rare lineages were rare only in those samples, likely diluted during transport away from other habitats where they may have been dominant. This scenario depicts the microbial community in the deep hot subsurface as occupying a unique niche, distinct from that of the diffuse flow and sulfide microbial communities found more distantly in the fluid flow, but one which we can glimpse through the rare biosphere.

Gene flow and community structuring at hydrothermal vents across the globe

Comparing microbial community structure across the globe also allowed us to gain insight into gene flow and biogeographic distribution of rare and abundant OTUs throughout the ocean basins. Generally, the bacterial and rare archaeal ecotypes appeared to be seamount-specific. Given that location was a stronger indicator of community similarity than chemistry in previous work (Opatkiewicz *et al.*, 2009), this result most likely reflects restricted dispersal. However, as discussed above, the more abundant lineages appeared to be more widely dispersed among samples, even across ocean basins, implying some degree of transport of organisms between vent sites. Deep ocean currents provide a source of connectivity between vent systems, potentially explaining the incidence of shared OTUs between geographically distant vent systems.

The overall picture that emerges is a complex interplay between niche specialization, geographic location, and reproductive success. Diffuse fluids tended to group according to seamount, with certain successful lineages displaying high abundance and wide geographic dispersal. Rare archaeal lineages showed high biogeographic restriction, but rare archaeal lineages from a deep hot sample showed similarities to sulfide samples, indicating potential seeding from the deep hot biosphere. Finally, the successful archaeal lineages tended to dominate within and across samples to a much greater degree than the successful bacterial lineages, implying a fundamental difference in evolutionary mechanisms or behavioral response between the two domains. Future work can reveal the extent of genome heterogeneity within OTU groupings, and whether ostensibly rare groups are actually dominant in microzones within porous minerals or particle aggregates. Taken together, these analyses present a more nuanced view of the rare biosphere, and urge caution with overly broad applications of the concept that "everything is everywhere."

Table 2.1. Metadata for all samples used in this study. Metadata for publicly available samples were obtained from the VAMPS database.

Sample name	Sample type	Depth (m)	Location	Latitude	Longitude	Domain	Temp (˚C)
sulf_01	active sulfide chimney	2707	Lau Basin, South Pacific Ocean	–20.316686	–176.1363	Bacteria	2.712
sulf_02	active sulfide chimney	2714	Lau Basin, South Pacific Ocean	–20.317851	–176.13737	Bacteria	2.712
sulf_03	active sulfide chimney	2139	Lau Basin, South Pacific Ocean	–20.761027	–176.19081	Bacteria	2.712
sulf_04	active sulfide chimney	1908	Lau Basin, South Pacific Ocean	–22.180673	–176.60124	Bacteria	2.736
sulf_05	microbial mat	1918	Lau Basin, South Pacific Ocean	–22.180185	–176.60081	Bacteria	2.736
sulf_06	active sulfide flange	1918	Lau Basin, South Pacific Ocean	–22.180185	–176.60081	Bacteria	2.736
sulf_07	active sulfide flange	1875	Lau Basin, South Pacific Ocean	–21.989609	–176.56809	Bacteria	2.731
sulf_08	active sulfide chimney	2619	Lau Basin, South Pacific Ocean	–20.053045	–176.13374	Bacteria	2.706
sulf_08	active sulfide chimney	2707	Lau Basin, South Pacific Ocean	–20.316686	–176.1363	Archaea	2.712
sulf_10	active sulfide chimney	2714	Lau Basin, South Pacific Ocean	–20.317851	–176.13737	Archaea	2.712
sulf_12	active sulfide chimney	1908	Lau Basin, South Pacific Ocean	–22.180673	–176.60124	Archaea	2.736
sulf_16	active sulfide chimney	2619	Lau Basin, South Pacific Ocean	–20.053045	–176.13374	Archaea	2.706
sulf_17	active sulfide chimney-bottom	2707	Lau Basin, South Pacific Ocean	–20.316686	–176.1363	Bacteria	2.712
sulf_18	active sulfide chimney-bottom	2707	Lau Basin, South Pacific Ocean	–20.316686	–176.1363	Archaea	2.712
sulf_19	active sulfide chimney-top	2707	Lau Basin, South Pacific Ocean	–20.316686	–176.1363	Bacteria	2.712
sulf_20	active sulfide chimney-top	2707	Lau Basin, South Pacific Ocean	–20.316686	–176.1363	Archaea	2.712
FS317	Hydrothermal fluids	1526	Axial Volcano, North Pacific Ocean	45.9227283	–129.9882383	Both	26.6
FS389	Hydrothermal fluids	1546	Axial Volcano, North Pacific Ocean	45.933583	–130.013583	Both	32.5
FS392	Hydrothermal fluids	1546	Axial Volcano, North Pacific Ocean	45.9337	—130.013617	Both	68.2
FS430	Hydrothermal fluids	1449	Forecast, Philippine Sea	13.394633	143.920096	Both	71
FS431	Hydrothermal fluids	1448	Forecast, Philippine Sea	13.394632	143.920083	Both	6
FS432	Hydrothermal fluids	1451	Forecast, Philippine Sea	13.39532	143.919902	Both	6.5
FS433	Hydrothermal fluids	1447	Forecast, Philippine Sea	13.395265	143.919873	Both	40
FS434	Background seawater	195	Forecast, Philippine Sea	13.3811	143.9021	Both	
FS435	Background seawater	1342.5	Forecast, Philippine Sea	13.3811	143.9021	Both	
FS445	Hydrothermal	560	NW Rota, Philippine	14.600912	144.775483	Both	19.7

	fluids		Sea				
FS446	Hydrothermal fluids	534	NW Rota, Philippine Sea	14.60085	144.77632	Both	48
FS447	Hydrothermal fluids	521	NW Rota, Philippine Sea	14.601177	144.775618	Both	29
FS448	Hydrothermal fluids	584	NW Rota, Philippine Sea	14.60084	144.7773	Both	25
FS449	Hydrothermal fluids	568	NW Rota, Philippine Sea	14.60081	144.77751	Both	15.1
FS462	Hydrothermal fluids	353	E Diamante, North Pacific Ocean	15.94277	145.68141	Both	22.5
FS467	Hydrothermal fluids	1612	Eifuku, North Pacific Ocean	21.48742	144.04163	Both	42.9
FS468	Hydrothermal fluids	1578	Eifuku, North Pacific Ocean	21.487248	144.042123	Both	45.1
FS469	Hydrothermal fluids	1578	Eifuku, North Pacific Ocean	21.487248	144.042123	Both	33.7
FS473	Hydrothermal fluids	438	Daikoku, North Pacific Ocean	21.324536	144.19293	Both	15.3
FS475	Hydrothermal fluids	414	Daikoku, North Pacific Ocean	21.324962	144.19139	Both	45.5
FS479	Hydrothermal fluids	458	Nikko, North Pacific Ocean	23.081017	142.325483	Both	80.2
FS480	Hydrothermal fluids	445	Nikko, North Pacific Ocean	23.07913	142.326433	Bacteria	24.1
FS481	Hydrothermal fluids	413	Nikko, North Pacific Ocean	23.07977	142.32687	Archaea	32.6
FS482	Background seawater	344	Nikko, North Pacific Ocean	23.077802	142.325151	Both	14.7
FS501	Background seawater	1526	Axial Volcano, North Pacific Ocean	45.94667	–129.98439	Bacteria	2.4
FS502	Hydrothermal fluids	1529	Axial Volcano, North Pacific Ocean	45.94632	–129.98398	Archaea	83.4
FS503	Hydrothermal fluids	1530	Axial Volcano, North Pacific Ocean	45.94364	–29.98519	Both	
FS505	Hydrothermal fluids	1524	Axial Volcano, North Pacific Ocean	45.93319	–129.98223	Both	
FS509	Hydrothermal fluids	1546	Axial Volcano, North Pacific Ocean	45.93357	–130.01329	Both	24.6
FS510	Hydrothermal fluids	1546	Axial Volcano, North Pacific Ocean	45.93331	–130.01334	Both	49
FS511	Hydrothermal fluids	1546	Axial Volcano, North Pacific Ocean	45.93364	–130.01329	Both	96.8
FS518	Hydrothermal fluids	1546	Axial Volcano, North Pacific Ocean	45.93357	–130.01329	Both	
FS519	Hydrothermal fluids	1538	Axial Volcano, North Pacific Ocean	45.91724	–129.99299	Both	29.6
FS520	Hydrothermal fluids	1536	Axial Volcano, North Pacific Ocean	45.91631	–129.98916	Both	14.9
FS521	Hydrothermal fluids	1524	Axial Volcano, North Pacific Ocean	45.92279	–129.98838	Both	27.5
LOIHI-PP1	Hydrothermal fluids	1272	Loihi Seamount, North Pacific Ocean	18.900833	–155.261389	Both	
LOIHI-PP2	Hydrothermal fluids	1302	Loihi Seamount, North Pacific Ocean	18.910278	–155.25111	Both	
LOIHI-PP4	Hydrothermal fluids	4983	Loihi Seamount, North Pacific Ocean	18.703056	–155.180833	Bacteria	
LOIHI-PP5	Hydrothermal fluids	1308	Loihi Seamount, North Pacific Ocean	18.31222	–155.26111	Both	
LOIHI-PP6	Hydrothermal fluids	4988	Loihi Seamount, North Pacific Ocean	18.703056	–155.180833	Both	

LOIHI-CTD03	Background seawater	1100	Loihi Seamount, North Pacific Ocean	18.911667	–155.26194	Bacteria	
CTDBtl 12	Background seawater		NW Rota, Philippine Sea	14.644167	–144.56667	Bacteria	6.3
Hulk	Hydrothermal fluids	2178	Juan de Fuca Ridge, North Pacific Ocean	47.9500	–129.0968	Both	125

Table 2.2. Evenness indices for each of the samples used in this study. Both the Shannon and Simpson indices are reported for both domains. All indices were calculated in mothur (Schloss *et al*., 2009).

	Bacteria		**Archaea**	
Group	*Simpson*	*Shannon*	*Simpson*	*Shannon*
DF_Ax_FS317	0.037	0.73	0.0096	0.31
DF_Ax_FS389	0.012	0.66	0.011	0.38
DF_Ax_FS392	0.012	0.51	0.017	0.21
DF_Ax_FS501	0.010	0.20		
DF_Ax_FS502	0.0060	0.16	0.010	0.36
DF_Ax_FS503	0.0041	0.40	0.0092	0.29
DF_Ax_FS505	0.011	0.59	0.010	0.36
DF_Ax_FS509	0.017	0.62	0.035	0.54
DF_Ax_FS510	0.025	0.63	0.010	0.31
DF_Ax_FS511	0.021	0.55	0.020	0.29
DF_Ax_FS518	0.061	0.73	0.027	0.50
DF_Ax_FS519	0.035	0.72	0.010	0.26
DF_Ax_FS520	0.031	0.73	0.019	0.16
DF_Ax_FS521	0.037	0.71	0.015	0.22
DF_NP_FS467	0.028	0.66	0.019	0.50
DF_NP_FS468	0.011	0.64	0.010	0.35
DF_NP_FS469	0.037	0.74	0.018	0.48
DF_NP_FS473	0.0037	0.48	0.013	0.25
DF_NP_FS475	0.012	0.49	0.019	0.28
DF_NP_FS479	0.023	0.63	0.012	0.21
DF_NP_FS480	0.031	0.60		
DF_NP_FS481	0.026	0.66	0.010	0.36
DF_NP_FS482	0.022	0.71	0.014	0.39
DF_Lo_PP1	0.026	0.63	0.018	0.47
DF_Lo_PP2	0.031	0.69	0.016	0.49
DF_Lo_PP4	0.015	0.62		
DF_Lo_PP5	0.020	0.65	0.0036	0.24
DF_Lo_PP6	0.023	0.66	0.024	0.43
DF_PS_FS430	0.079	0.74	0.022	0.26
DF_PS_FS431	0.0085	0.66	0.024	0.43
DF_PS_FS432	0.024	0.75	0.019	0.31
DF_PS_FS433	0.039	0.70	0.019	0.13
DF_PS_FS434	0.011	0.60	0.044	0.49
DF_PS_FS435	0.0098	0.60	0.018	0.40
DF_PS_FS445	0.045	0.69	0.048	0.45
DF_PS_FS446	0.012	0.47	0.023	0.36
DF_PS_FS447	0.024	0.68	0.015	0.36
DF_PS_FS448	0.034	0.72	0.016	0.50
DF_PS_FS449	0.027	0.66	0.029	0.52

DF_PS_FS462	0.0060	0.48	0.0079	0.24
Hulk high temp	0.020	0.60	0.011	0.33
sulf_01	0.0065	0.44		
sulf_02	0.023	0.57		
sulf_03	0.037	0.71		
sulf_04	0.019	0.63		
sulf_05	0.012	0.55		
sulf_06	0.011	0.60		
sulf_07	0.0071	0.46		
sulf_08	0.037	0.68		
sulf_09			0.039	0.52
sulf_10			0.015	0.27
sulf_12			0.022	0.33
sulf_16			0.016	0.18
sulf_17	0.0073	0.42		
sulf_19	0.021	0.57		
sulf_20			0.029	0.47
CTDBTL12	0.048	0.72		
LOIHI_CTD03	0.016	0.58		
Average	**0.023**	**0.61**	**0.019**	**0.35**

Table 2.3. ANOSIM results for bacterial and archaeal datasets, grouped according to environment as listed in Figures 2.4 and 2.5. ANOSIM was conducted as a one-way analysis on a resemblance matrix of Bray-Curtis dissimilarity among samples. P-values here are reported in percent. A test is considered significant if $p \leq 0.1\%$.

Domain	**Grouping**	**With sulfides**			**Without sulfides**		
		R statistic	*p-value*	*Significant?*	*R statistic*	*P-value*	*Significant?*
Bacteria	All	0.574	<0.1	Yes	0.563	<0.1	Yes
	Abundant	0.534	<0.1	Yes	0.527	<0.1	Yes
	Rare	0.552	<0.1	Yes	0.476	<0.1	Yes
Archaea	All	0.288	0.4	No	0.084	19.8	No
	Abundant	0.246	1.1	No	0.031	38.6	No
	Rare	0.558	<0.1	Yes	0.442	<0.1	Yes

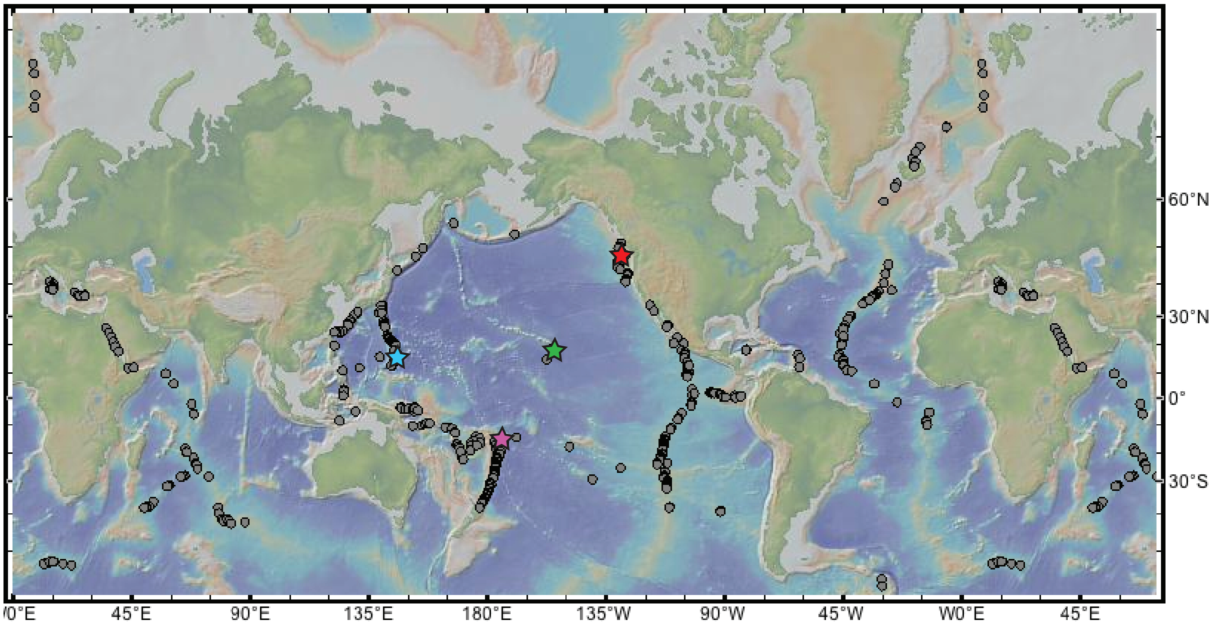


Figure 2.1. Approximate locations of sampling sites at hydrothermal vents worldwide. Red star indicates Main Endeavour Field and Axial Seamount, Juan de Fuca Ridge; blue star, Eifuku, Daikoku, Nikko, Forecast, NW Rota, and E Diamante seamounts, and Mariana Arc; green star, Loihi Seamount, Hawaii; and purple star, Lau Basin. Map was generated using GeoMapApp (http://www.geomapapp.org/).

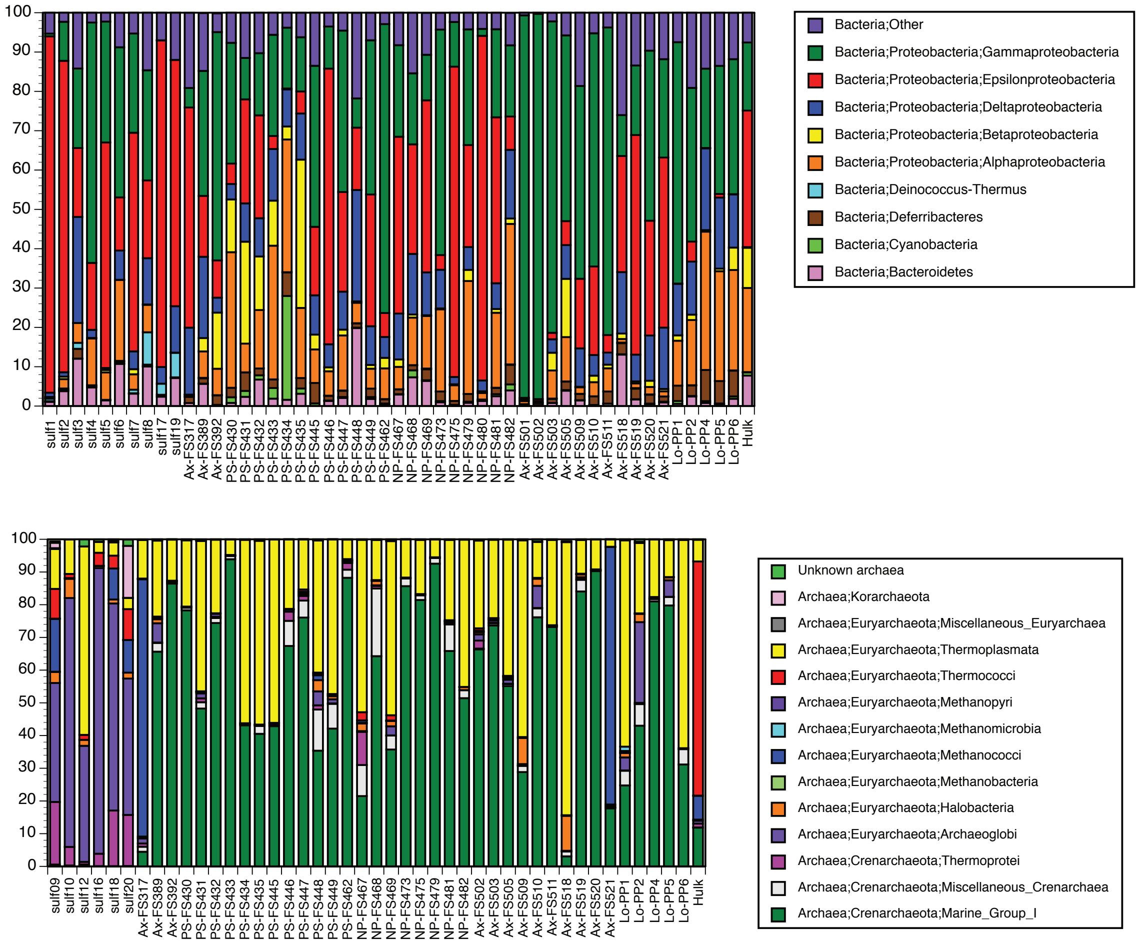


Figure 2.2. Bar charts of bacterial (A) and archaeal (B) taxonomy for all samples. Taxonomy was assigned in VAMPS by the GAST process (Huse *et al*., 2008). Hulk archaeal sample is classified based on v4–v6 sequence; all others are classified based on v6 sequence.

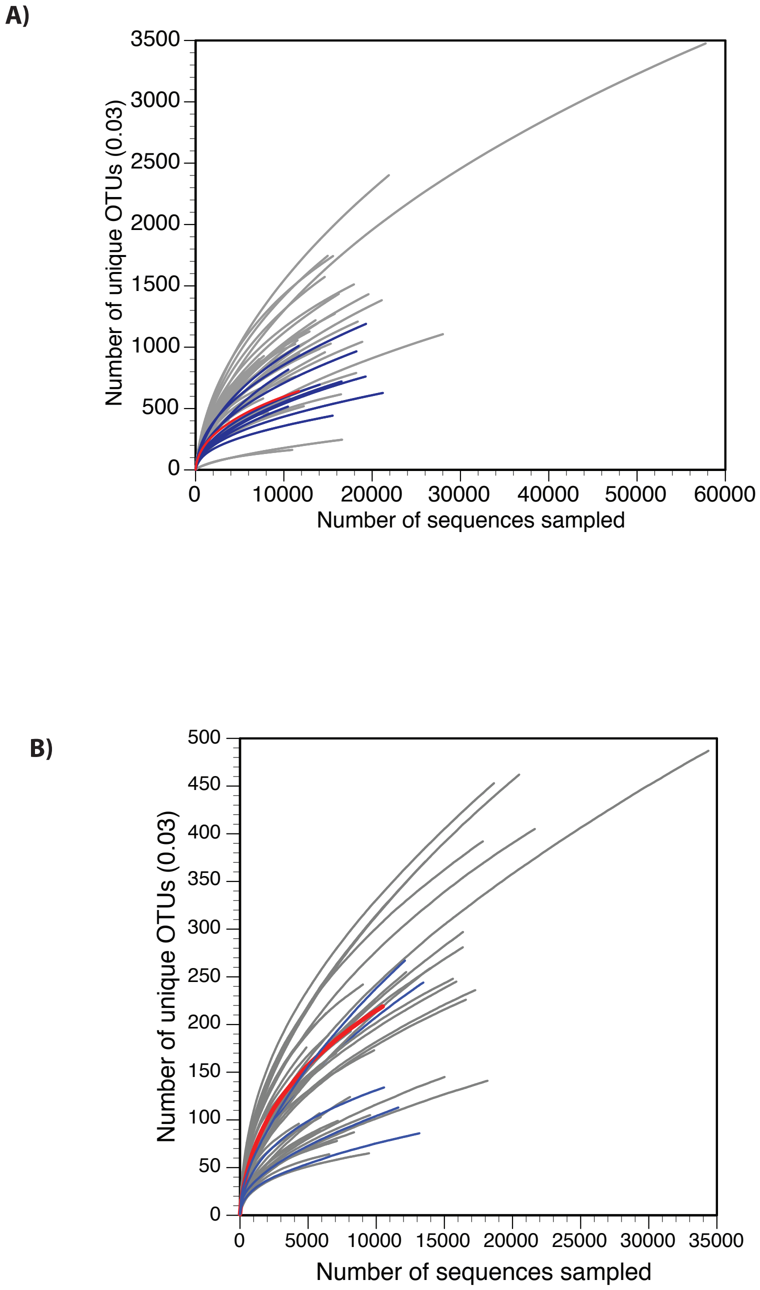
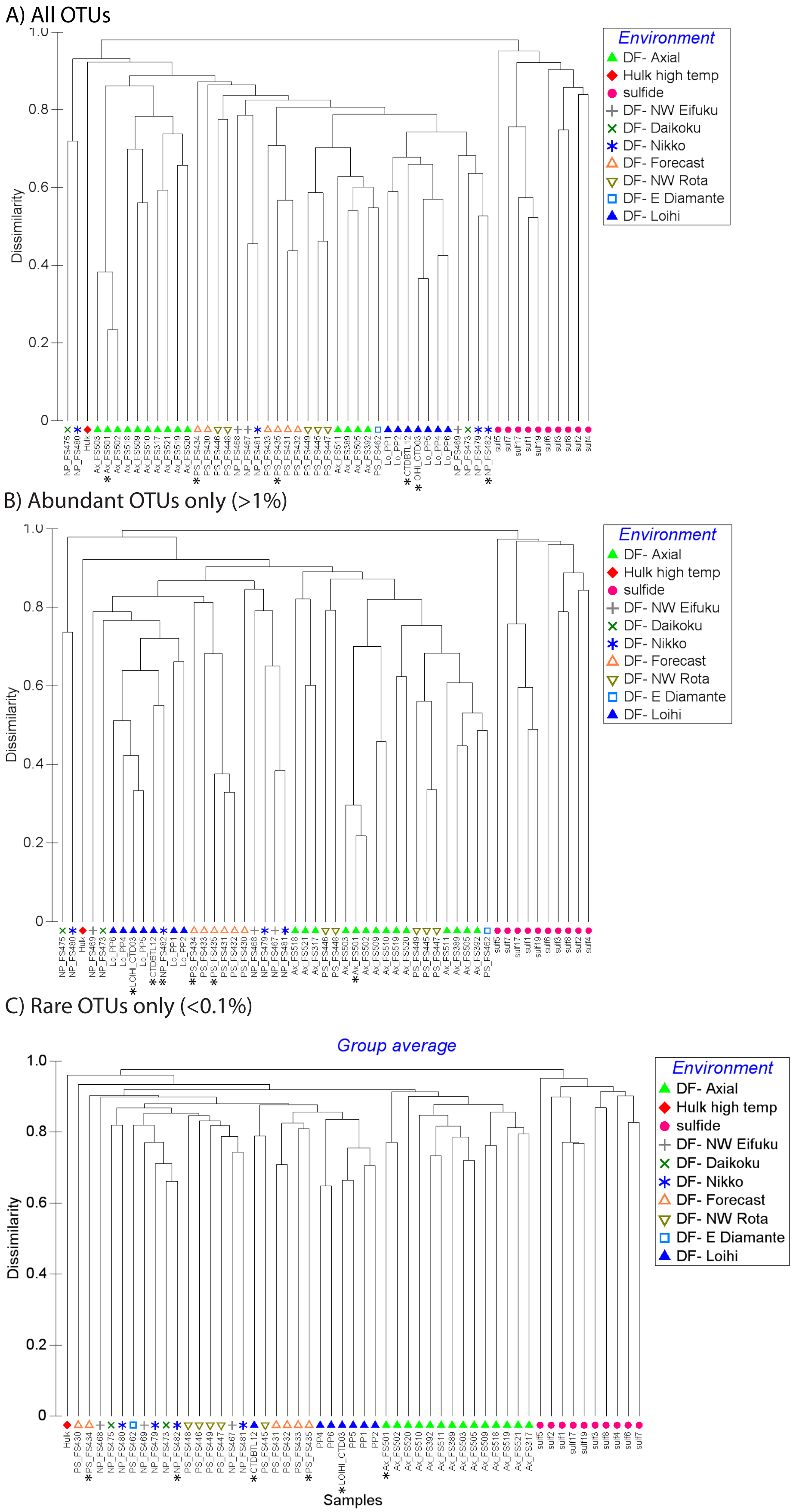


Figure 2.3. Rarefaction curves of bacterial (A) and archaeal (B) samples. Diffuse flow samples are noted in grey; sulfide samples, in blue; and Hulk sample, in red. Rarefaction curves were generated in mothur (Schloss *et al*., 2009).

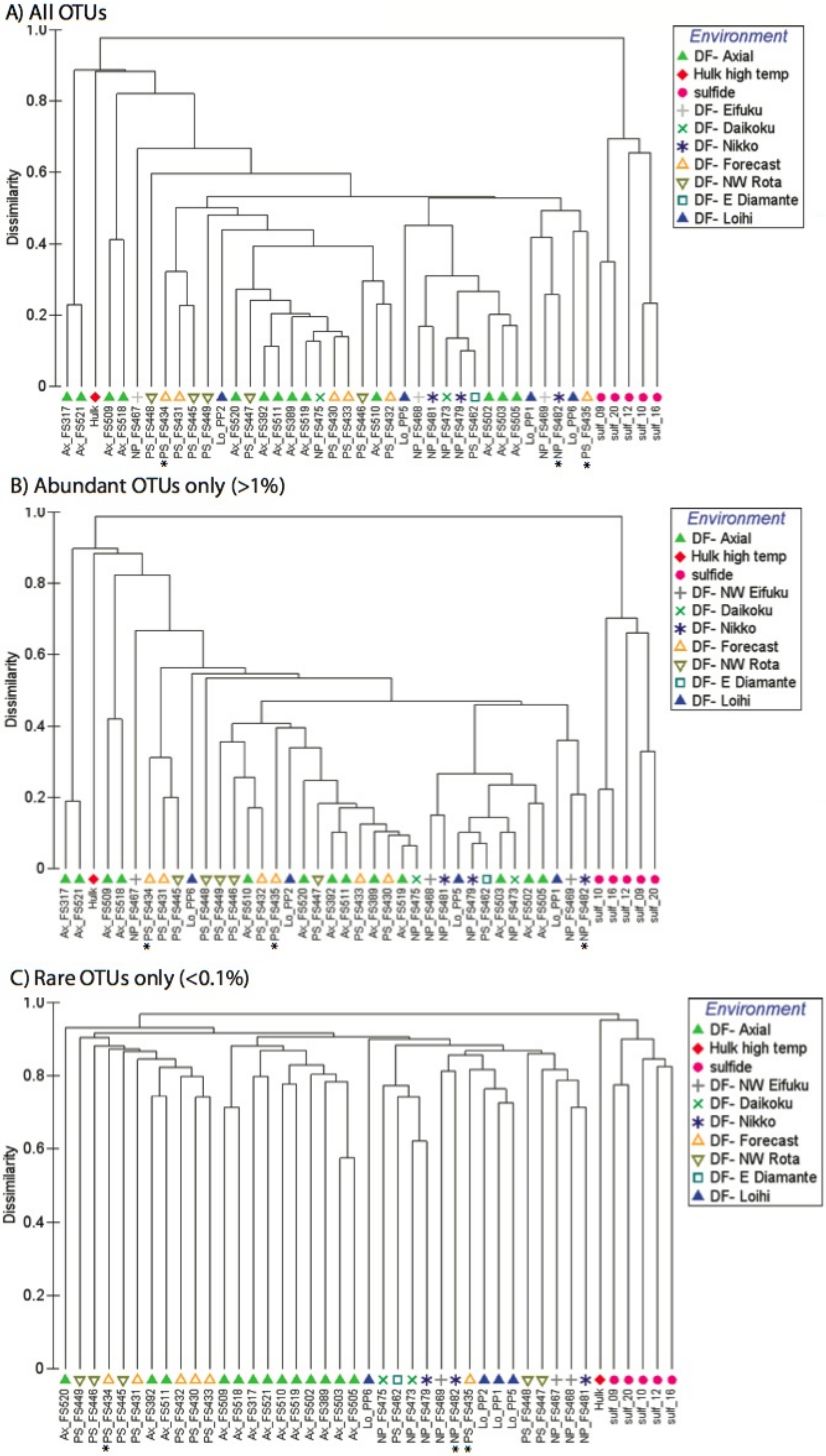


A) All OTUs

B) Abundant OTUs only (>1%)

C) Rare OTUs only (<0.1%)

Figure 2.4. (Previous page) Cluster dendrograms of diffuse flow and sulfide bacterial samples. Cluster dendrograms were created with group average method using the Bray-Curtis dissimilarity index. Operational taxonomic units are defined at the 3% distance for these analyses: A) analysis including all OTUs in each sample; B) analysis including only abundant OTUs (representing 1% or more of all sequences in each sample); and C) analysis including only rare OTUs (representing 0.1% or less of all sequences in each sample). Background samples are marked by asterisks. Samples are labeled according to fluid sample number, seamount, and region: NP = North Pacific, Ax = Axial Seamount, PS = Philippine Sea, Lo = Loihi Seamount.



A) All OTUs

B) Abundant OTUs only (>1%)

C) Rare OTUs only (<0.1%)

Figure 2.5. (Previous page) Cluster dendrograms of diffuse flow and sulfide archaeal samples. Cluster dendrograms were created with group average method using the Bray-Curtis dissimilarity index. Operational taxonomic units are defined at the 3% distance for these analyses: A) analysis including all OTUs in each sample; B) analysis including only abundant OTUs (representing 1% or more of all sequences in each sample); and C) analysis including only rare OTUs (representing 0.1% or less of all sequences in each sample). Background samples are marked by asterisks. Samples are labeled according to fluid sample number, seamount, and region: NP = North Pacific, Ax = Axial Seamount, PS = Philippine Sea, Lo = Loihi Seamount.

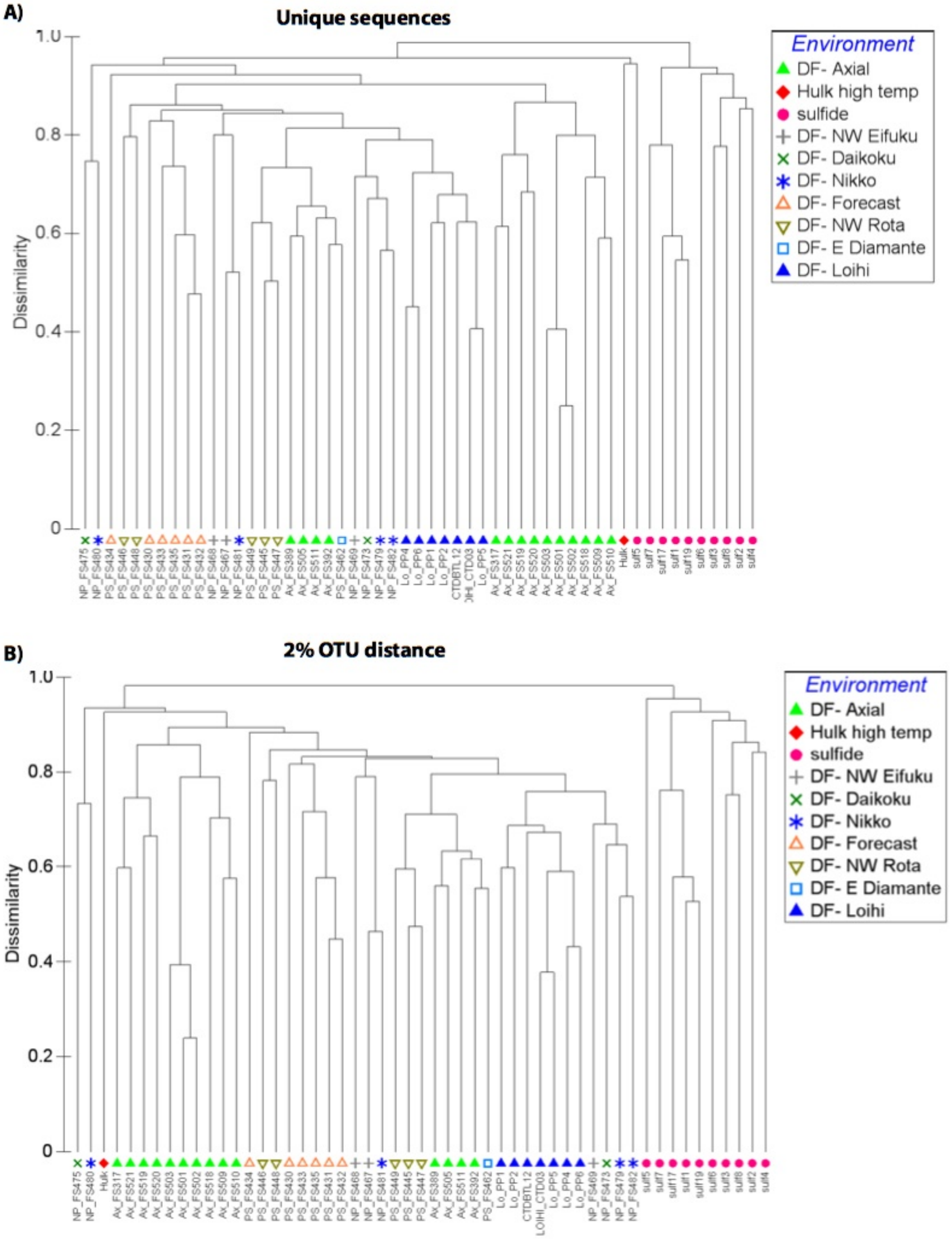


Figure 2.6. Cluster dendrograms of bacterial samples, clustered for unique sequences (A) and at the 2% OTU distance (B).

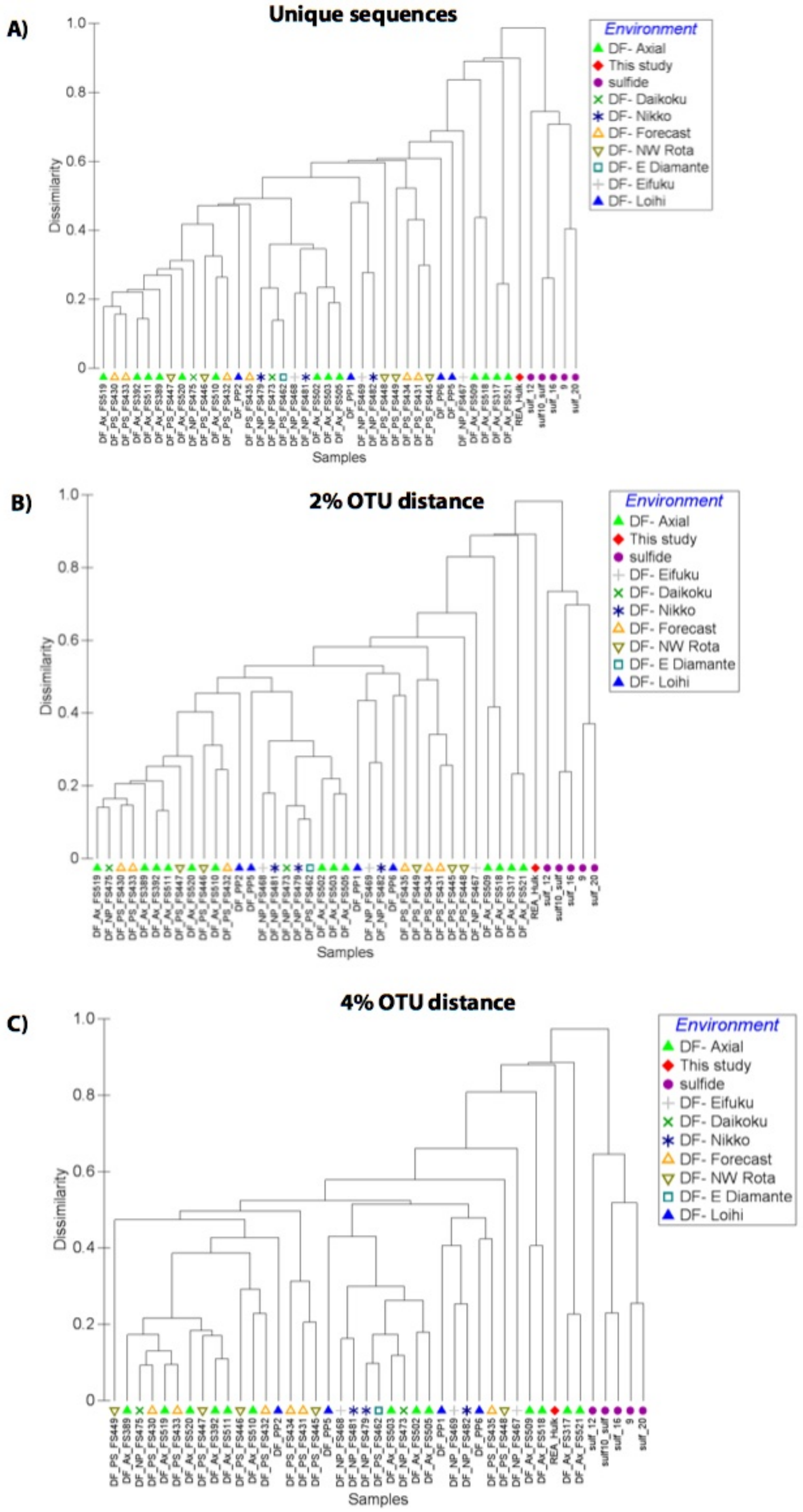


Figure 2.7. Cluster dendrograms of archaeal samples, clustered for unique sequences (A) and at the 2% (B) and 4% (C) OTU distances.

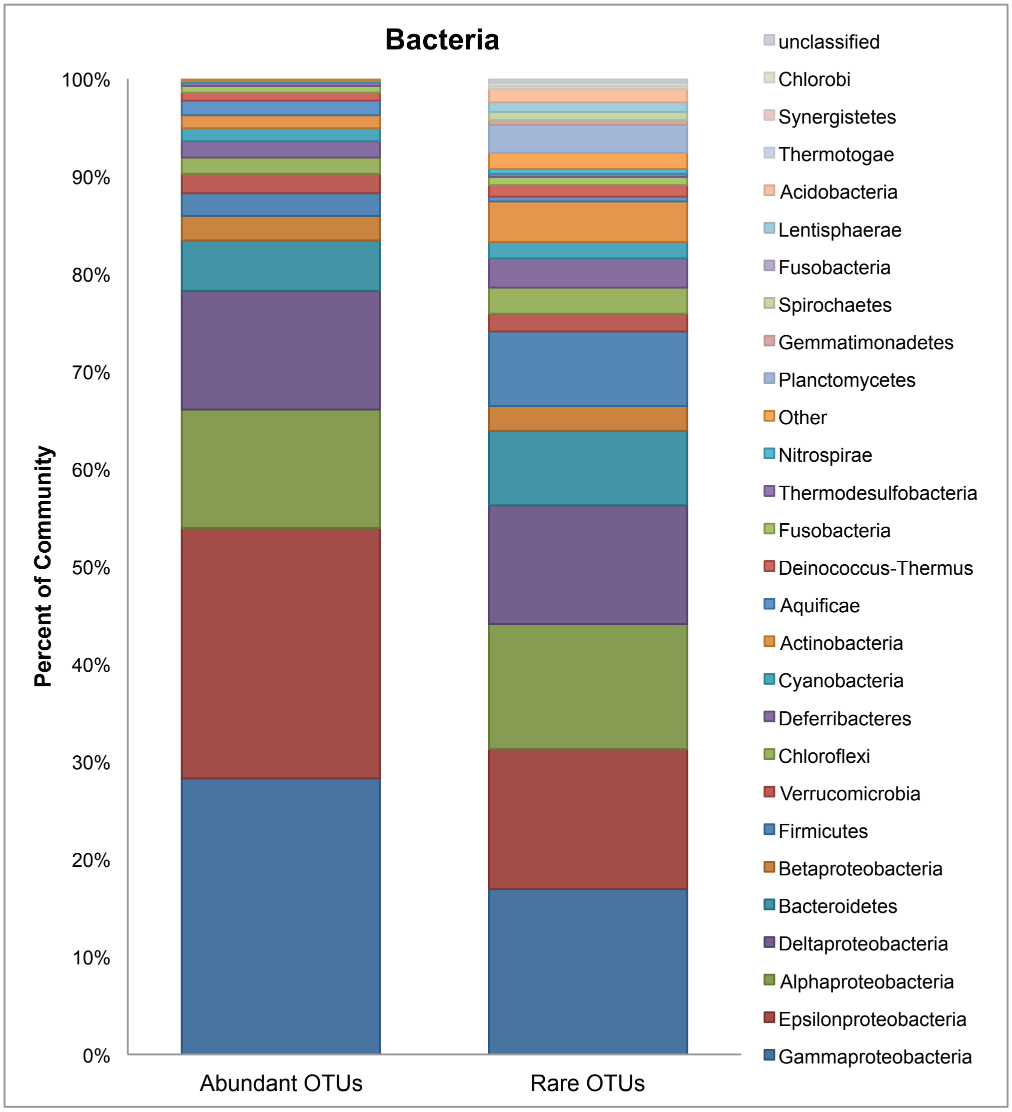


Figure 2.8. Bar charts of bacterial taxonomy for abundant and rare OTUs. Each category includes OTUs that were abundant or rare in at least one sample. There were 434 OTUs that were abundant in at least one sample, and 21733 OTUs that were rare in at least one sample. Taxonomy was assigned in mothur according to assignment to the SILVA database.

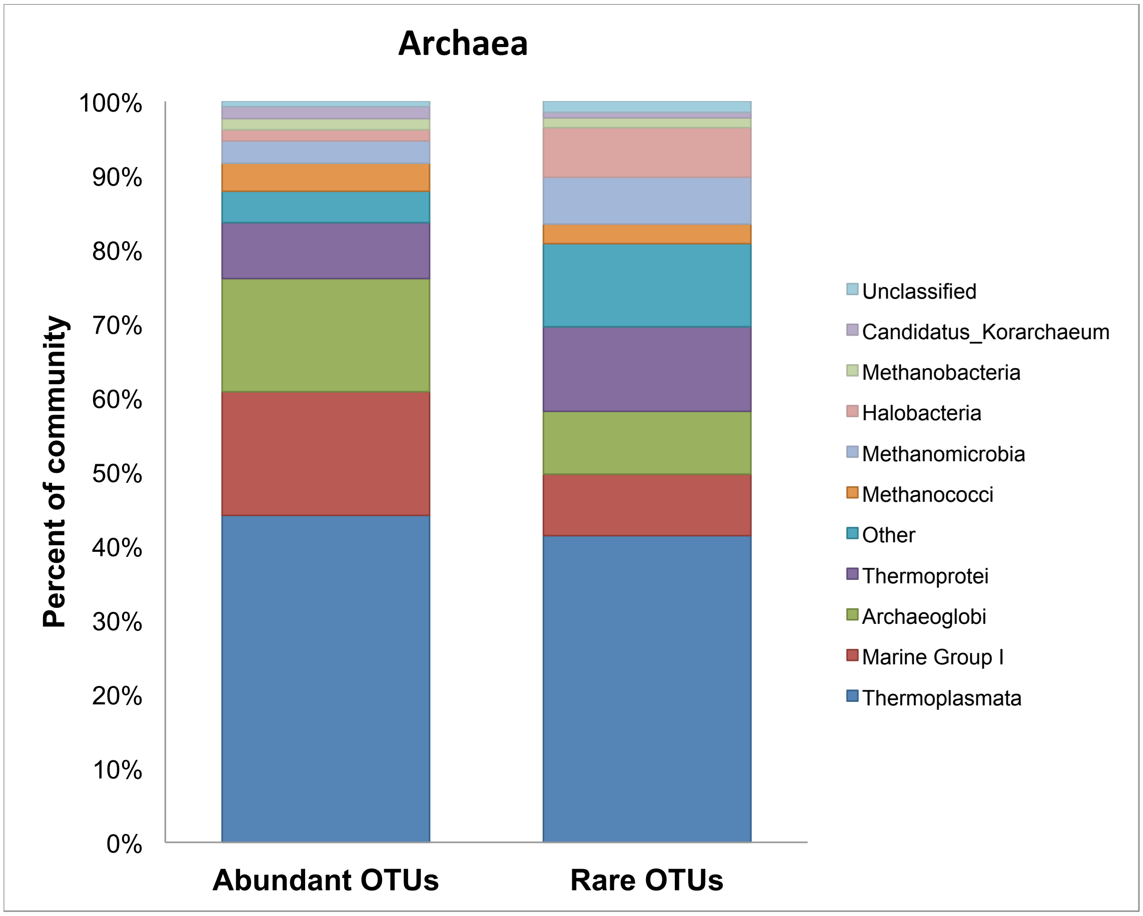


Figure 2.9. Bar charts of archaeal taxonomy for abundant and rare OTUs. Each category includes OTUs that were abundant or rare in at least one sample. There were 263 OTUs that were abundant in at least one sample, and 3435 OTUs that were rare in at least one sample. Taxonomy was assigned in mothur according to assignment to the SILVA database.

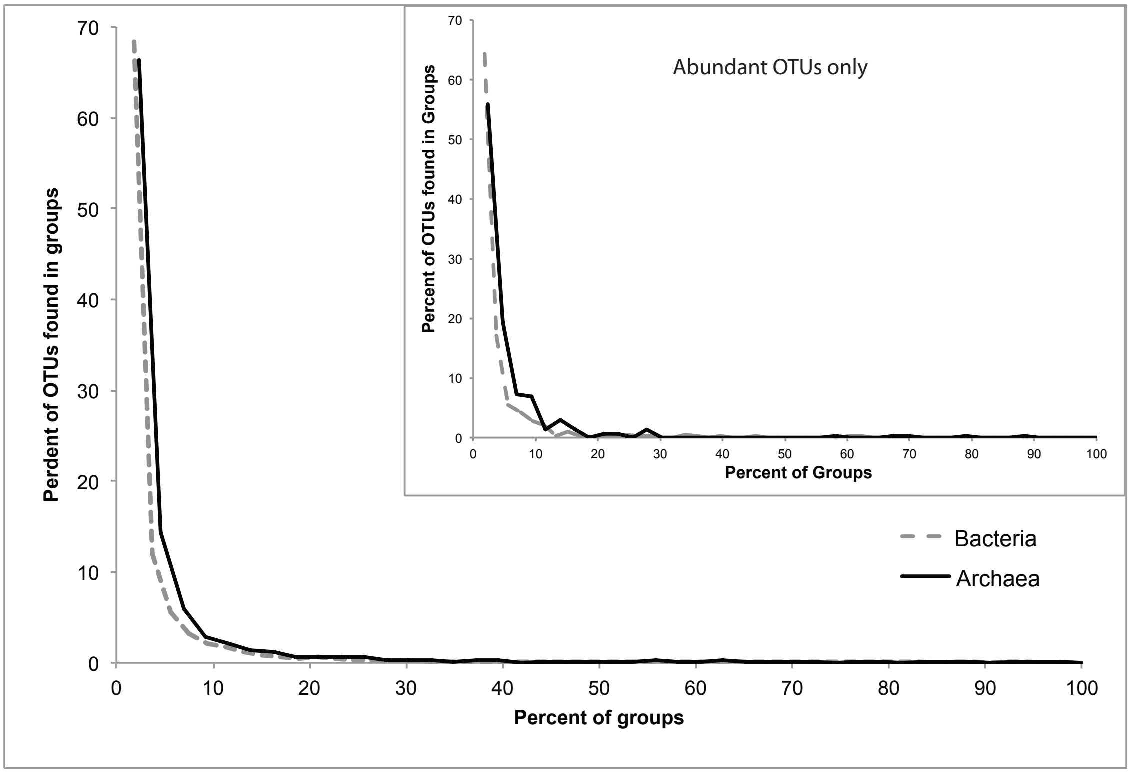


Figure 2.10. The percent of OTUs that were found in different proportions of samples within the dataset. Inset shows the same plot, but depicts only those OTUs that were classified as "abundant" in at least one sample.

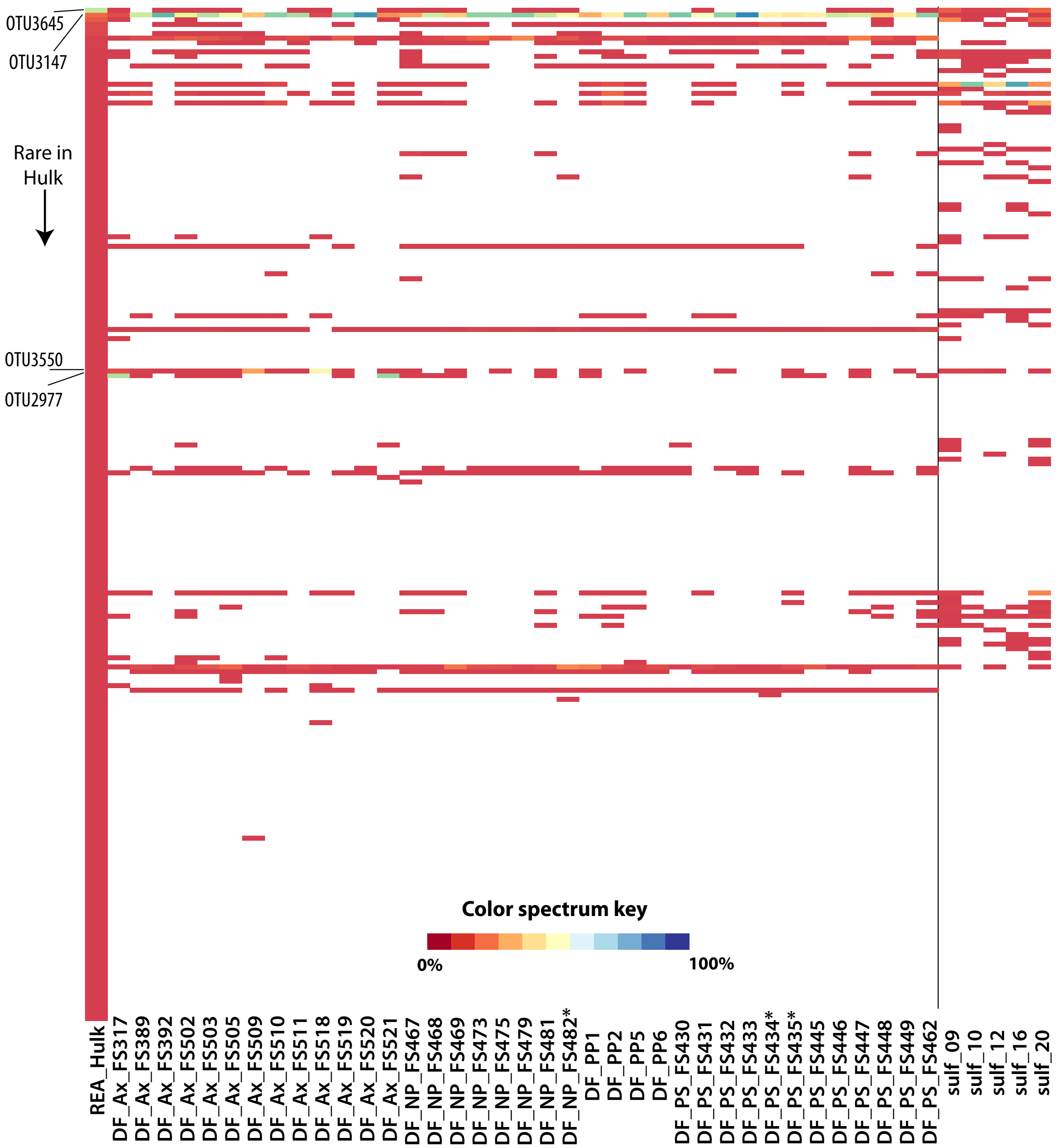

OTU3645
OTU3147
Rare in Hulk
OTU3550
OTU2977
Color spectrum key
0%
100%
REA_Hulk
DF_Ax_FS317
DF_Ax_FS389
DF_Ax_FS392
DF_Ax_FS502
DF_Ax_FS503
DF_Ax_FS505
DF_Ax_FS509
DF_Ax_FS510
DF_Ax_FS511
DF_Ax_FS518
DF_Ax_FS519
DF_Ax_FS520
DF_Ax_FS521
DF_NP_FS467
DF_NP_FS468
DF_NP_FS469
DF_NP_FS473
DF_NP_FS475
DF_NP_FS479
DF_NP_FS481
DF_NP_FS482*
DF_PP1
DF_PP2
DF_PP5
DF_PP6
DF_PS_FS430
DF_PS_FS431
DF_PS_FS432
DF_PS_FS433
DF_PS_FS434*
DF_PS_FS435*
DF_PS_FS445
DF_PS_FS446
DF_PS_FS447
DF_PS_FS448
DF_PS_FS449
DF_PS_FS462
sulf_09
sulf_10
sulf_12
sulf_16
sulf_20

Figure 2.11. (Previous page) Heatmap depicting the relative abundance of archaeal OTUs found in Hulk vent compared to other samples. OTUs are ordered according to their abundance in Hulk. OTUs falling roughly at or below the “Rare in Hulk” marker on the heatmap were present at 0.1% abundance or lower in the Hulk sample. White colors indicate that the OTU was not found in a given sample. Background samples are marked with asterisks.

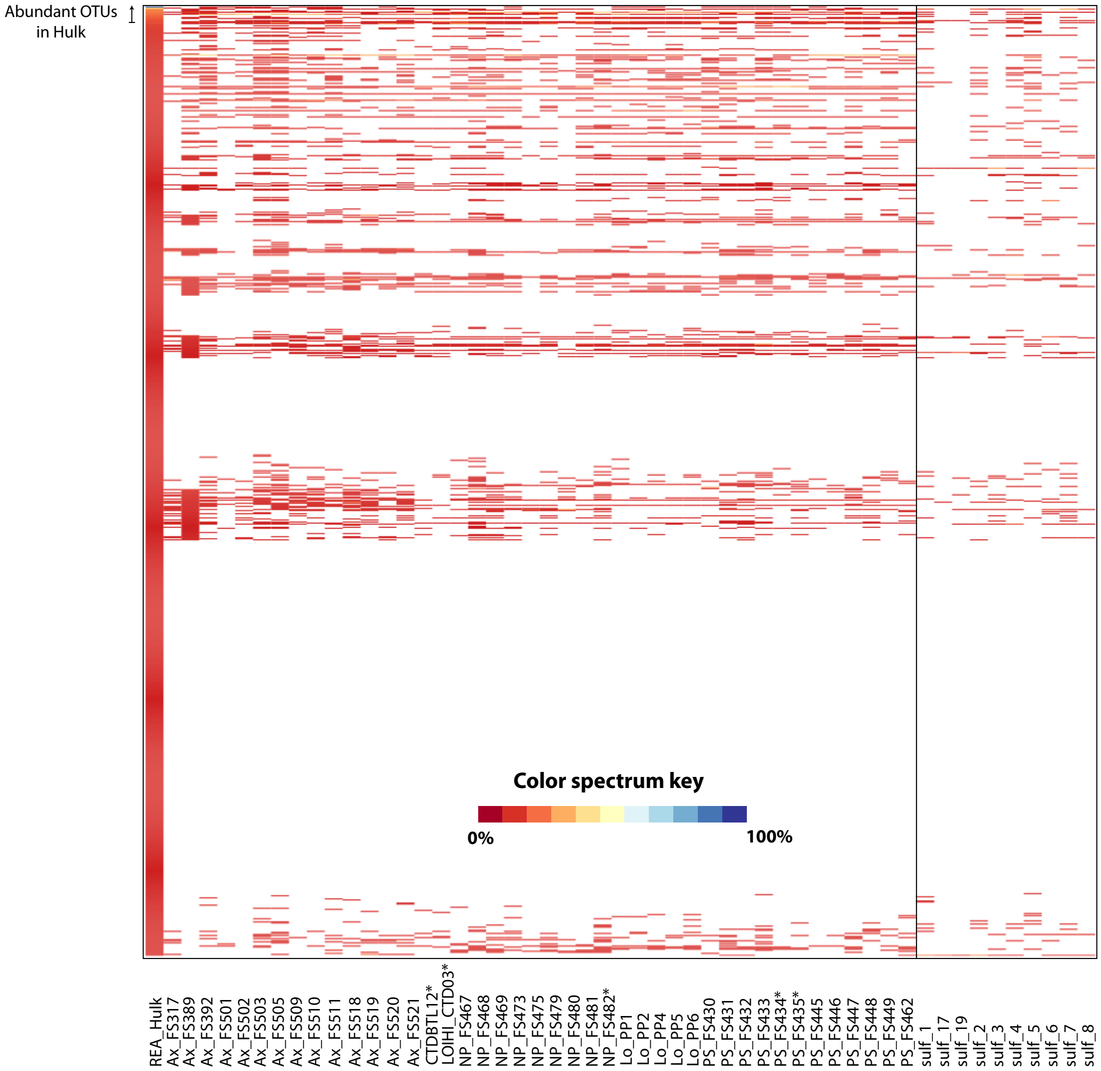


Figure 2.12. Heatmap depicting the relative abundance of bacterial OTUs found in Hulk vent compared to other samples. OTUs are ordered according to their abundance in Hulk. Abundant OTUs (those with a 1% abundance or higher) are indicated in brackets in the heatmap above. White colors indicate that OTU was not found in a given sample. Background samples are marked with asterisks.

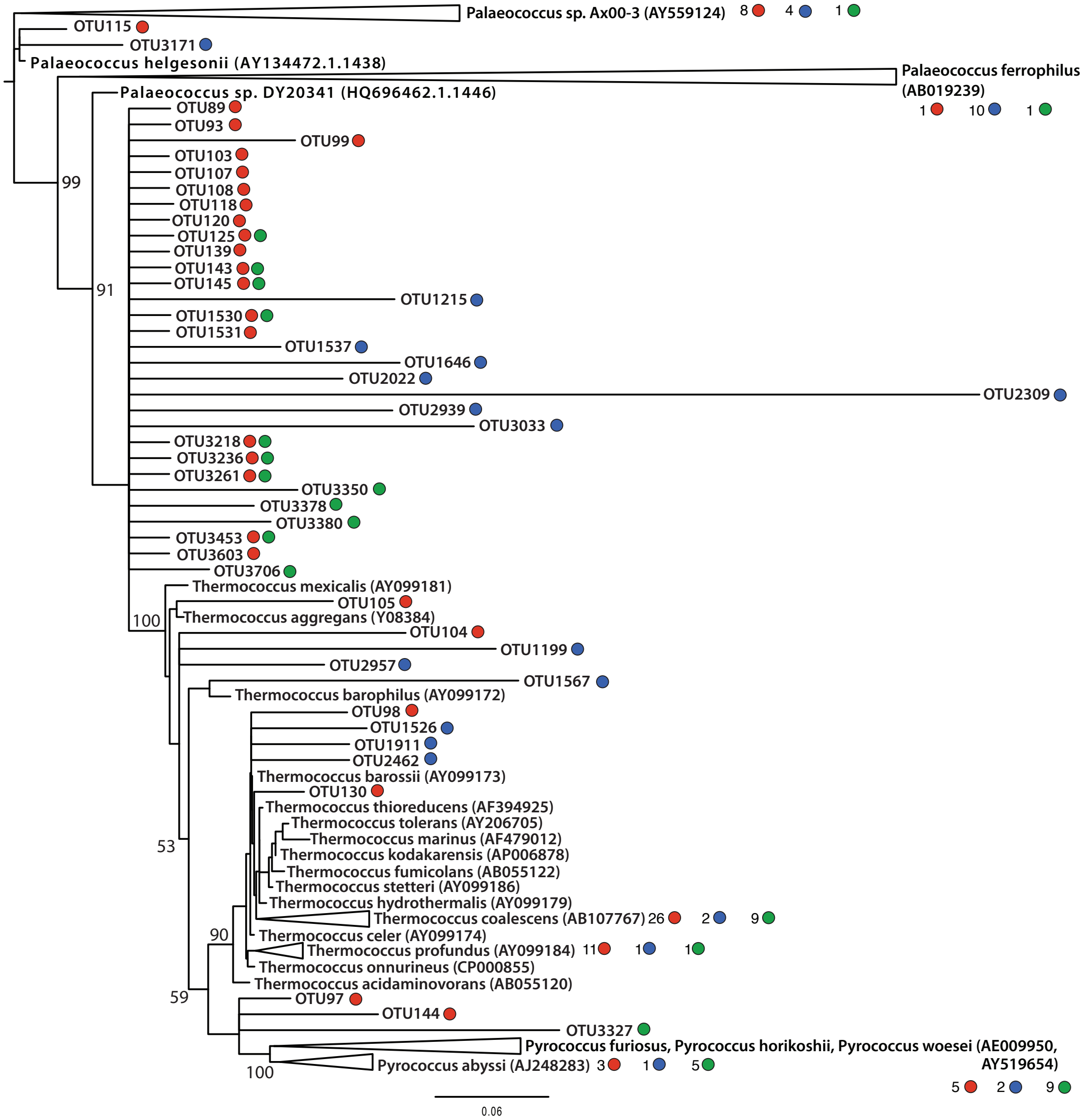

Palaeococcus sp. Ax00-3 (AY559124) 8 ● 4 ● 1 ●
OTU115 ●
OTU3171 ●
Palaeococcus helgesonii (AY134472.1.1438)
Palaeococcus ferrophilus (AB019239) 1 ● 10 ● 1 ●
Palaeococcus sp. DY20341 (HQ696462.1.1446)
OTU89 ●
OTU93 ●
OTU99 ●
OTU103 ●
OTU107 ●
OTU108 ●
OTU118 ●
OTU120 ●
OTU125 ● ●
OTU139 ●
OTU143 ● ●
OTU145 ● ●
OTU1215 ●
OTU1530 ● ●
OTU1531 ●
OTU1537 ●
OTU1646 ●
OTU2022 ●
OTU2309 ●
OTU2939 ●
OTU3033 ●
OTU3218 ● ●
OTU3236 ● ●
OTU3261 ● ●
OTU3350 ●
OTU3378 ●
OTU3380 ●
OTU3453 ● ●
OTU3603 ●
OTU3706 ●
99
91
Thermococcus mexicalis (AY099181)
OTU105 ●
100
Thermococcus aggregans (Y08384)
OTU104 ●
OTU1199 ●
OTU2957 ●
OTU1567 ●
Thermococcus barophilus (AY099172)
OTU98 ●
OTU1526 ●
OTU1911 ●
OTU2462 ●
Thermococcus barossii (AY099173)
OTU130 ●
Thermococcus thioreducens (AF394925)
Thermococcus tolerans (AY206705)
Thermococcus marinus (AF479012)
Thermococcus kodakarensis (AP006878)
Thermococcus fumicolans (AB055122)
Thermococcus stetteri (AY099186)
Thermococcus hydrothermalis (AY099179)
Thermococcus coalescens (AB107767) 26 ● 2 ● 9 ●
53
90
Thermococcus celer (AY099174)
Thermococcus profundus (AY099184) 11 ● 1 ● 1 ●
Thermococcus onnurineus (CP000855)
Thermococcus acidaminovorans (AB055120)
59
OTU97 ●
OTU144 ●
OTU3327 ●
Pyrococcus furiosus, Pyrococcus horikoshii, Pyrococcus woesei (AE009950, AY519654) 5 ● 2 ● 9 ●
100
Pyrococcus abyssi (AJ248283) 3 ● 1 ● 5 ●
0.06

Figure 2.13. Phylogenetic tree of *Thermococcales* based on 16S rRNA gene sequences, with pyrotag sequences added to the reference tree. Red dots indicate a sequence found in the high-temperature Hulk sample; blue dots, sequences found in diffuse flow samples; and green dots, sequences found in sulfide samples. Collapsed wedges are annotated with the number of sequences in each cluster that was found in each respective environment. Evolutionary history was inferred using a rapid bootstrap, maximum likelihood method with 100 alternative runs on distinct starting trees, using the GTR+ optimization of substitution rates and the GAMMA model of heterogeneity in RAxML (Stamatakis 2006). The Evolutionary Placement Algorithm (Berger *et al*., 2011) was used to insert short reads into the reference tree. Bootstrap values for the reference tree are labeled where they are over 50.

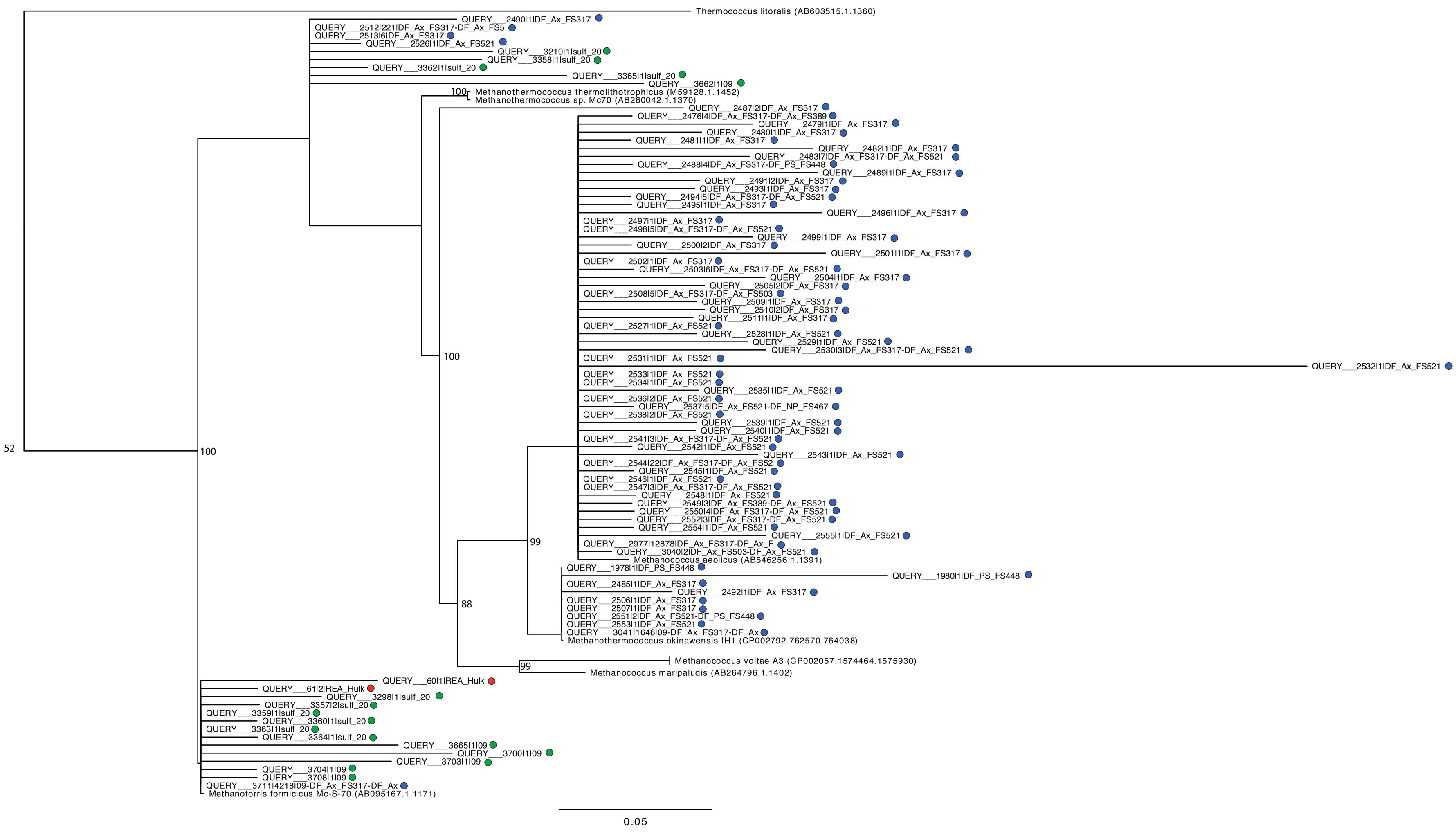

Thermococcus litoralis (AB603515.1.1360)
QUERY___2490|1|DF_Ax_FS317 ●
QUERY___2512|221|DF_Ax_FS317-DF_Ax_FS5 ●
QUERY___2513|6|DF_Ax_FS317 ●
QUERY___2526|1|DF_Ax_FS521 ●
QUERY___3210|1|sulf_20 ●
QUERY___3358|1|sulf_20 ●
QUERY___3362|1|sulf_20 ●
QUERY___3365|1|sulf_20 ●
QUERY___3662|1|09 ●
100
Methanothermococcus thermolithotrophicus (M59128.1.1452)
Methanothermococcus sp. Mc70 (AB260042.1.1370)
QUERY___2487|2|DF_Ax_FS317 ●
QUERY___2476|4|DF_Ax_FS317-DF_Ax_FS389 ●
QUERY___2479|1|DF_Ax_FS317 ●
QUERY___2480|1|DF_Ax_FS317 ●
QUERY___2481|1|DF_Ax_FS317 ●
QUERY___2482|1|DF_Ax_FS317 ●
QUERY___2483|7|DF_Ax_FS317-DF_Ax_FS521 ●
QUERY___2488|4|DF_Ax_FS317-DF_PS_FS448 ●
QUERY___2489|1|DF_Ax_FS317 ●
QUERY___2491|2|DF_Ax_FS317 ●
QUERY___2493|1|DF_Ax_FS317 ●
QUERY___2494|5|DF_Ax_FS317-DF_Ax_FS521 ●
QUERY___2495|1|DF_Ax_FS317 ●
QUERY___2496|1|DF_Ax_FS317 ●
QUERY___2497|1|DF_Ax_FS317 ●
QUERY___2498|5|DF_Ax_FS317-DF_Ax_FS521 ●
QUERY___2499|1|DF_Ax_FS317 ●
QUERY___2500|2|DF_Ax_FS317 ●
QUERY___2501|1|DF_Ax_FS317 ●
QUERY___2502|1|DF_Ax_FS317 ●
QUERY___2503|6|DF_Ax_FS317-DF_Ax_FS521 ●
QUERY___2504|1|DF_Ax_FS317 ●
QUERY___2505|2|DF_Ax_FS317 ●
QUERY___2508|5|DF_Ax_FS317-DF_Ax_FS503 ●
QUERY___2509|1|DF_Ax_FS317 ●
QUERY___2510|2|DF_Ax_FS317 ●
QUERY___2511|1|DF_Ax_FS317 ●
QUERY___2527|1|DF_Ax_FS521 ●
QUERY___2528|1|DF_Ax_FS521 ●
QUERY___2529|1|DF_Ax_FS521 ●
QUERY___2530|3|DF_Ax_FS317-DF_Ax_FS521 ●
QUERY___2531|1|DF_Ax_FS521 ●
100
QUERY___2532|1|DF_Ax_FS521 ●
QUERY___2533|1|DF_Ax_FS521 ●
QUERY___2534|1|DF_Ax_FS521 ●
QUERY___2535|1|DF_Ax_FS521 ●
QUERY___2536|2|DF_Ax_FS521 ●
QUERY___2537|5|DF_Ax_FS521-DF_NP_FS467 ●
QUERY___2538|2|DF_Ax_FS521 ●
QUERY___2539|1|DF_Ax_FS521 ●
QUERY___2540|1|DF_Ax_FS521 ●
QUERY___2541|3|DF_Ax_FS317-DF_Ax_FS521 ●
QUERY___2542|1|DF_Ax_FS521 ●
QUERY___2543|1|DF_Ax_FS521 ●
52
100
QUERY___2544|22|DF_Ax_FS317-DF_Ax_FS52 ●
QUERY___2545|1|DF_Ax_FS521 ●
QUERY___2546|1|DF_Ax_FS521 ●
QUERY___2547|3|DF_Ax_FS317-DF_Ax_FS521 ●
QUERY___2548|1|DF_Ax_FS521 ●
QUERY___2549|3|DF_Ax_FS389-DF_Ax_FS521 ●
QUERY___2550|4|DF_Ax_FS317-DF_Ax_FS521 ●
QUERY___2552|3|DF_Ax_FS317-DF_Ax_FS521 ●
QUERY___2554|1|DF_Ax_FS521 ●
QUERY___2555|1|DF_Ax_FS521 ●
99
QUERY___2977|12878|DF_Ax_FS317-DF_Ax_F ●
QUERY___3040|2|DF_Ax_FS503-DF_Ax_FS521 ●
Methanococcus aeolicus (AB546256.1.1391)
QUERY___1978|1|DF_PS_FS448 ●
QUERY___1980|1|DF_PS_FS448 ●
QUERY___2485|1|DF_Ax_FS317 ●
QUERY___2492|1|DF_Ax_FS317 ●
88
QUERY___2506|1|DF_Ax_FS317 ●
QUERY___2507|1|DF_Ax_FS317 ●
QUERY___2551|2|DF_Ax_FS521-DF_PS_FS448 ●
QUERY___2553|1|DF_Ax_FS521 ●
QUERY___3041|1646|09-DF_Ax_FS317-DF_Ax ●
Methanothermococcus okinawensis IH1 (CP002792.762570.764038)
99
Methanococcus voltae A3 (CP002057.1574464.1575930)
Methanococcus maripaludis (AB264796.1.1402)
QUERY___60|1|REA_Hulk ●
QUERY___61|2|REA_Hulk ●
QUERY___3298|1|sulf_20 ●
QUERY___3357|2|sulf_20 ●
QUERY___3359|1|sulf_20 ●
QUERY___3360|1|sulf_20 ●
QUERY___3363|1|sulf_20 ●
QUERY___3364|1|sulf_20 ●
QUERY___3665|1|09 ●
QUERY___3700|1|09 ●
QUERY___3703|1|09 ●
QUERY___3704|1|09 ●
QUERY___3708|1|09 ●
QUERY___3711|4218|09-DF_Ax_FS317-DF_Ax ●
Methanotorris formicicus Mc-S-70 (AB095167.1.1171)
0.05

Figure 2.14. (Previous page) Phylogenetic tree of *Methanococcales* based on 16S rRNA gene sequences, with pyrotag sequences added to the reference tree. Red dots indicate a sequence found in the high-temperature Hulk sample; blue dots, sequences found in diffuse flow samples; and green dots, sequences found in sulfide samples. Collapsed wedges are annotated with the number of sequences in each cluster that was found in each respective environment. Evolutionary history was inferred using a rapid bootstrap, maximum likelihood method with 100 alternative runs on distinct starting trees, using the GTR+ optimization of substitution rates and the GAMMA model of heterogeneity in RAxML (Stamatakis 2006). The Evolutionary Placement Algorithm (Berger *et al*., 2011) was used to insert short reads into the reference tree. Bootstrap values for the reference tree are labeled where they are over 50.

References

Allen, E.E., Tyson, G.W., Whitaker, R.J., Detter, J.C., Richardson, P.M., and Banfield, J.F. (2007) Genome dynamics in a natural archaeal population. Proc Natl Acad Sci U S A **104**: 1883–8.

Aller, J.Y. and Kemp, P.F. (2008) Are Archaea inherently less diverse than Bacteria in the same environments? FEMS Microbiol Ecol **65**: 74–87.

Anderson, R.E., Beltrán, M.T., Hallam, S.J., and Baross, J.A. (2013) Microbial community structure across fluid gradients in the Juan de Fuca Ridge hydrothermal system. FEMS Microbiol Ecol **83**: 324–339.

Anderson, R.E., Brazelton, W.J., and Baross, J.A. (2011) Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. FEMS Microbiol Ecol **77**: 120–133.

Baas Becking, L. (1934) Geobiologie of Inleiding Tot de Milieukunde [Geobiology or Introduction to the Science of the Environment] W.P. Van Stockum & Zoon, The Hague, Netherlands.

Beiko, R.G., Harlow, T.J., and Ragan, M. a (2005) Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U S A **102**: 14332–7.

Berger, S.A., Krompass, D., and Stamatakis, A. (2011) Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. Syst Biol **60**: 291–302.

Bouvier, T. and del Giorgio, P.A. (2007) Key role of selective viral-induced mortality in determining marine bacterial community composition. Environ Microbiol **9**: 287–97.

Brazelton, W.J., Ludwig, K.A., Sogin, M.L., Andreishcheva, E.N., Kelley, D.S., Shen, C.-C., *et al.* (2010) Archaea and bacteria with surprising microdiversity show shifts in dominance over 1,000-year time scales in hydrothermal chimneys. Proc Natl Acad Sci U S A **107**: 1612–1617.

Brazelton, W.J., Schrenk, M.O., Kelley, D.S., and Baross, J.A. (2006) Methane-and sulfur-metabolizing microbial communities dominate the Lost City hydrothermal field ecosystem. Appl Environ Microbiol **72**: 6257.

Campbell, B.J., Yu, L., Heidelberg, J.F., and Kirchman, D.L. (2011) Activity of abundant and rare bacteria in a coastal ocean. Proc Natl Acad Sci U S A **108**: 12776–12781.

Clarke, K.R. and Gorley, R.N. (2006) PRIMER v6: User Manual/tutorial. Primer-E Ltd Plymouth. text.

Deming, J. and Baross, J. (1993) Deep-sea smokers: Windows to a subsurface biosphere? Geochim Cosmochim Acta **57**: 3219–3230.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26**: 2460–1.

Embley, R.W., Baker, E.T., Chadwick, W.W., Lupton, J.E., Resing, J.A., Massoth, G.J., and Nakamura, K. (2004) Explorations of Mariana Arc volcanoes reveal new hydrothermal systems. Eos, Trans Am Geophys Union **85**: 37.

Flores, G.E., Campbell, J.H., Kirshtein, J.D., Meneghin, J., Podar, M., Steinberg, J.I., *et al.* (2011) Microbial community structure of hydrothermal deposits from geochemically different vent fields along the Mid-Atlantic Ridge. Environ Microbiol **13**: 2158–71.

Flores, G.E., Shakya, M., Meneghin, J., Yang, Z.K., Seewald, J.S., Wheat, G.C., *et al.* (2012) Inter-field variability in the microbial communities of hydrothermal vent deposits from a back-arc basin. Geobiology **10**: 333–46.

Frigaard, N.U., Martinez, A., Mincer, T.J., and DeLong, E.F. (2006) Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. Nature **439**: 847–850.

Fuhrman, J.A. (2009) Microbial community structure and its functional implications. Nature **459**: 193–9.

Galand, P.E., Casamayor, E.O., Kirchman, D.L., and Lovejoy, C. (2009) Ecology of the rare microbial biosphere of the Arctic Ocean. Proc Natl Acad Sci U S A **106**: 22427–32.

Gibbons, S.M., Caporaso, J.G., Pirrung, M., Field, D., Knight, R., and Gilbert, J.A. (2013) Evidence for a persistent microbial seed bank throughout the global ocean. Proc Natl Acad Sci U S A **110**: 4651–4655.

Huber, J.A., Butterfield, D.A., and Baross, J.A. (2003) Bacterial diversity in a subseafloor habitat following a deep-sea volcanic eruption. FEMS Microbiol Ecol **43**: 393–409.

Huber, J.A., Butterfield, D.A., and Baross, J.A. (2006) Diversity and distribution of subseafloor Thermococcales populations in diffuse hydrothermal vents at an active deep-sea volcano in the northeast Pacific Ocean. J Geophys Res **111**: G04016.

Huber, J.A., Butterfield, D.A., and Baross, J.A. (2002) Temporal changes in archaeal diversity and chemistry in a mid-ocean ridge subseafloor habitat. Appl Environ Microbiol **68**: 1585–1594.

Huber, J.A., Cantin, H. V, Huse, S.M., Welch, D.B.M., Sogin, M.L., and Butterfield, D.A. (2010) Isolated communities of Epsilonproteobacteria in hydrothermal vent fluids of the Mariana Arc seamounts. FEMS Microbiol Ecol **73**: 538–49.

Huber, J.A., Mark Welch, D.B., Morrison, H.G., Huse, S.M., Neal, P.R., Butterfield, D.A., and Sogin, M.L. (2007) Microbial population structures in the deep marine biosphere. Science (80-) **318**: 97–100.

Huber, J.A., Merkel, A., Holden, J.F., Lilley, M.D., and Butterfield, D.A. (2009) Molecular diversity and activity of methanogens in the subseafloor at deep-sea hydrothermal vents of the Pacific Ocean. AGU Fall Meet Abstr **-1**: 08.

Huse, S.M., Dethlefsen, L., Huber, J.A., Mark Welch, D., Welch, D.M., Relman, D.A., and Sogin, M.L. (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet **4**: e1000255.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., and Mark Welch, D.B. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol **8**: R143.

Jones, S.E. and Lennon, J.T. (2010) Dormancy contributes to the maintenance of microbial diversity. Proc Natl Acad Sci U S A **107**: 5881–6.

Kelley, D.S., Baross, J.A., and Delaney, J.R. (2002) Volcanoes, fluids, and life at mid-ocean ridge spreading centers. Annu Rev Earth Planet Sci **30**: 385–491.

Koonin, E. V., Makarova, K.S., and Aravind, L. (2001) Horizontal gene transfer in prokaryotes: Quantification and classification. Annu Rev Microbiol **55**: 709–742.

Kormas, K.A., Tivey, M.K., Von Damm, K., and Teske, A. (2006) Bacterial and archaeal phylotypes associated with distinct mineralogical layers of a white smoker spire from a deep-sea hydrothermal vent site (9˚N, East Pacific Rise). Environ Microbiol **8**: 909–20.

Lennon, J.T. and Jones, S.E. (2011) Microbial seed banks: the ecological and evolutionary implications of dormancy. Nat Rev Microbiol **9**: 119–30.

Moyer, C.L., Dobbs, F.C., and Karl, D.M. (1994) Estimation of diversity and community structure through restriction fragment length polymorphism distribution analysis of bacterial 16S rRNA genes from a microbial mat at an active, hydrothermal vent system, Loihi Seamount, Hawaii. Appl Envir Microbiol **60**: 871–879.

Nercessian, O., Reysenbach, A.-L., Prieur, D., and Jeanthon, C. (2003) Archaeal diversity associated with in situ samplers deployed on hydrothermal vents on the East Pacific Rise (13˚N). Environ Microbiol **5**: 492–502.

Opatkiewicz, A.D., Butterfield, D.A., and Baross, J.A. (2009) Individual hydrothermal vents at Axial Seamount harbor distinct subseafloor microbial communities. FEMS Microbiol Ecol **70**: 81–92.

Pedrós-Alió, C. (2006) Marine microbial diversity: can it be determined? Trends Microbiol **14**: 257–63.

Pruesse, E., Peplies, J., and Glöckner, F.O. (2012) SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics **28**: 1823–9.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res **41**: D590–6.

R Core, T. (2013) R: A Language and Environment for Statistical Computing.

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature **424**: 1042–7.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol **75**: 7537–41.

Schrenk, M.O., Kelley, D.S., Delaney, J.R., and Baross, J.A. (2003) Incidence and diversity of microorganisms within the walls of an active deep-sea sulfide chimney. Appl Environ Microbiol **69**: 3580–3592.

Slobodkin, A., Campbell, B., Cary, S.C., Bonch-Osmolovskaya, E., and Jeanthon, C. (2001) Evidence for the presence of thermophilic Fe(III)-reducing microorganisms in deep-sea hydrothermal vents at 13 degrees N (East Pacific Rise). FEMS Microbiol Ecol **36**: 235–243.

Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R., *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A **103**: 12115–20.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**: 2688–90.

Sul, W.J., Oliver, T.A., Ducklow, H.W., Amaral-Zettler, L.A., and Sogin, M.L. (2013) Marine bacteria exhibit a bipolar distribution. Proc Natl Acad Sci U S A **110**: 2342–7.

Summit, M. and Baross, J.A. (2001) A novel microbial habitat in the mid-ocean ridge subseafloor. Proc Natl Acad Sci U S A **98**: 2158–63.

Takai, K. and Horikoshi, K. (1999) Genetic diversity of archaea in deep-sea hydrothermal vent environments. Genetics **152**: 1285–1297.

Takai, K., Komatsu, T., Inagaki, F., and Horikoshi, K. (2001) Distribution of archaea in a black smoker chimney structure. Appl Environ Microbiol **67**: 3618–29.

Valentine, D.L. (2007) Adaptations to energy stress dictate the ecology and evolution of the Archaea. Nat Rev Microbiol **5**: 316–23.

Vergin, K.L., Beszteri, B., Monier, A., Cameron Thrash, J., Temperton, B., Treusch, A.H., *et al.* (2013) High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. ISME J **7**: 1322–1332.

CHAPTER THREE

Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage[3]

Summary

Metagenomic analyses of viruses have revealed widespread diversity in the viriosphere, but it remains a challenge to identify specific hosts for a viral assemblage. To address this problem, we analyze the viral metagenome of a northeast Pacific hydrothermal vent with a comprehensive database of spacers derived from the CRISPR (Clustered Regularly Interspaced Short Palindromic Repeat) putative immune system. CRISPR spacer matches to the marine vent virome suggest that viruses infecting hosts from diverse taxonomic groups are present in this vent environment. Comparative virome analyses show that CRISPR spacers from vent isolates and from thermophiles in general have a higher percentage of matches to the vent virome than to other marine or terrestrial hot spring viromes. However, a high percentage of hits to spacers from mesophilic hosts, combined with a moderately high modeled alpha diversity, suggest that the marine vent virome is comprised of viruses that have the potential to infect diverse taxonomic groups of multiple thermal regimes in both the bacterial and archaeal domains.

Introduction

Viruses play important ecological, biogeochemical and evolutionary roles throughout the world's ecosystems, particularly in the oceans (Suttle, 2005). Several viral metagenomes, or viromes, have been published from a wide range of marine and terrestrial environments (Breitbart *et al.*, 2002; Angly *et al.*, 2006; Bench *et al.*, 2007; Desnues *et al.*, 2008; Schoenfeld *et al.*, 2008; Williamson *et al.*, 2008; Lopez-Bueno *et al.*, 2009; Santos *et al.*, 2010). These reports have demonstrated the mobility of viral genes between environments as well as the tremendous diversity of genes encoded by the global viriosphere (Breitbart and Rohwer, 2005; Dinsdale *et al.*, 2008; Kristensen *et al.*, 2009).

[3] Previously published in FEMS Microbiology Ecology, Vol. 77, p. 120-133, 2011.

These metagenomic analyses face several challenges, however. First, in most viromes published to date, the vast majority of reads have no match to existing databases, while the majority of the rest have matches to bacterial or archaeal genes (*see for ex.* Angly *et al.*, 2006). This high percentage of unknown sequences renders further identification of viral types or viral genes more challenging. Moreover, the isolation of viral particles independently of their hosts, as is done for most viromes, makes it difficult to identify which hosts are targeted by the viral assemblage. Yet identification of viral hosts is crucial to understanding the role of viruses in an ecosystem. Identifying viral hosts would aid in determining how viruses impact the microbial diversity of a given ecosystem, for example, or which microbes may be sharing genes through virally mediated horizontal gene transfer. To address this problem, we have analyzed a viral metagenome from a diffuse flow hydrothermal vent with a comprehensive database created from the CRISPR (Clustered Regularly Interspaced Palindromic Repeat) immune system.

The CRISPR system is a putative antiviral immunity mechanism found in both archaea and bacteria (Barrangou *et al.*, 2007; Brouns *et al.*, 2008; Sorek *et al.*, 2008; van der Oost, J. *et al.*, 2009; Horvath *et al.*, 2010; Labrie *et al.*, 2010; Marraffini and Sontheimer, 2010). CRISPR loci generally consist of a series of short repeats, each approximately 20-50 bp in length, interspersed by spacers of about 25-75 bp in length (Grissa *et al.*, 2007b). CRISPR loci are thought to create immunity when short sequences derived from invaders such as viruses or plasmids are incorporated as spacers between the repeat sequences by genes involved in the CRISPR response, known as CRISPR-associated (CAS) genes. When introduced genetic elements, such as viruses or plasmids, have a 100% match to a preexisting CRISPR spacer sequence in the host genome, these elements are recognized as pathogenic invaders (Makarova *et al.*, 2003; Bolotin *et al.*, 2005; Haft *et al.*, 2005; Mojica *et al.*, 2005; Pourcel *et al.*, 2005; Marraffini and Sontheimer, 2010; Hale *et al.*, 2009). In response, the CRISPR/Cas system cleaves the invading nucleic acid (Garneau *et al.*, 2010).

CRISPR loci effectively act as libraries of previous viral infection; thus analyses of CRISPR spacers across species have great potential for microbial and viral ecology. Previous studies have examined CRISPRs in an ecological context, focusing on variability and distribution in acid mine drainage biofilms (Andersson and Banfield, 2008; Tyson and Banfield, 2007) as well as in terrestrial hot springs (Heidelberg *et al.*, 2009; Held and Whitaker, 2009; Held *et al.*, 2010). These studies have found a high degree of variability within CRISPR spacer sequences, implying a rapid rate of host-virus co-evolution. These studies have also demonstrated a clear biogeographic distribution in CRISPR spacers. Additionally, Snyder *et al.* (2010) have designed microarrays using CRISPR spacer sequences to detect viruses in environmental samples.

CRISPR spacers provide a means to analyze and compare viral sequences for which we have some host genomic context (*i.e.* the complete genomes of isolates), whereas metagenomics provides information about the genetic content of a viral assemblage at a particular location and point in time. Here, by comparing a database of CRISPR spacers from all published archaeal and bacterial genomes with reads from a viral metagenome, we are able to infer what types of hosts might be infected by the viruses in the viral assemblage, even if their sequences have no close BLAST matches in available databases.

The environment we have chosen as our focus for this analysis is a diffuse flow hydrothermal vent system in the Main Endeavour Field in the northeast Pacific Ocean. As in other marine environments, the virus to cell ratio is approximately ten to one in vents at the Main Endeavour Field (Ortmann and Suttle, 2005), yet induction experiments have shown that vent communities of the East Pacific Rise display a higher incidence of lysogeny than other marine environments (Williamson *et al.*, 2008). While it is evident that viruses play a prominent role in the vent environment, until now the diversity, structure and genomic content of vent viral communities have not been assessed.

The dynamic, gradient-dominated nature of the vent environment makes it a particularly attractive site for studies of viral ecology and evolution. In these environments, ambient seawater mixes with high-temperature hydrothermal fluid enriched in reduced compounds, creating gradients in pH, temperature, chemical composition, and mineralogy both above and below the seafloor (Baross and Hoffman, 1985). These gradients set up a series of microenvironments, providing niches for diverse communities of microorganisms (Huber *et al.*, 2003; Schrenk *et al.*, 2003). Continuous circulation of the hydrothermal fluid both above and below the seafloor enables potentially frequent contact among these microbial communities and their accompanying viral assemblages. In such an environment, viruses of a diverse array of hosts could also potentially come into frequent contact with each other. As viruses are known vectors of horizontal gene exchange, the presence of a wide diversity of viruses and their hosts could facilitate widespread gene transfer. This analysis, with an emphasis on identifying potential viral hosts, provides a new perspective on a viral assemblage whose unique signature mirrors the dynamic yet extreme environment it inhabits.

Materials and Methods

Diffuse flow hydrothermal fluid virus sampling

170 L of hydrothermal vent fluid were collected with a barrel sampler from diffuse flow at the base of Hulk vent in the Main Endeavour Field on the Juan de Fuca ridge (approximately 450 km west of Washington state in the Pacific Ocean). The sample funnel was placed atop a clump of tubeworms on a sulfide structure venting diffuse flow hydrothermal vent fluid (Figure 3.1). Chemical and physical metadata from Hulk vent are summarized in Table 3.1. The minimum temperature of the sample was 13˚C, as measured through a temperature probe on a hydrothermal fluid sampler (HFS) at the sample site. However, the average chemistry-derived temperature of the barrel sample, calculated based on dissolved silica content, was much higher. Measured silica content of hydrothermal fluid from a 300˚C black smoker about 10 meters away from the sample site was 15,199 µM, whereas seawater silica content was 185 µM at 1.8˚C. From this, our sample temperature was calculated to be approximately 125˚C. While this is much higher than the diffuse flow temperatures recorded by the HFS, it is possible that this is because diffuse flow measured by the intake nozzle of the HFS retained much higher amounts of seawater than that taken in by the intake funnel of the barrel sampler, which may have had a better seal on the sulfide structure and therefore pulled in higher temperature fluid.

Upon recovery, several 20 ml fluid subsamples were collected for cell and virus counts. Samples were fixed with 10% formaldehyde and stored at 4˚C for two weeks until counted. Cell and viral counts were conducted by filtering 1 ml of a 1/10 diluted sample onto a 25mm 0.02 µm Anodisc filter (Whatman Inc., Kent, United Kingdom) backed by a GF/F nitrocellulose filter at <20 kpa pressure. Filters were placed on a drop of 1-5X SYBR Gold and allowed to sit for 15 minutes prior to mounting on slides with filtered PBS/glycerol/ascorbate solution. At least 200 cells and viruses were counted in a minimum of 20 fields of view.

For pyrosequencing, the sample was filtered with a 0.22 µm Steripak filter unit (Millipore, Massachusetts, United States) on ice to remove cells. The filtrate was concentrated through tangential flow filtration (30-kD cutoff) to approximately 400 mL (Thurber *et al.*, 2009) in a 4˚C cold room. Samples were stored in 50-mL fractions and frozen at –80˚C. Upon thawing, 10% wt vol^{-1} PEG 8000 was added to one 50-mL fraction and incubated at 4˚C overnight. Each sample was pelleted by centrifugation at 104,400 RPM for 50 min., resuspended in TE and incubated for 15 min. with 0.7 volume of chloroform to lyse any remaining cellular contamination. After centrifugation for 10 min. at 4˚C to remove chloroform, the aqueous fraction was incubated with 10% DNAse I for two hours at 37˚C to eliminate any free DNA in solution. DNAse was inactivated by adding EDTA to a final concentration of 0.02M. Viral DNA was extracted with the QIAAmp MinElute Virus Spin Kit (Qiagen Inc., California, United States). Samples were sent to the Broad Institute for 454 Titanium pyrosequencing (454 Life Sciences, Branford, Connecticut, United States).

Bioinformatics

Phylogenetic assignments of reads in the marine vent virome were carried out through the MG-RAST pipeline (Meyer *et al.*, 2008). Reads were compared to the SEED database with tblastx with a maximum e-value cutoff of 10^{-5}. Reads with a significant match to a viral sequence according to these parameters were categorized into families as defined by the International Commission on Taxonomy of Viruses (ICTV) 2009 release of Virus Taxonomy (http://www.ictvonline.org/virusTaxonomy.asp?bhcp=1). Marine vent virome contigs were assembled and analyzed with Geneious (Drummond *et al.*, 2009) (www.geneious.com). Contigs were assembled with the "Medium Sensitivity" method with a word length of 14, maximum gap size of 2, maximum gaps per read of 15, and maximum mismatches of 15. Contig taxonomy for each read was defined according to the consensus taxonomy as defined by the taxonomy of the majority of reads within each contig. Read taxonomy was assigned through the MG-RAST pipeline by comparing to the SEED database, with a maximum e-value cutoff of 10^{-5}.

Modeling uncultured viral assemblage diversity

The alpha diversity of each virome was estimated using the Phage Communities from Contig Spectrum (PHACCS) online tool (http://biome.sdsu.edu/phaccs), described in previous publications (Angly *et al.*, 2005, 2006). Briefly, 10,000 random sequences were assembled using Minimo (98% identity over at least 35bp overlap). Circonspect (Angly *et al.*, 2006) (http://sourceforge.net/projects/circonspect/) was used to calculate a contig spectrum by calculating the number of contigs of each size, using a minimum metagenome coverage of 2, minimum dinucleotide entropy of 2.0, low complexity filter window length of 21, and with varying trim and discard sizes depending on the average read length of the metagenome. The average viral genome length was estimated using GAAS (Angly *et al.*, 2009) through a CAMERA 2.0 alpha diversity workflow (http://calit.camera2.net).

CRISPR spacer analyses

All genomes analyzed in this study (1083 archaea and bacteria) were downloaded from the NCBI ftp server on 20 April 2010. CRISPRs were identified in each of these genomes with the CRISPR Recognition Tool (CRT) (Bland *et al.*, 2007) using default parameters, and the number of CRISPR loci and total CRISPR spacers per genome were tabulated. CRISPR spacers were compiled into a single database and categorized by genome. To compare the spacers in the database to each other, we performed a blastn comparison of the set of spacers within an individual genome against the set of spacers in each of the other genomes and then compiled these results into a resemblance matrix. From this we determined what proportion of all CRISPR spacers between the two genomes was shared. All spacer similarities were calculated with blastn (Altschul *et al.*, 1990). Spacer ”matches” were defined as matches of 100% identity along at least 20 base pairs of the spacer sequence. To calculate the percentage of reads with a match to a spacer, only unique queries (reads) were counted. For the analysis in which we identified the taxonomy of potential hosts, we included matches of multiple reads to the same spacer, as well as multiple spacers to the same read.

For analysis of CRISPR spacer matches from each temperature regime, each genome in the NCBI database was sorted according to thermal regime as defined by a genome properties list downloaded from the NCBI ftp server. “Vent isolates” were characterized as all strains, both thermophilic and mesophilic, that had been isolated from either a shallow or deep-sea hydrothermal vent. These are listed in Table 3.2.

To compare average growth temperature with CRISPR abundance, archaea and bacteria were grouped as thermophiles (optimal growth temperature of 60˚C or above; includes hyperthermophiles) or mesophiles (optimal growth temperature between 25˚C and 60˚C), and some bacteria were designated as psychrophiles (optimal growth temperature below 25˚C).

Results and Discussion

The structure of our analysis focused first on ensuring virome quality through contig analysis. BLAST analyses were conducted on virome reads to gain an overall picture of both the structure and content of the marine vent viral assemblage, and to determine which viral families were present. Next, we modeled the richness and evenness of the viral assemblage and compared this with previously sequenced marine and hot springs viromes. Finally, to provide host context for these results, we queried the marine vent virome with a comprehensive CRISPR spacer database to identify potential microbial hosts of viruses in the vent viral assemblage.

Matches of marine vent virome reads to known sequences

Of 228,698 reads, the majority (67.14%) of reads in the marine vent virome yielded no matches to the SEED database (e-value cutoff 10^{-5}) (Figure 3.2). This viral metagenome has a smaller percentage of unknown reads than found in some previous marine viral metagenomes (Table 3.3). However, this may be an artifact of read length: this metagenome, sequenced with 454 Titanium technology, had an average read length of 334 bp, whereas marine viral metagenomes sequenced with 454 FLX technology averaged approximately 100 bp (Angly *et al.*, 2006). Longer reads are more likely to have significant matches to existing databases. Similar percentages of unknown reads have been found in viral metagenomes with longer read lengths (Bench *et al.*, 2007; Schoenfeld *et al.*, 2008), though this is not true for all cases (Lopez-Bueno *et al.*, 2009). It is possible that contamination with cellular sequences may contribute to the relatively how percentage of unknown sequences; however, we believe that a significant portion of the metagenome was viral. This is discussed in greater detail below.

Of the reads in the marine vent virome with a significant database match, 25.87% matched bacterial sequences, 4.40% matched archaeal sequences, and only 0.69% matched known viral sequences. Similar proportions have been found in previously sequenced viromes. In general, the abundance of bacterial and archaeal matches may be explained by the larger number of archaeal and bacterial sequences in the database and possibly also by a high rate of horizontal gene transfer between viruses and their hosts, resulting in the presence of microbial genes in viral genomes and vice versa (Angly *et al.*, 2006).

We next examined the presence of specific viral families based on reads with matches to known viral sequences. The results, shown in Figure 3.3, suggest that the viral assemblage at marine vents is more similar to other marine viral assemblages than to those in terrestrial hot springs. As only DNA was sequenced, this analysis would necessarily miss RNA viruses or retroviruses, but the presence of DNA viruses among different biomes can be compared. The majority of viral reads in the marine vent virome belonged to the *Myoviridae* family, as is the case with many other marine viromes (Figure 3.3). Other tailed viruses common to marine viromes, the *Podoviridae* and *Siphoviridae*, were also relatively common in the marine vent virome. Recent studies have shown that ssDNA viruses such as *Microviridae* predominate in temperate marine waters such as the Sargasso Sea and the Bay of British Columbia (Angly *et al.*, 2006), yet sequences matching the *Microviridae* family were largely absent from the marine vent virome. However, unlike other viromes, our sample was not amplified with Phi29 polymerase, which is biased toward amplification of ssDNA viruses, and may explain the relative lack of ssDNA viruses in this virome (Kim *et al.*, 2008).

Viruses known to infect archaea such as the *Rudiviridae*, *Fueselloviridae*, and *Lipothrixviridae*, commonly found in hot springs viral assemblages (Prangishvili *et al.*, 2006; Schoenfeld *et al.*, 2008), were largely absent from the marine vent virome. The abundance of archaea in marine vents would suggest that it is unlikely that archaeal viruses are absent from the marine vent assemblage, and therefore this implies that archaeal viruses present in the marine vent assemblage were unlike any sequenced strains found in terrestrial hot springs. Therefore, marine vent systems may play host to novel archaeal viruses not yet discovered.

In total, eleven different virus families were found in the marine vent assemblage, which is higher than any of the other viromes compared in this analysis, with the exception of the Arctic Ocean (Figure 3.3). This supports the notion that a wide range of viral types is present in the marine vent viral assemblage.

Marine vent virome assembly

Assembly of the marine vent virome yielded several large contigs. Figure 3 shows the mean coverage and length of each contig. Many of the longest contigs in the vent virome contained reads matching bacterial genes (Figure 3.4A). However, the longest contigs in the virome had relatively low mean coverage**.** Reads with the highest mean coverage tended to be slightly shorter and contained reads with no matches to the SEED database. One possible explanation for this pattern is that shorter reads with higher coverage were derived from viral genomes, whereas the longer reads with low mean coverage were derived from bacterial or archaeal genomes. Contigs with high coverage tended to contain reads with no matches to existing databases (Figure 3.4A). A BLAST search of the contig with the highest coverage revealed hits only to short segments at each end of the contig, most of which corresponded to DNA ligases, further supporting the notion that these high-coverage contigs were viral.

Additionally, sequenced genomes from a range of viral types have been found to contain sequences with high similarity to archaeal or bacterial genes *(see for ex*. Mann *et al.*, 2003; Filée *et al.*, 2007; Geslin *et al.*, 2007; Fischer *et al.*, 2010), and therefore some contigs that were assigned to bacterial or archaeal taxa may actually lie within viral genomes.

Modeling richness and evenness of the viral assemblage

We modeled the alpha diversity of the marine vent virome and compared it with the diversity of six previously sequenced viromes: Bear Paw and Octopus Spring from Yellowstone National Park (Schoenfeld *et al.*, 2008), for a high temperature comparison, and four marine viromes: the Sargasso Sea, the Gulf of Mexico, the Bay of British Columbia, and the Arctic Ocean (Angly *et al.*, 2006), for a marine comparison. We re-modeled the alpha diversity of each virome to maintain consistent parameters in the Circonspect and PHACCS models to enable direct comparison. The modeled diversity values thus differ from original published results due to changes in both the Circonspect and PHACCS software (Angly, personal communication). In each of the metagenomes sequenced with 454 technology (resulting in over 100,000 reads with reads ranging from 100-300 bp), the trim size and discard size were set to 100, and the sample size in Circonspect was set to 10,000 reads. For the metagenomes sequenced with Sanger technology (resulting in only 8000-22,000 reads of about 1000 bp long), the trim size and discard size were set to 650 due to longer read lengths, but the 10,000 read sample size was only possible for one of the metagenomes. Therefore, comparing richness across viromes sequenced by different technologies must be done cautiously, as the different read lengths and number of reads alter the output values. We sought to minimize error in the analysis while retaining reasonable read lengths and sample sizes, given the sequencing technology.

Our results (Table 3.4) indicate that the richness of the marine vent virome is comparable to that of other high temperature or marine environments. Our results also show that the evenness of the marine vent virome is higher than that of any other virome, and thus the viral assemblage is not dominated by any single genotype. We also modeled the alpha diversity of the marine vent virome after removing all reads contained within contigs longer than 3000 bp in order to test whether the presence of long, low-coverage contigs (possibly derived from archaea and bacteria) influenced the results. The results, labeled (b) in Table 3.4, were not significantly altered.

Using the CRISPR spacer database to identify potential hosts

The previous analyses have not provided specific information about what types of hosts are infected by the viral assemblage in marine hydrothermal vents. To address this, we created a database of the CRISPR spacers contained within all sequenced organisms in the NCBI database, consisting of 81,260 spacers from 1083 genomes. As each CRISPR spacer is thought to be derived from a viral (or plasmid) sequence, this database serves as a repository of sequences from viruses that have infected these organisms. Moreover, since each of these spacer sequences is derived from the genome of a particular organism, we can match the viral sequence to the host. A similar CRISPR spacer search was conducted by (Garrett *et al.*, 2010) to query hyperthermophilic viral enrichments; however, rather than targeting specific hosts, this CRISPR spacer database was designed to identify potential hosts for the viruses represented by our assembled metagenomic sequences.

For our initial analysis, we conducted a blastn search between the CRISPR spacer database we generated and the marine vent virome, searching for 98% identity across the entire spacer sequence. Zero matches were found with these parameters, which attests to the diversity of viral sequences and the speed at which they mutate.

However, phage genomes are known to be mosaic in nature (Hendrix *et al.*, 2000; Hendrix, 2003), and it is thought that viruses can evade the CRISPR system by scrambling their sequences through the process of recombination (Andersson and Banfield, 2008). Thus we searched for 100% alignments of CRISPR spacers across a portion of the spacer sequence rather than the full sequence, choosing as our cutoff 100% alignment across at least 20 base pairs. This 20 base pair cutoff was chosen to be lenient enough to find matches that are significant, but stringent enough to preclude false matches to CRISPR spacers.

Control dataset: comparing spacers within the database

To test the significance of these parameters, spacers from all of the sequenced bacteria and archaea in our CRISPR spacer database were compared to each other with blastn, searching for matches of 100% identity across 20 bp. The set of CRISPR spacers within each of the 578 CRISPR-containing archaeal and bacterial genomes was compared to the set of CRISPR spacers in each of the other 578 genomes. Of the 166,753 unique genome comparisons, 262 had one or more matching spacers at 100% identity across 20 bp. Of these, 249 (95%) were between spacers from genomes of the same genus, and of these, 155 (63%) were spacer matches between spacers from genomes of the same species. This provides strong evidence that a sequence with a match to a CRISPR spacer at this level (100% identity across 20 bp) is most likely derived from a virus that infected a host of the same genus or species as the CRISPR spacer it matches.

Querying the marine vent virome with the CRISPR spacer database

When comparing the CRISPR spacer database to the marine vent virome, a total of 290 different spacers out of 81,260 spacers in the database had a match to the marine vent virome at 100% identity across 20 base pairs. 382 different reads out of 228,698 (0.167%) in the marine vent virome contained a match to one of these CRISPR spacers. While these reads represent a low percentage of the total, the conservative parameters were retained to minimize the possibility of false matches. At this stringency level, there is a $(0.25)^4$, or 9.09 $x10^{-13}$, chance that a random sequence would match, and therefore out of 228,698 reads, one would expect 2.08×10^{-7} reads to have a match. Thus, the result of 382 different virome reads with a match to a spacer cannot be due to random sequence similarity.

To compare this result with that of cellular metagenomes, we conducted the same BLAST search of the CRISPR spacer database against several other cellular metagenomes taken from the MG-RAST database. These metagenomes represent a range of GC content, read length, and number of reads. Results are shown in Table 3.5 in terms of the ratio of spacer matches to base pairs to normalize for differences in read length and number. The results show that the average ratio of matches to base pairs is 4.027×10^{-6} for the cellular metagenomes, whereas it is 6.27×10^{-6} for the marine vent virome. This suggests that there was a higher proportion of spacer hits to this virome than to these cellular metagenomes, despite the presence of CRISPR loci and possible viral contamination in the cellular metagenomes. Numbers of CRISPR-associated (*cas*) genes identified in each metagenome are also listed in this table. While our results indicate that the number of *cas* genes identified in a given metagenome can vary widely, these results do indicate that CRISPRs were present in the cellular metagenomes and may have contributed to the total number of spacer matches. While some *cas* genes were found in the marine vent virome as well, the reads matching these *cas* genes fell on only five contigs consisting of 3 or more reads; of these, each of these *cas*-gene-containing contigs had relatively low coverage (maximum 3.2).

To further test for the presence of contaminating bacterial or archaeal reads that may have contained CRISPR loci, we searched for evidence of CRISPR direct repeats in the vent virome to act as a proxy for CRISPR loci derived from cellular genomes. CRISPR direct repeat sequences, unlike spacer sequences, do not correspond to viral sequences and are much more highly conserved among loci and among taxa (Kunin *et al.*, 2007). The marine vent virome contained only 58 reads with a match to a CRISPR repeat sequence, compared to 382 reads with a match to a CRISPR spacer. A "match" is here again defined as 100% identity over 20 base pairs. If the vent virome had a high proportion of contaminating CRISPR loci from bacterial or archaeal genomes, we would have expected a relatively higher number of matches to CRISPR direct repeats.

Figure 3.4B shows which contigs contained a read with a match at this level to a CRISPR spacer. While some were found within reads assigned to bacteria, 76% of the contigs with matches to CRISPR spacers contained a majority of reads with no match to the SEED database. Additionally, nearly half of the reads with matches to the spacer database belonged to these unidentified contigs, which contain only about one-third of the total virome reads. This supports the notion that the reads with matches to the CRISPR spacer database represent viral sequences.

Identification of potential hosts for the marine vent viral assemblage

To identify potential hosts for the marine vent viral assemblage, we grouped all of the spacers matching the virome according to the taxonomic group of the strain from which they were derived. Table 3.6 depicts the distribution of BLAST hits between the CRISPR spacers of each group and the marine vent virome. Most notable about the results is the wide range of both archaeal and bacterial taxonomic groups that had CRISPR spacers matching the marine vent virome, with no single taxonomic group dominating. This suggests that the viruses in the vent assemblage have the potential to infect a wide range of taxonomic groups. The groups with the most matches between their CRISPR spacers and the marine vent virome were the *Firmicutes*, the *Bacteroidetes/Chlorobi*, and the *Gammaproteobacteria*; however, the high number of hits from these groups may be attributed partially to the high number of CRISPR spacers from these groups in the database. Interestingly, a relatively small percentage of spacers from the *Proteobacteria* (particularly α-, β-, γ-, and δ*-proteobacteria*) had matches to the marine vent virome, despite the large number of spacers from *Proteobacteria* in the spacer database, and despite the prevalence of these taxa at this site (Huber *et al.*, 2007). Therefore, this result may reflect a surprising lack of viruses infecting *Proteobacteria* in our sample. Matches to archaeal CRISPR spacers are common within the marine vent virome (Table 3.6), despite the relative absence of known archaeal virus families in the metagenome (Figure 3.3). However, known archaeal virus families have predominantly been cultured from terrestrial hot springs. Since the archaeal domain is known to be well-represented in marine hydrothermal fluids (Huber *et al.*, 2007), these data suggest that the archaeal viruses present in the marine vent virome are unlike those found in terrestrial hot springs and were therefore undetectable with traditional BLAST searches, but may have been detected by our CRISPR spacer analysis.

To more closely examine species known to be endemic to marine hydrothermal vent ecosystems, we determined the relative numbers of matches between the marine vent virome and the CRISPR spacers from genomes of vent isolates. While these spacer matches do not necessarily indicate that viruses infecting these specific species have been identified, we can state that sequences similar to those that have infected these species in the past are present in this virome. It is also interesting to note that the species listed in this table were isolated from a wide range of different vent environments; therefore, these results may instead give some indication of the similarity of viruses across different vent types, given that spacer sequences unique to distinct hosts have been identified in the same virome. Moreover, it is interesting to note that the results (Table 3.7) show that two-thirds of the vent isolate spacer hits were from *Methanocaldococcus* species, despite the fact that *Methanocaldococcus* strains only comprise about 15% of sequenced vent isolates.

CRISPR spacers in Methanocaldococcus genomes

The high abundance of CRISPR spacer matches from *Methanocaldococcus* species can be attributed in part to the high number of CRISPR spacers in individual

Methanocaldococcus genomes. For example, the genome of *Methanocaldococcus* sp. FS406-22, a hyperthermophilic methanogen that fixes nitrogen at 92˚C (Mehta and Baross, 2006), has the highest number of CRISPR loci of all sequenced isolates to date: 23 were identified with the CRISPRFinder application (Grissa *et al.*, 2007b, 2007a), and 20 were identified with the CRISPR Recognition Tool (CRT) (Bland *et al.*, 2007). *M. vulcanius* M7 and *M. jannaschii* DSM 2661, also isolated from marine hydrothermal vents, contain the second and third highest numbers of CRISPR loci of all sequenced isolates, respectively. As described above, no spacers were shared among genomes. Interestingly, nearly all spacers (94–99%) were also unique within each thermophilic methanogen genome, even in those containing high numbers of CRISPR loci. In other words, almost none of these genomes contained a duplicate CRISPR spacer. It seems unlikely that typical recombination and mutation events could cause this level of diversity in the spacer sequences but not in the CRISPR repeats, all of which are identical or nearly identical within each CRISPR locus. Instead, it is likely that these methanogens gained their multitude of CRISPR spacers through distinct infection events. It is not clear whether spacer diversity correlates with infecting viral diversity, however, because of the apparent semi-random nature of the CRISPR mechanism. Proto-spacer adjacent motifs (PAMs) are thought to act as recognition sequences for CRISPR genes. Most PAMs are 2-3 nucleotides long, resulting in a large number of potential spacer sites on a viral genome (Mojica *et al.*, 2009). Therefore, the lack of duplicate sequences indicates a large number of distinct infection events, but it is unclear whether it also implies high diversity of infecting viruses. While the reasons for the abundance of CRISPR loci in thermophilic methanogens are unknown, it is part of a larger trend in thermophiles that is discussed further below.

Nevertheless, while the abundance of CRISPRs in *Methanocaldococcus* genomes is striking, it does not fully explain the large percentage of matches between the marine vent virome and *Methanocaldococcus* spacers. CRISPR spacers from *Methanocaldococcus* species represent 28% of the CRISPR spacers from vent isolates (Table 3.7), yet they represented over 50% of the vent isolate spacer matches to the marine vent virome. This suggests that viruses of *Methanocaldococcus* species were particularly prevalent in this diffuse flow sample.

CRISPR spacers as a probe of host thermal regime

We next performed a blastn search of our CRISPR spacer database with five other previously published viromes, with a particular emphasis on thermal regime. We compared the CRISPR spacer database with four marine viromes and two Yellowstone hot springs viromes (combined together for this analysis). Only a single match with 100% similarity over the full length of the spacer sequence was found: a spacer from *Synechococcus* sp. JA-2-3B'a(2-13), isolated from Octopus Spring in Yellowstone, had a match to the virome from the same site. No other perfect matches between the CRISPR spacer database and any of these viromes were found.

We next searched for 100% alignments of 20 bp or above, as before. A total of 901 out of 81,260 spacers, or 1.11% of the spacers in the CRISPR database, had a match to one or more of the six viromes. We grouped these hits according to the thermal regime of the host from which the spacer was derived (Figure 3.5**)**. Our results show that 1.84% of spacers specific to vent isolates had a match to the marine vent virome. This constitutes a higher percentage than to viromes in other environments, suggesting that there is a unique vent virus "signature," perhaps due to a particular sequence or set of sequences that is shared among vent viruses. Notably, the vent isolates included in this analysis were isolated from marine hydrothermal vents around the globe, indicating that this vent "signature" is not unique to a particular marine vent location or depth.

Thermophilic strains also had a high percentage of spacer matches to the vent virome (0.98%) relative to other viromes. Several *Sulfolobus* spacers, for example, had matches to the marine vent virome despite being endemic to terrestrial hot springs. Again, this is an interesting contrast to the relative lack of marine vent virome read matches to archaeal virus families found in terrestrial hot springs (Figure 3.3). It is possible that the *Sulfolobus* spacers with matches to the marine vent virome are derived from viruses that have not been isolated or sequenced, and therefore had no matches in existing databases. As a natural "library" of viral infection, the CRISPR spacer dataset does not rely upon isolation of individual virus-host systems and is therefore able to identify potential hosts for viruses in the assemblage with no cultured relatives.

Finally, the marine vent virome had a relatively high proportion of matches to spacers from non-vent and non-thermophilic organisms (Fig. 4). This result highlights the multitude of microenvironments present in marine diffuse flow hydrothermal systems. Because gradients in temperature, pH, chemical composition, and mineralogy are known to dominate vent systems (Baross and Hoffman, 1985), our vent fluid sample likely was a composite of fluids that experienced a variety of environmental conditions in the subsurface. These fluids may have experienced temperatures ranging from that of ambient seawater, at around 2˚C, up to 135˚C or possibly even higher. The pH could have ranged from that of ambient seawater, between pH 7-8, to much lower pH values typical of high temperature hydrothermal fluids, at around pH 2 or 3. Therefore, it is not surprising that the diverse microbial communities inhabiting vents play host to diverse viral communities as well.

Correlation of CRISPR locus abundance per genome and growth temperature

Our analyses indicated that CRISPR spacers from thermophiles are common in all viromes, which may be attributed to the abundance of CRISPR spacers from thermophiles in our database. Closer examination of this trend shows that thermophilic strains, on average, have higher numbers of CRISPR loci in their genomes than mesophiles. Early literature on CRISPR loci made brief note of this trend (Makarova *et al.*, 2003, 2006), but it has not yet been given extensive treatment. This is an important consideration when using CRISPR spacers for metagenomic analysis, however, because this indicates that CRISPR spacers are not distributed evenly among bacteria and archaea. Any attempts to quantify viral hosts using the CRISPR spacer database must bear this in mind.

To examine this trend more explicitly, we calculated numbers of CRISPR loci per genome (as determined by CRT) and binned the isolates according to growth temperature. The genomes of bacteria and archaea isolated from high temperature environments contain higher numbers of CRISPR loci, on average, than mesophilic or psychrophilic organisms (Fig. 5a). The trend is evident in both the bacteria and the archaea.

However, the number of spacers contained within each CRISPR locus is not constant: while most CRISPR loci contain an average of 30–40 spacers, some contain as few as one or two, while others, such as a CRISPR locus in *Haliangium ochraceum,* contain as many as 600 spacers in a single locus. We calculated the total number of spacers encoded within all CRISPR loci for each genome and correlated this with growth temperature, as before. The trend of increased CRISPR locus abundance in thermophiles (Figure 3.6a) held for CRISPR spacers as well (Figure 3.6b).

This trend is not an artifact of high CRISPR abundance in specific taxonomic groups. For example, this trend can be seen across thermal groups in the methanogens (Fig. 3.6c). The trend also holds within single genera: for example, of the 11 sequenced *Synechococcus* isolates, only three possess CRISPR loci. Two of these three CRISPR-possessing strains are the only thermophilic *Synechococcus* isolates with sequenced genomes.

Outliers do exist in each category (Table 3.8). Most notably, a small number of mesophilic bacteria contain relatively large numbers of spacers in their genomes. While these cases are unusual and should be studied in further detail, they constitute a small minority of mesophilic bacteria: 85% of the over 800 sequenced mesophilic bacteria have between 0 and 2 CRISPR loci.

The reasons for this temperature trend are not yet clear. It is unlikely that the CRISPR overabundance in thermophiles is due to higher diversity among viruses infecting thermophilic hosts, as our diversity modeling results indicate that this is not universally the case (Table 3.2). We also do not expect that the high abundance of CRISPR loci and spacers in thermophiles can be attributed to higher rates of infection in high temperature environments, as studies thus far indicate that virus-to-cell ratios are not necessarily higher in the vent and hot spring environments than in other environments (Srinivasiah *et al.*, 2008). It is possible that CRISPRs are the predominant immunity system in thermophiles, whereas mesophiles favor other types of immunity mechanisms; alternatively, it is possible that the abundance of CRISPR loci in thermophiles is the result of high rates of horizontal gene transfer at high temperatures. However, our current understanding of viral immune systems across the bacteria and archaea and of horizontal gene transfer in different thermal regimes is not thorough enough to distinguish between these possibilities at the present time.

Conclusion

Our results indicate that the 1840 genotypes present in the viral assemblage of this marine diffuse flow hydrothermal vent represent a range of viruses with the potential to infect mesophilic and thermophilic hosts across both the archaeal and bacterial domains. The high evenness of the vent viral assemblage indicates that each of the viral types is fairly equally represented. Therefore, it is likely that viruses infecting a diverse range of hosts are relatively evenly represented in the viral assemblage. This is reflective of the dynamic hydrothermal vent environment, which enables potentially frequent interactions among diverse and extreme microbial communities and their associated viral communities. No other environment possesses the range of physiochemical gradients that characterizes the subsurface vent system, nor the means by which to bring such a wide range of taxonomic groups into close contact. Moreover, the abundance of CRISPR spacers in thermophiles, especially in vent methanogens, suggests that viruses play a unique role in the vent environment, yet the sheer diversity of these spacers attests to the rapid rates of virus-host evolution in these environments. This study, by pairing traditional metagenomic analyses with a novel comparison to a comprehensive CRISPR spacer database, has provided the first insight into the infection potential of the viral assemblage at vents, and opens the way for further studies into how these viruses impact the ecology and evolution of their microbial hosts.

Table 3.1. Summary of physical, chemical, and biological attributes of Hulk vent in the Main Endeavour Field. Temperature minimum was measured by temperature probes on a hydrothermal fluid sampler, temperature maximum was extrapolated based on dissolved silica concentrations.

Physical characteristic	**Value**
Latitude, Longitude	47˚ 57.00' N, 129˚ 5.81' W
Temperature	13–130˚C
Dissolved silica	6378 mmol/L
H_2	2.44 μM
CH_4	28.99 μM
Bacterial counts	1.69 x 10^7 cells/ml[1]
Viral counts	6.80 x 10^6 VLPs/ml[2]

[1] Cell counts prior to filtration through the Steripak filter unit.

[2] Viral-like-particle (VLP) counts after filtration through the Steripak filter unit. VLPs were not reliably countable prior to filtration due to the abundance of biomass and exopolysaccharide material in the sample.

Table 3.2. List of all sequenced strains categorized as hydrothermal vent isolates for this study.

Strain	Number of CRISPR loci	Number of CRISPR spacers	Temperature range
Aciduliprofundum boonei T469	2	23	Thermophilic
Archaeoglobus profundus DSM 5631	0	0	Hyperthermophilic
Ferroglobus placidus DSM 10642	6	101	Hyperthermophilic
Pyrococcus furiosus DSM 3638	7	200	Hyperthermophilic
Pyrococcus horikoshii OT3	6	149	Hyperthermophilic
Staphylothermus marinus F1	12	119	Hyperthermophilic
Archaeoglobus fulgidus DSM 4304	3	149	Hyperthermophilic
Aquifex aeolicus VF5	7	23	Hyperthermophilic
Methanocaldococcus sp. FS406-22	20	238	Hyperthermophilic
Methanopyrus kandleri AV19	0	0	Hyperthermophilic
Methanocaldococcus vulcanius M7	19	219	Hyperthermophilic
Methanocaldococcus fervens AG86	7	77	Hyperthermophilic
Nitratiruptor sp. SB155-2	0	0	Thermophilic
Persephonella marina EX-H1	4	36	Thermophilic
Rhodothermus marinus DSM 4252	9	237	Thermophilic
Thermus thermophilus HB8	10	111	Thermophilic
Hyperthermus butylicus DSM 5456	2	94	Hyperthermophilic
Ignicoccus hospitalis KIN4/I	8	73	Hyperthermophilic
Methanocaldococcus jannaschii DSM 2661	15	177	Hyperthermophilic
Nanoarchaeum equitans Kin4-M	2	41	Hyperthermophilic
Pyrobaculum aerophilum str. IM2	5	131	Hyperthermophilic
Pyrococcus abyssi GE5	4	58	Hyperthermophilic
Thermococcus gammatolerans EJ3	3	40	Hyperthermophilic
Thermococcus kodakarensis KOD1	3	75	Hyperthermophilic
Thermotoga neapolitana DSM 4359	7	58	Hyperthermophilic
Deferribacter desulfuricans SSM1	4	69	Thermophilic
Idiomarina loihiensis L2TR	0	0	Mesophilic
Shewanella loihica PV-4	0	0	Mesophilic
Sulfurovum sp. NBC37-1	0	0	Mesophilic
Thiomicrospira crunogena XCL-2	0	0	Mesophilic

Table 3.3. Percentages of sequences with matches to the SEED database, with a maximum e-value of 10^{-5}. Data was obtained from the MG-RAST database (metagenomics.nmpdr.org).

	Marine Hydrothermal Vent	**Yellowstone-Octopus Spring**	**Yellowstone-Bear Paw**	**Sargasso Sea**	**Gulf of Mexico**	**Bay of British Columbia**	**Arctic**
Unknown	67.14	69.67	38.43	97.58	95.23	97.08	88.14
Bacteria	25.87	16.81	53.50	0.79	4.58	2.89	11.78
Archaea	4.40	9.78	4.25	0.01	0.01	0.00	0.01
Eukaryota	1.90	0.92	2.73	0.02	0.03	0.02	0.06
Viruses	0.69	2.73	0.99	1.60	0.15	0.00	0.01
Plasmids	0.00	0.09	0.10	0.00	0.00	0.00	0.00

Table 3.4. Diversity indices for seven viral metagenomes as calculated by PHACCS (biome.sdsu.edu/phaccs). Hulk hydrothermal vent diversity was modeled twice: (a) modeled diversity from all reads; (b) modeled diversity only from reads in contigs shorter than 3 kbp. See text for details. Model error describes the difference between the modeled contig spectrum from each diversity model and the actual contig spectrum.

Virome	**# reads**	**Avg. Genome Size (bp)**	**Avg. read length for Circonspect**	**Rank-Abundance model**	**Model error**	**Richness** (# genotypes)	**Evenness**	**Most abundant genotype** (% of the community)	**Shannon-Wiener Index**	**Ref**
Hulk hydrothermal vent (a)	228,698	56013	100	Logarithmic	2.62	1730	0.970	2.81	7.23	This study
Hulk hydrothermal vent (b)	216,966	53838	100	Logarithmic	1.22	1840	0.973	2.51	7.32	This study
Yellowstone hot springs-Bear Paw	8,352	35340	650	power	313	1610	0.855	6.27	6.32	Schoenfeld *et al.*, (2008)
Yellowstone hot springs-Octopus	22,272	29086	650	power	281	2340	0.94	2.03	7.29	Schoenfeld *et al.*, (2008)
Arctic Ocean	688,590	57927	100	Lognormal	1.58	886	0.888	3.26	6.03	Angly *et al.,* (2006)
Bay of British Columbia	416,456	65922	100	Logarithmic	35.3	2020	0.952	4.45	7.25	Angly *et al.*, (2006)
Gulf of Mexico	263,907	60921	100	Logarithmic	159	9020	0.906	8.31	8.25	Angly *et al.,* (2006)
Sargasso Sea	399,343	65104	100	Logarithmic	89.4	2990	0.890	10.1	7.12	Angly *et al.*, (2006)

Table 3.5. Results of comparing the CRISPR spacer database against cellular metagenomes and against the marine vent virome. Spacer match results are reported in raw matches versus total base pairs to control for differences in read length and read number. A spacer "match" here is defined as 100% identity across at least 20 base pairs. *cas* genes were identified through the MG-RAST pipeline with a 10^{-5} e-value cutoff. Metagenomes were downloaded from the MG-RAST database. The whale fall and acid mine metagenomes used here were combined from several datasets in the database.

Meta-genome	**MG-RAST number**	**Number of base pairs**	**Average read length (bp)**	**GC Content (%)**	**# spacer matches**	**Ratio spacers** [1]	**No. *cas* genes in meta-genome**	**Ratio cas genes** [2]
Waseca farm soil	4441091.3	154,475,569	1116.58	55.80	740	4.79	17	1.10
Whale fall	4441619.3, 4441656.4, 4441620.3	118,310,832	1017.83, 990.68, 1016.07	46.19	430	3.63	27	2.28
Acid mine drainage	4441137.3, 4441138.3	325,778,024	1041.56, 1004.75	47.97	910	2.79	835	25.6
HOT/ ALOHA upper euphotic	4441051.3	7,482,115	954.72	47.59	30	4.01	0	0
Alvinell a epibiont	4441663.3	290,371,756	990.81	39.90	888	3.06	40	1.38
North Atlantic spring bloom	4443725.3	55,847,247	228.76	48.61	331	5.93	4	0.72
Fishgut	4441695.3	5,076,977	98.59	56.81	30	5.91	0	0
Micro-biolites	4440061.3	26,691,593	103.63	47.26	56	2.10	0	0
Marine vent virome	This study	76,424,561	334.17	37.80	479	6.27	39	5.10

[1] Ratio spacers corresponds to the ratio of total number of spacer matches to base pairs, e-value of E-06.

[2] Ratio *cas* genes corresponds to the ratio of total *cas* gene hits to base pairs, e-value of E-07.

Table 3.6. CRISPR spacer database matches in the marine vent virome. First column lists the groups with spacers having a match to the marine vent virome; second column lists the number of hits in the vent virome to spacers in that group; third column lists the total number of spacers from that group in the CRISPR spacer database.

Group	Number of matches in vent virome to group	Number of spacers from group in database
Firmicutes	109	7796
Bacteroidetes/Chlorobi	78	1556
Gammaproteobacteria	64	5076
Euryarchaeota	63	4195
Crenarchaeota	33	4038
Chloroflexi	24	3188
Thermotogae	21	1392
Cyanobacteria	14	1935
Aquificae	13	519
Actinobacteria	12	3662
Betaproteobacteria	11	1301
Deinococcus-Thermus	7	453
Deltaproteobacteria	5	2365
Alphaproteobacteria	3	1368
Dictyoglomi	3	245
Epsilonproteobacteria	2	302
Fusobacteria	2	47
Nanoarchaeota	2	41
Nitrospirae	2	182
Thermobaculum	2	206
Deferribacteres	1	19
Planctomycetes	1	30
Spirochaetes	1	88

Table 3.7. CRISPR spacer database matches in the marine vent virome, focusing only on species endemic to hydrothermal vents. First column lists the vent species with spacers having a match to the marine vent virome; second column lists the number of hits in the vent virome to spacers in that species; third column lists the total number of spacers from that species in the CRISPR spacer database.

Species	**Number of matches in vent virome to species**	**Number of spacers from species in database**
Methanocaldococcus vulcanius M7	18	219
Methanocaldococcus sp. FS406-22	5	238
Hyperthermus butylicus DSM 5456	3	94
Methanocaldococcus jannaschii DSM 2661	3	177
Thermococcus kodakarensis KOD1	3	75
Methanocaldococcus fervens AG86	2	77
Nanoarchaeum equitans Kin4-M	2	41
Persephonella marina EX-H1	2	38
Thermococcus onnurineus NA1	2	118
Pyrobaculum aerophilum str. IM2	1	131
Pyrococcus horikoshii OT3	1	149

Table 3.8. Top five organisms with the highest number of CRISPR loci per genome for each of the temperature categories listed in Figure 3.6.

Strain	Temp category/group	No. CRISPR loci	No. CRISPR spacers	Isolation environment
Methanocaldococcus sp. FS406-22	Thermophilic archaea	20	238	Hydrothermal vent
Methanocaldococcus vulcanius M7	Thermophilic archaea	19	219	Hydrothermal vent
Methanocaldococcus jannaschii DSM 2661	Thermophilic archaea	15	177	Hydrothermal vent
Staphylothermos marinus F1	Thermophilic archaea	12	119	Hydrothermal vent
Thermofilum pendens Hrk 5	Thermophilic archaea	12	182	Solfataric hot spring
Methanobrevibacter ruminantium M1	Mesophilic archaea	7	129	Bovine rumen
Methanosarcina acetivorans C2A	Mesophilic archaea	6	77	Marine canyon sediments
Methanospirillum hangatei JF-1	Mesophilic archaea	6	263	Sewage sludge
Haloarcula marismortui ATCC 43049	Mesophilic archaea	5	134	Dead Sea
Methanosarcina barkeri str. fusaro	Mesophilic archaea	5	128	Freshwater lake sediments
Roseiflexus castenholzii DSM 13941	Thermophilic bacteria	14	553	Hot springs
Sulfurihydrogenibium azorense Az-Fu1	Thermophilic bacteria	14	158	Hot springs
Thermomonospora curvata DSM 43183	Thermophilic bacteria	14	344	Straw compost
Roseiflexus sp. RS-1	Thermophilic bacteria	13	526	Hot springs
Thermobifida fusca YX	Thermophilic bacteria	12	247	Soil
Cyanothece sp. PCC 7424	Mesophilic bacteria	13	359	Rice fields
Herpetosiphon aurantiacus ATCC 23779	Mesophilic bacteria	12	300	Freshwater lake
Nostoc sp. PCC 7120	Mesophilic bacteria	12	115	Soil
Rhodospirillum rubrum ATCC 11170	Mesophilic bacteria	12	215	Marine
Clostridium difficile 630	Mesophilic bacteria	11	119	Pathogen

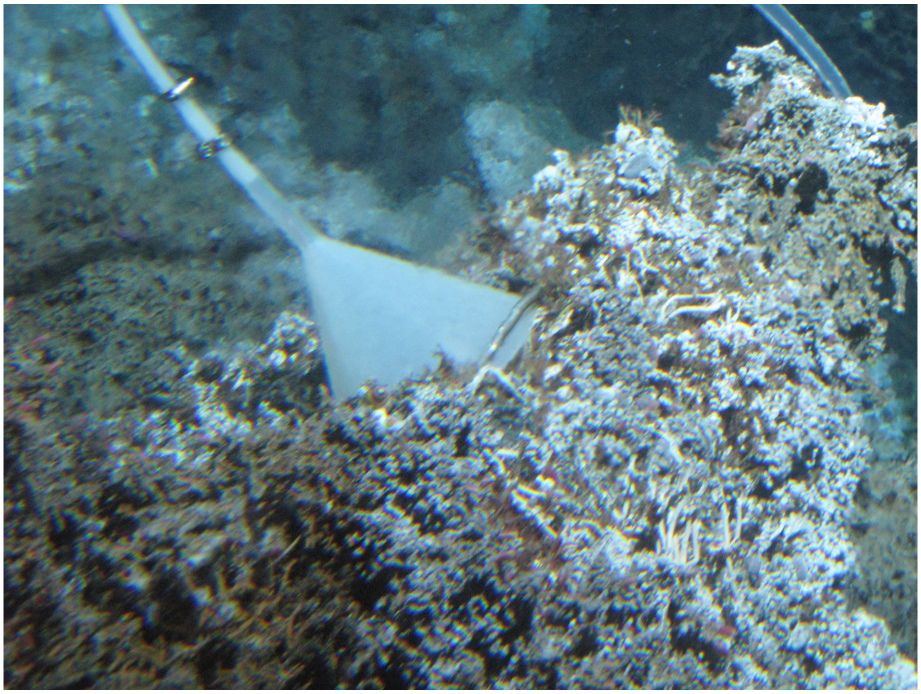


Figure 3.1. Image of the sample intake funnel of the barrel sampler atop a sulfide structure on the side of Hulk vent in the Main Endeavour Field. Blurred background lines in the photo indicate the diffuse fluid that is issuing from the structure, which was the source fluid for the virome in this analysis.

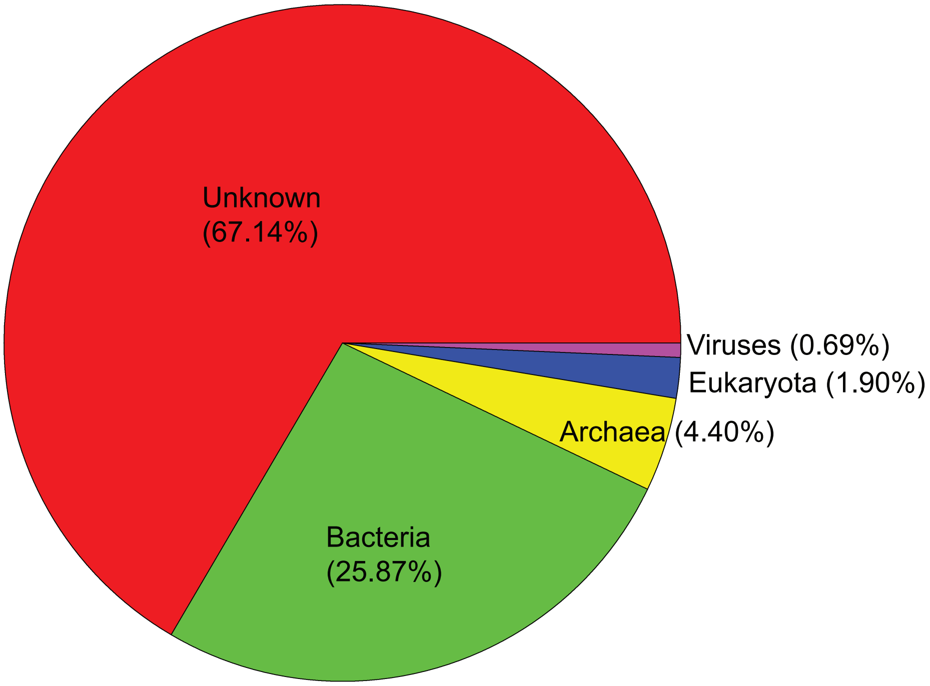


Figure 3.2. Distribution of reads in the hydrothermal vent virome with matches to the SEED database, maximum e-value 10^{-5}. All analyses were performed through the MG-RAST pipeline.

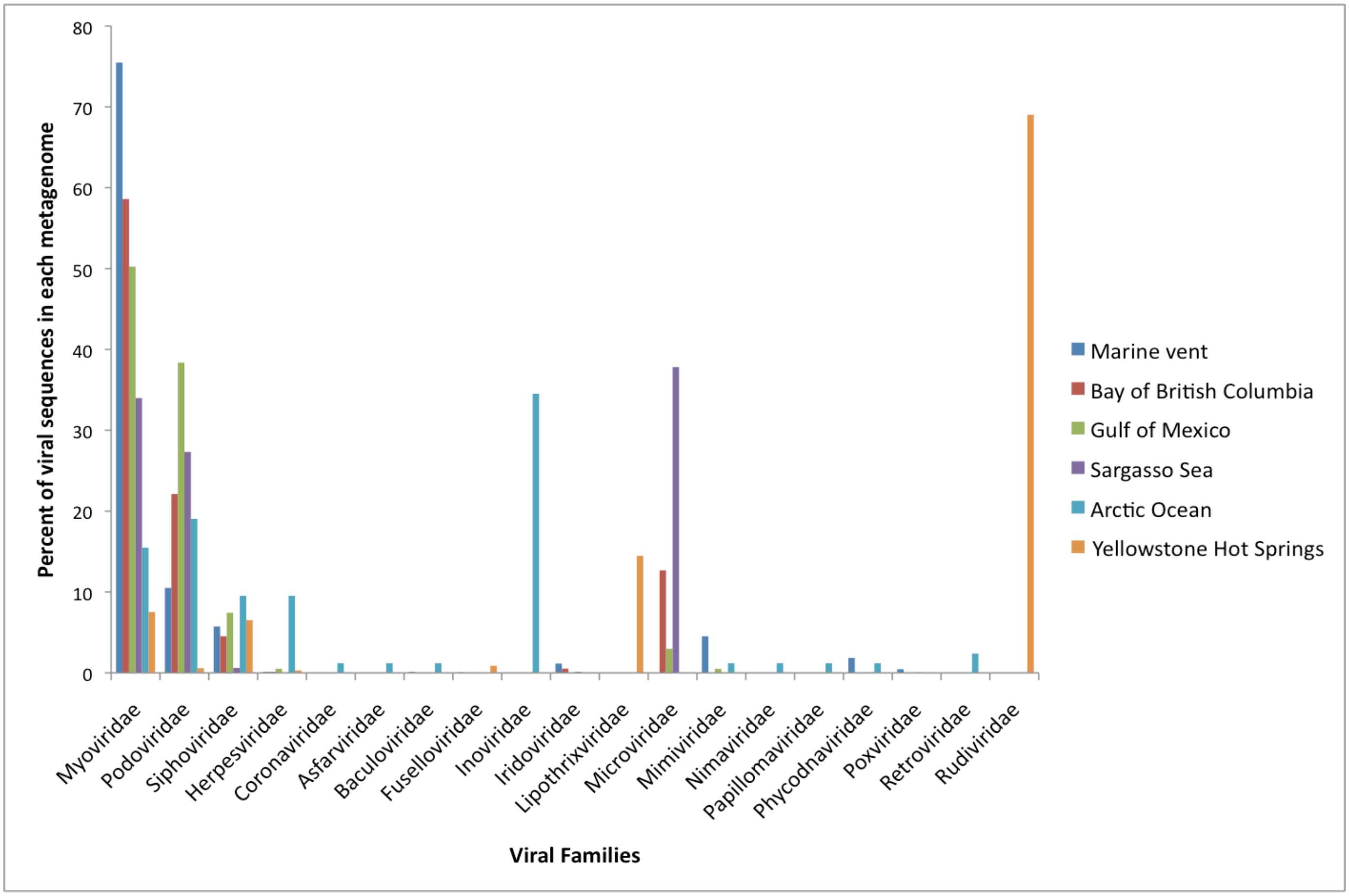


Figure 3.3. Comparison of viral family types present in the hydrothermal vent viral assemblage as well as that of four marine biomes and terrestrial hot springs. Marine viral metagenomes from Angly *et al.* (2006), and Yellowstone hot springs viral metagenomes from Schoenfeld *et al.* (2008).

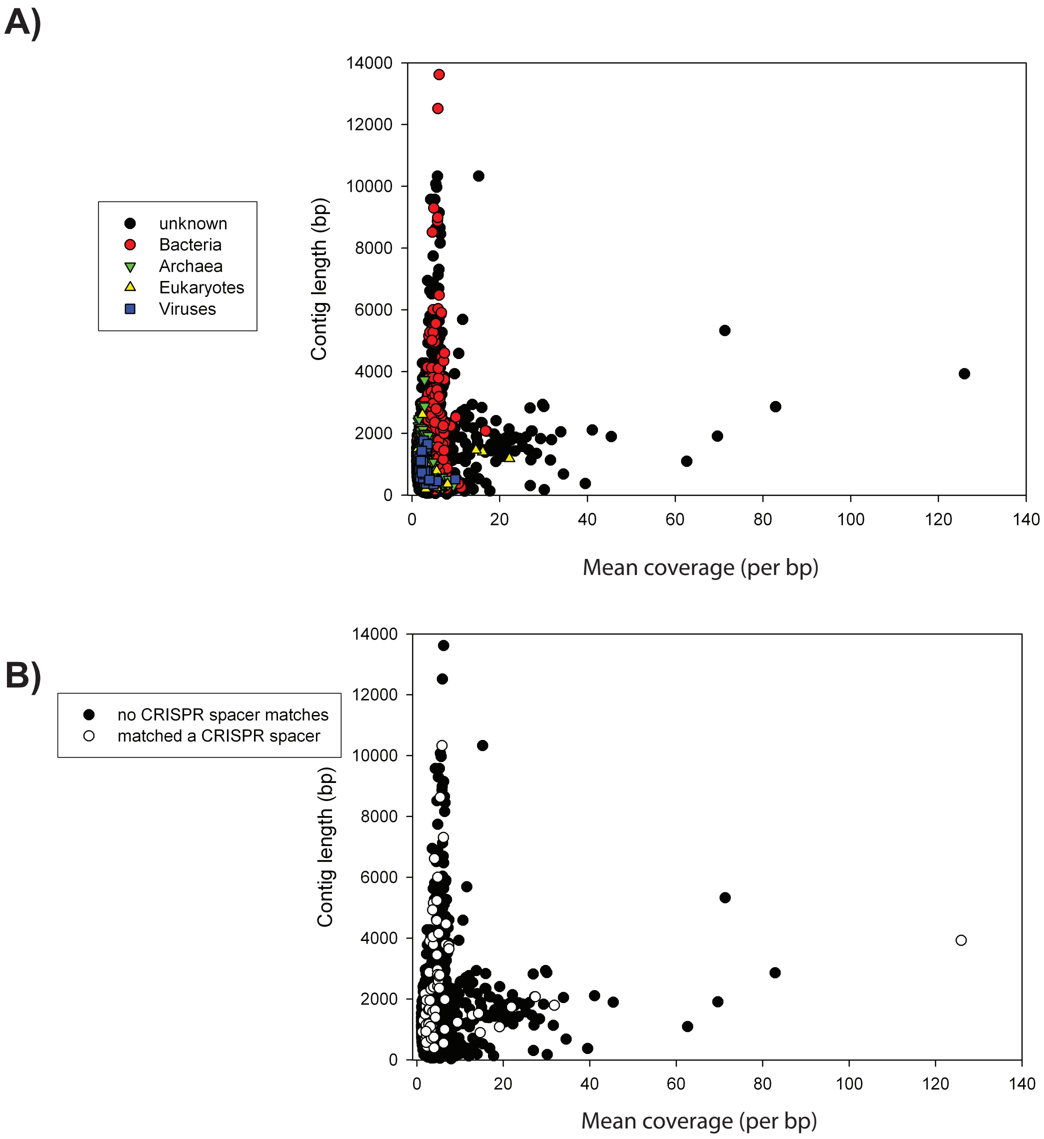


Figure 3.4. Assembly of marine hydrothermal vent virome reads. Contigs were assembled using Geneious (see Methods). Only contigs containing three or more reads are shown. (A), contigs labeled according to domain. Reads were assigned taxa by comparison with the SEED database; contigs were labeled according to the most common taxonomic grouping among constituent reads. (B), contigs labeled according to whether the contig contained a read with a 100% identity alignment of at least 20 bp to a CRISPR spacer.

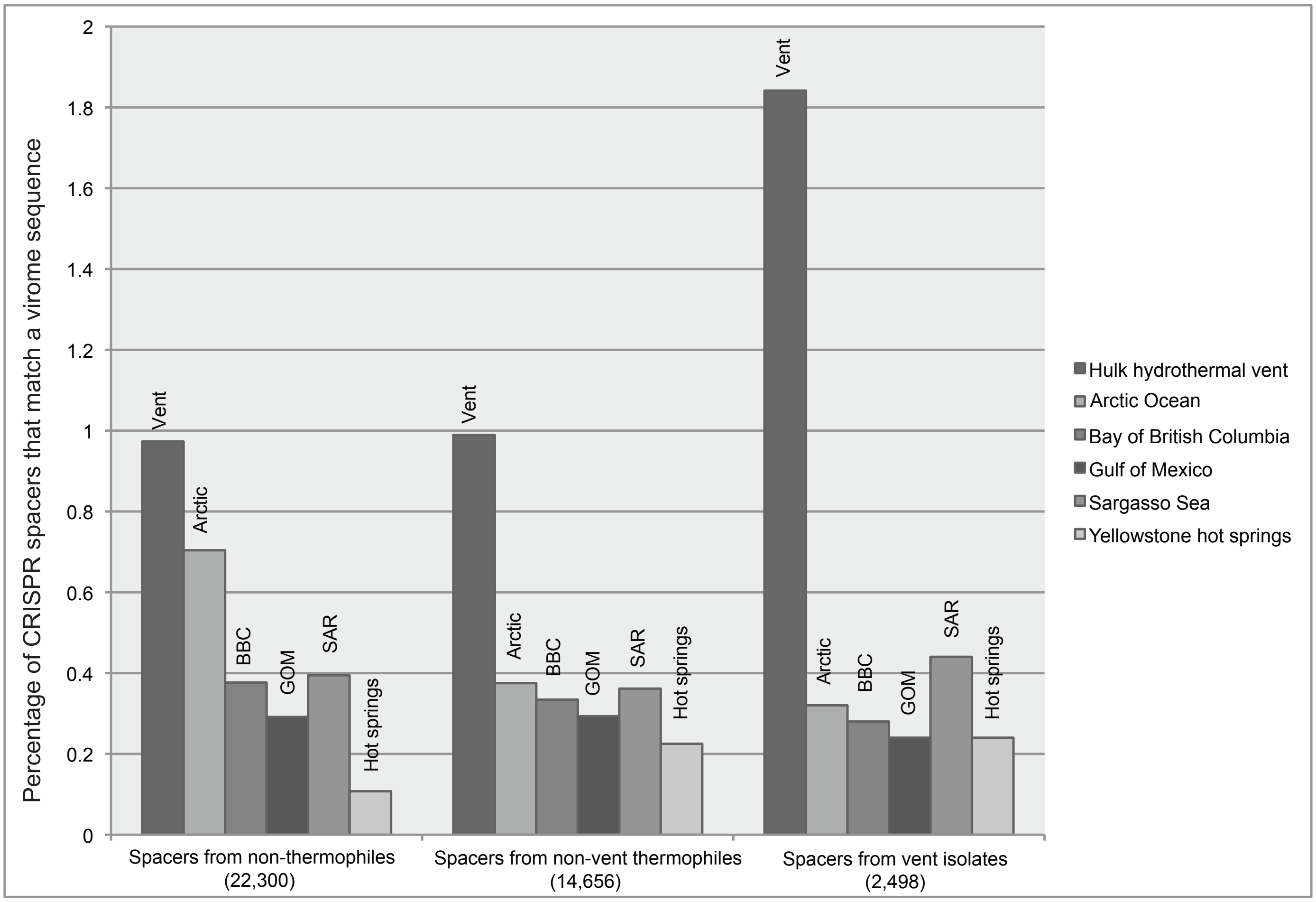


Figure 3.5. CRISPR spacer matches to other marine or hot springs viromes. CRISPR spacers were grouped as derived from vent isolates, non-vent thermophiles, and all other isolates. Sequence similarity searches were performed with blastn, and a “match” was defined as a 100% match across an alignment of 20 base pairs or greater. Numbers under each category on the X-axis indicate the number of spacers in each group.

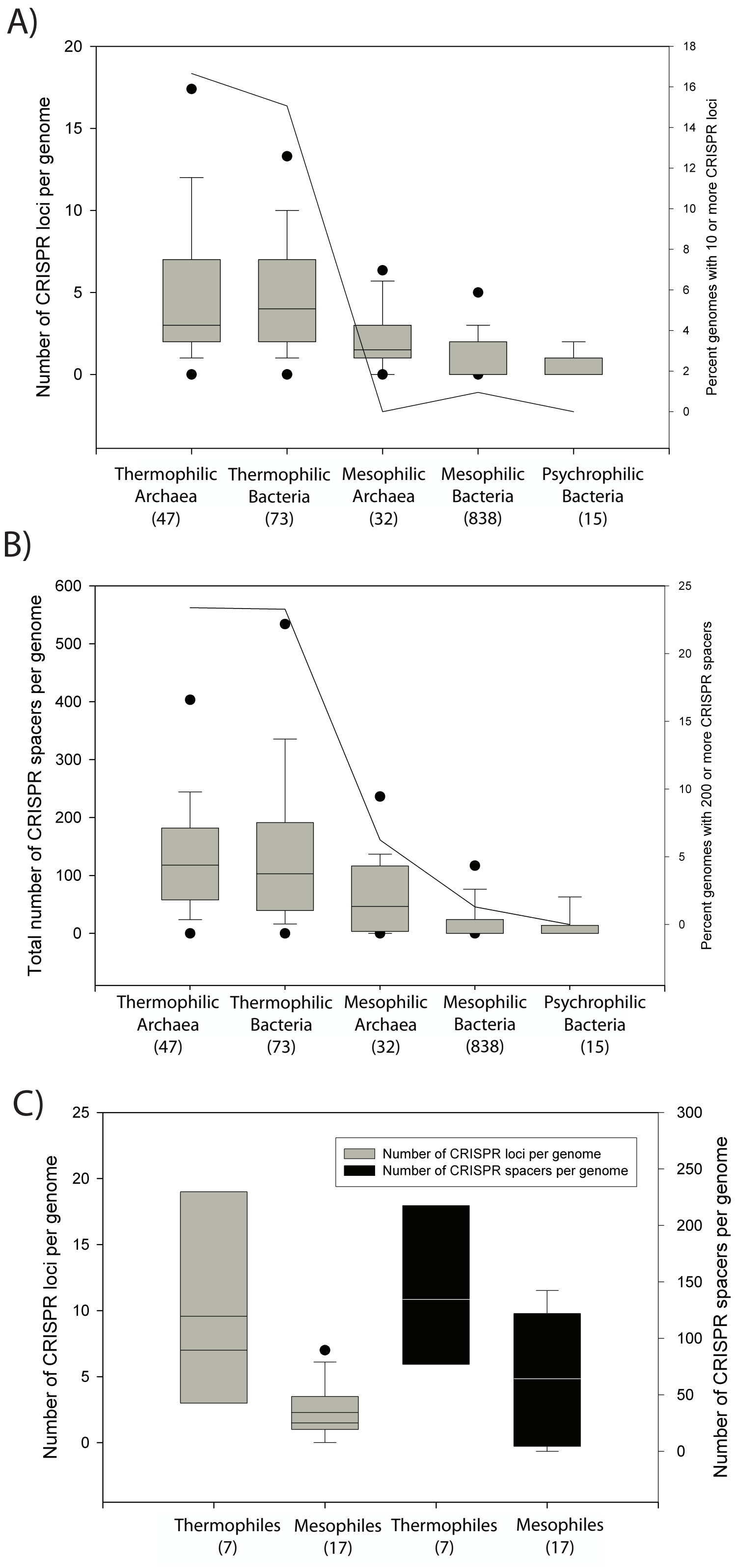

A)
Number of CRISPR loci per genome
20
15
10
5
0
Percent genomes with 10 or more CRISPR loci
18
16
14
12
10
8
6
4
2
0
Thermophilic Archaea (47)
Thermophilic Bacteria (73)
Mesophilic Archaea (32)
Mesophilic Bacteria (838)
Psychrophilic Bacteria (15)
B)
Total number of CRISPR spacers per genome
600
500
400
300
200
100
0
Percent genomes with 200 or more CRISPR spacers
25
20
15
10
5
0
Thermophilic Archaea (47)
Thermophilic Bacteria (73)
Mesophilic Archaea (32)
Mesophilic Bacteria (838)
Psychrophilic Bacteria (15)
C)
Number of CRISPR loci per genome
25
20
15
10
5
0
Number of CRISPR spacers per genome
300
250
200
150
100
50
0
Number of CRISPR loci per genome
Number of CRISPR spacers per genome
Thermophiles (7)
Mesophiles (17)
Thermophiles (7)
Mesophiles (17)

Figure 3.6. (previous page) Abundances of CRISPR loci and spacers in different thermal groups. Numbers below temperature categories list the number of genomes in that category. Box boundaries represent the 25th and 75th percentiles, a line within the box marks the median. Error bars above and below the box indicate the 90th and 10th percentiles. Outlying points represent the 5th and 95th percentiles. Dashed line shows the percent of genomes within each group containing 10 or more CRISPR loci or 100 or more CRISPR spacers. A) Number of CRISPR loci per genome; B) Number of CRISPR spacers per genome; C) Number of CRISPR loci and spacers per genome in methanogens only.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. J Mol Biol **215**: 403–410.

Andersson, A.F. and Banfield, J.F. (2008) Virus population dynamics and acquired virus resistance in natural microbial communities. Science **320**: 1047–50.

Angly, F., Rodriguez-Brito, B., Bangor, D., McNairnie, P., Breitbart, M., Salamon, P., *et al.* (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics **6**: 41.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., *et al.* (2006) The marine viromes of four oceanic regions. PLoS Biol **4**: e368.

Angly, F.E., Willner, D., Prieto-Davó, A., Edwards, R.A., Schmieder, R., Vega-Thurber, R., *et al.* (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Comput Biol **5**: e1000593.

Baross, J.A. and Hoffman, S.E. (1985) Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. Origins Life Evol B **15**: 327–345.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., *et al.* (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science **315**: 1709–1712.

Bench, S.R., Hanson, T.E., Williamson, K.E., Ghosh, D., Radosovich, M., Wang, K., and Wommack, K.E. (2007) Metagenomic characterization of Chesapeake Bay virioplankton. Appl Environ Microbiol **73**: 7629.

Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., and Hugenholtz, P. (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics **8**: 209.

Bolotin, A., Quinquis, B., Sorokin, A., and Ehrlich, S.D. (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology **151**: 2551 –2561.

Breitbart, M. and Rohwer, F. (2005) Here a virus, there a virus, everywhere the same virus? Trends Microbiol **13**: 278–284.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., *et al.* (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A **99**: 14250 –14255.

Brouns, S.J.. J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.. H., Snijders, A.P.. L., *et al.* (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. Science **321**: 960 –964.

Desnues, C., Rodriguez-Brito, B., Rayhawk, S., Kelley, S., Tran, T., Haynes, M., *et al.* (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature **452**: 340–343.

Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M., *et al.* (2008) Functional metagenomic profiling of nine biomes. Nature **452**: 629–632.

Filée, J., Siguier, P., and Chandler, M. (2007) I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. Trends Genet **23**: 10–15.

Fischer, M.G., Allen, M.J., Wilson, W.H., and Suttle, C.A. (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. Proc Natl Acad Sci U S A **107**: 19508.

Garneau, J.E., Dupuis, M.È., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., *et al.* (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature **468**: 67–71.

Garrett, R.A., Prangishvili, D., Shah, S.A., Reuter, M., Stetter, K.O., and Peng, X. (2010) Metagenomic analyses of novel viruses and plasmids from a cultured environmental sample of hyperthermophilic neutrophiles. Environ Microbiol **12**: 2918–2930.

Geslin, C., Gaillard, M., Flament, D., Rouault, K., Le Romancer, M., Prieur, D., and Erauso, G. (2007) Analysis of the first genome of a hyperthermophilic marine virus-like particle, PAV1, isolated from *Pyrococcus abyssi*. J Bacteriol **189**: 4510–9.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007a) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res **35**: W52–W57.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007b) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics **8**: 172.

Haft, D.H., Selengut, J., Mongodin, E.F., and Nelson, K.E. (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol **1**: e60.

Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., *et al.* (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell **139**: 945–956.

Heidelberg, J.F., Nelson, W.C., Schoenfeld, T., and Bhaya, D. (2009) Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. PloS ONE **4**: e4169.

Held, N.L., Herrera, A., Cadillo-Quiroz, H., Whitaker, R.J., and Planet, P.J. (2010) CRISPR Associated diversity within a population of *Sulfolobus islandicus*. PLoS ONE **5**: e12988.

Held, N.L. and Whitaker, R.J. (2009) Viral biogeography revealed by signatures in *Sulfolobus islandicus* genomes. Environ Microbiol **11**: 457–466.

Hendrix, R.W. (2003) Bacteriophage genomics. Curr Opin Microbiol **6**: 506–511.

Hendrix, R.W., Lawrence, J.G., Hatfull, G.F., and Casjens, S. (2000) The origins and ongoing evolution of viruses. Trends Microbiol **8**: 504–508.

Horvath, P., Barrangou, R., and Hovarth, P. (2010) CRISPR/Cas, the immune system of bacteria and archaea. Science **327**: 167–170.

Huber, J.A., Butterfield, D.A., and Baross, J.A. (2003) Bacterial diversity in a subseafloor habitat following a deep-sea volcanic eruption. FEMS Microbiol Ecol **43**: 393–409.

Huber, J.A., Mark Welch, D.B., Morrison, H.G., Huse, S.M., Neal, P.R., Butterfield, D.A., and Sogin, M.L. (2007) Microbial population structures in the deep marine biosphere. Science **318**: 97–100.

Kim, K.-H., Chang, H.-W., Nam, Y.-D., Roh, S.W., Kim, M.-S., Sung, Y., *et al.* (2008) Amplification of uncultured single-stranded DNA viruses from rice paddy soil. Appl Environ Microbiol **74**: 5975–5985.

Kristensen, D.M., Mushegian, A.R., Dolja, V. V, and Koonin, E. V (2009) New dimensions of the virus world discovered through metagenomics. Trends Microbiol **18**: 11–19.

Kunin, V., Sorek, R., and Hugenholtz, P. (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biology **8**: R61.

Labrie, S.J., Samson, J.E., and Moineau, S. (2010) Bacteriophage resistance mechanisms. Nat Rev Microbiol **8**: 317–327.

Lopez-Bueno, A., Tamames, J., Velazquez, D., Moya, A., Quesada, A., and Alcami, A. (2009) High diversity of the viral community from an Antarctic lake. Science **326**: 858.

Makarova, K.S., Grishin, N. V, Shabalina, S.A., Wolf, Y.I., and Koonin, E. V (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biology Direct **1**: 7.

Makarova, K.S., Wolf, Y.I., and Koonin, E. V (2003) Potential genomic determinants of hyperthermophily. Trends Genet **19**: 172–176.

Mann, N.H., Cook, A., Millard, A., Bailey, S., and Clokie, M. (2003) Marine ecosystems: Bacterial photosynthesis genes in a virus. Nature **424**: 741.

Marraffini, L.A. and Sontheimer, E.J. (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet **11**: 181–190.

Mehta, M.P. and Baross, J.A. (2006) Nitrogen fixation at 92˚C by a hydrothermal vent archaeon. Science **314**: 1783.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., *et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics **9**: 386.

Mojica, F.J.., Díez-Villaseñor, C., García-Martínez, J., and Soria, E. (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol **60**: 174–182.

Mojica, F.J.M., Diez-Villasenor, C., Garcia-Martinez, J., and Almendros, C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology **155**: 733.

Van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M., Brouns, S.J.J., van der Oost, J., *et al.* (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem Sci **34**: 401–407.

Ortmann, A.C. and Suttle, C.A. (2005) High abundances of viruses in a deep-sea hydrothermal vent system indicates viral mediated microbial mortality. Deep-Sea Res Pt I **52**: 1515–1527.

Pourcel, C., Salvignol, G., and Vergnaud, G. (2005) CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology **151**: 653.

Prangishvili, D., Forterre, P., and Garrett, R.A. (2006) Viruses of the Archaea: a unifying view. Nat Rev Microbiol **4**: 837–848.

Santos, F., Yarza, P., Parro, V., Briones, C., and Antón, J. (2010) The metavirome of a hypersaline environment. Environ Microbiol **12**: 2965–2976.

Schoenfeld, T., Patterson, M., Richardson, P.M., Wommack, K.E., Young, M., and Mead, D. (2008) Assembly of viral metagenomes from Yellowstone hot springs. Appl Environ Microbiol **74**: 4164.

Schrenk, M.O., Kelley, D.S., Delaney, J.R., and Baross, J.A. (2003) Incidence and diversity of microorganisms within the walls of an active deep-sea sulfide chimney. Appl Environ Microbiol **69**: 3580–3592.

Snyder, J.C., Bateson, M.M., Lavin, M., and Young, M.J. (2010) Use of cellular CRISPR (Clusters of Regularly Interspaced Short Palindromic Repeats) spacer-based microarrays for detection of viruses in environmental samples. Appl Environ Microbiol **76**: 7251.

Sorek, R., Kunin, V., and Hugenholtz, P. (2008) CRISPR — a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat Rev Microbiol **6**: 181–186.

Srinivasiah, S., Bhavsar, J., Thapar, K., Liles, M., Schoenfeld, T., and Wommack, K.E. (2008) Phages across the biosphere: contrasts of viruses in soil and aquatic environments. Res Microbiol **159**: 349–357.

Suttle, C.A. (2005) Viruses in the sea. Nature **437**: 356–361.

Thurber, R. V, Haynes, M., Breitbart, M., Wegley, L., and Rohwer, F. (2009) Laboratory procedures to generate viral metagenomes. Nature Protocols **4**: 470–483.

Tyson, G.W. and Banfield, J.F. (2007) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ Microbiol **0**: 200–207.

Williamson, S.J., Cary, S.C., Williamson, K.E., Helton, R.R., Bench, S.R., Winget, D., and Wommack, K.E. (2008) Lysogenic virus–host interactions predominate at deep-sea diffuse-flow hydrothermal vents. ISME J **2**: 1112–1121.

CHAPTER FOUR

Evolutionary strategies of viruses, bacteria and archaea in hydrothermal vent ecosystems revealed through metagenomics

Summary

The deep-sea hydrothermal vent habitat hosts a diverse community of archaea and bacteria that withstand extreme fluctuations in environmental conditions. Abundant viruses in these systems, a high proportion of which are lysogenic, must also withstand these environmental extremes. Here, we explore the evolutionary strategies of both microorganisms and viruses in hydrothermal systems through comparative analysis of a cellular and viral metagenome, collected by size fractionation of high temperature fluids from a diffuse flow hydrothermal vent. We detected numerous mobile elements in both the viral and cellular gene pools, as well as a large number of prophage in the cellular fraction relative to microorganisms in other environments. We show that the hydrothermal vent viral gene pool was significantly enriched in genes related to energy metabolism, a feature that appears unique to this viral gene pool compared to viral gene pools from other environments, indicating a potential for integrated prophage to enhance host metabolic flexibility. The observation of stronger purifying selection in the viral versus cellular gene pool suggests viral strategies that promote prolonged host integration. Our results support the hypothesis that in a diffuse flow vent environment, viruses maintain genes related to energy metabolism as a means of enhancing host genomic plasticity and adaptability in this extreme and dynamic environment.

Introduction

The deep subsurface below hydrothermal systems hosts a high diversity of archaea, bacteria and viruses that must tolerate extremely variable environmental conditions. High-temperature, reduced hydrothermal fluids mix with cold, oxidized seawater both above and below the seafloor to establish strong gradients in temperature, pH, and chemical and mineralogical composition (Anderson *et al.* 2013a; Anderson *et al.* 2013b; Anderson *et al.* 2011a; Baross & Hoffman 1985; Schrenk *et al.* 2003). Wide variations in environmental parameters can occur over centimeter scales. Constant fluid flux throughout and above the subsurface transports organisms from one region to the next, exposing them to a range of environmental conditions. Gradients that dominate this environment create a highly diverse microbial community that encompasses more than 38000 bacterial operational taxonomic units (OTUs) and over 2700 archaeal OTUs (Huber *et al.*, 2007). Physical and chemical parameters vary according to fluid mixing and volcanic activity, leading to niche partitioning in microbial communities across both space (Anderson *et al.*, 2013a) and time (Huber *et al.*, 2002, 2003). Moreover, hyperthermophiles are routinely cultured from fluids that exit at low temperatures (5–30˚C) (Holden *et al.*, 1998; Summit and Baross, 2001), indicating that organisms in vent systems are frequently flushed from their native habitats, most likely from the deep subsurface. Microbial communities in the subsurface most likely form biofilms along porous mineral structures, resulting high-density communities with high contact rates between organisms.

This dynamic, diverse and dense habitat may foster frequent exchange of genes within the microbial community. Previous work with vent samples has shown that the genes responsible for this process, including transposases and integrases, occur at high frequency in hydrothermal systems compared to other environments (Elsaied *et al.* 2007; Brazelton & Baross 2009). Fully sequenced genomes of thermophiles, including many from vent systems, provide evidence of frequent gene transfers that sometimes cross domains (Nelson *et al.* 1999; Beiko *et al.* 2005; Koonin *et al.* 2001). The prevalence of horizontal gene transfer in vent systems may expand the functional repertoire of the organism, allowing individual taxa to become more metabolically flexible. This expanded flexibility would provide a strong advantage in hydrothermal vent environments where fluid flux and environmental gradients expose communities to wide extremes in temperature, pH, and chemical composition.

Here, we use metagenomics to elucidate the role that viruses play in facilitating gene flow and manipulating host genetic potential in hydrothermal systems, environments not previously studied from this perspective. Viruses play pivotal roles in the transfer of genes and the alteration of host phenotype, particularly in the pelagic oceans (see Breitbart 2012 for review). Bacterial and archaeal viruses introduce foreign genetic material through transduction and expression of virally encoded genes during infection. Transduction, or virally-mediated horizontal gene transfer, occurs on a massive scale in the surface oceans. Up to 10^{14} transduction events can occur per year in Tampa Bay estuary (Jiang & Paul 1998), and virus-like particles that serve as gene transfer agents (GTAs) may boost these transduction rates by one million-fold (McDaniel *et al.* 2010). Viruses are known to encode auxiliary metabolic genes, or AMGs, which play critical roles in facilitating biochemical or metabolic processes (Breitbart *et al.*, 2007). For example, cyanophage transcribe and express photosynthetic genes during lytic infection of their cyanobacterial hosts (Lindell *et al.*, 2005, 2007; Clokie *et al.*, 2006). In lysogenic phage, the expression of virally encoded genes that persist over multiple generations can manipulate host phenotype, such as in the case of the cholera toxin expressed by a lysogenic bacteriophage integrated in the *Vibrio cholerae* genome (Waldor and Mekalanos, 1996). Selection should favor expression of genes within the integrated phage that enhance host fitness. It has been hypothesized that lysogenic phage conserve resources under low-energy or low-nutrient conditions by expressing genes that suppress host metabolism (Paul 2008). Here, we are testing the hypothesis that in the vent environment phage enhance metabolic flexibility by encoding genes not present on the host genome.

Despite increasing evidence that viruses play a crucial role in manipulating host genotype and phenotype throughout the oceans, this phenomenon has yet to be explored in the dynamic environment of hydrothermal vents. Viruses are abundant in hydrothermal systems (Ortmann & Suttle 2005) and have the potential to infect a wide range of hosts (Anderson *et al.* 2011a). In the deep ocean and at vent sites in particular, a high percentage of cells contain lysogenic prophage (Williamson *et al.* 2008). Considering the abundance of viruses in these systems, and lysogenic viruses in particular, several questions arise: are these viruses transferring genes between hosts? Are they expressing fitness factors while integrated as prophage? If so, which genes are expressed? Are viruses contributing to host genomic plasticity and facilitating their adaptation to changing conditions? Does selection act differently on virally expressed genes compared to cellular genes?

To address these questions, we used a cultivation-independent approach that provides a community-wide perspective of both the viral gene pool and the bacterial and archaeal gene pool (hereafter referred to as the “cellular” gene pool) in hydrothermal systems. Specifically, we analyzed the viral and cellular metagenomes of high-temperature diffuse flow hydrothermal fluid from Hulk hydrothermal vent in the Main Endeavour Field on the Juan de Fuca Ridge. We compared the relative content of each of these gene pools and inferred the modes of genetic interaction between viruses and their hosts. The fluid sample, which most likely sampled both cool seawater and hot fluid from the subsurface, contained organisms that inhabited niches spanning the associated environmental gradients. Given the potential to share viruses and exchange genes across these niches, these metagenomes may provide unique insights into the interactions within the communal gene pool of the hydrothermal vent microbial community.

Here, comparative analysis of the cellular and viral metagenomes from this sample addressed whether viruses facilitate adaptation to environmental dynamism by contributing to host genomic plasticity. The presence of genes facilitating horizontal gene transfer and prophage integration described the genetic potential for these processes in the vent environment. Searches for hypervariable genomic islands indicated what types of genes were successfully transferred in the vent environment and provided evidence of genomic plasticity in host genomes. We compared these results to the relative abundance of genes in the viral and cellular gene pools in order to determine whether these transferred genes were enriched in the viral gene pool. Finally, we asked how evolution has shaped the viral and cellular gene pools by examining relative selection pressures on viral and cellular genes. Together, these analyses provide insight into the broader question of how evolution has shaped virus-host interactions in some of the more extreme environments of the planet.

Materials and Methods

Sample collection and DNA extraction

The 170-L hydrothermal vent fluid sample was collected from Hulk vent at the Main Endeavour Field on the Juan de Fuca Ridge (47°57.00′ N, 129°5.81′ W) using a large barrel sampler, as described previously (Anderson *et al.* 2011a). The vent fluid was obtained using a large barrel sampler equipped with two 100-L sterile bags. A sample collection funnel attached to the sampler was placed atop a region of diffuse venting, adjacent to a colony of tube worms on the side of a large sulfide structure. While the tube worms were surrounded by fluid at measured temperatures of 13–30˚C, the average temperature of the metagenome fluid sample was calculated from its silica chemistry to be about 125˚C (Anderson *et al.* 2011a). This result indicates that this sample most likely pulled fluid from many different niches, including both cool background seawater and high-temperature hydrothermal fluid (up to 300˚C) from the sulfide structure adjacent to the sample site. The organisms collected in the sample therefore represent a range of habitats in the hydrothermal environment, all with the potential to come into contact through fluid flux.

The cellular fraction was collected by filtering the 170 L of hydrothermal vent fluid through three 0.22 µl Steripaks (Millipore, USA) while the sample and filtrate were held on ice. The filtrate was retained for subsequent virus sampling. Filters were frozen at –80˚C while shipboard and until sample processing. DNA was extracted from one Steripak using a modified DNA extraction procedure described by Anderson *et al.* (2013a). Briefly, DNA extraction buffer (0.1 M Tris-HCl, 0.2 M Na-EDTA, 0.1 M NaH2PO4, 1.5 M NaCl, and 1% cetyltrimethylammonium bromide) was added to each filter, then the filters were capped and freeze-thawed five times. Lysozyme (50 mg/mL solution), proteinase K (1% solution), and SDS (20% solution) were added to each filter and incubated. Lysate was removed from filters and centrifuged; DNA was extracted from the supernatant using a phenol/chloroform/isoamyl extraction method described by Anderson *et al.* (2013a).

For virus collection, the sample filtrate was concentrated using tangential flow filtration (30 kDa cutoff) to approximately 400 mL in a 4˚C cold room. Concentrated filtrate was frozen into six aliquots at –80˚C until further processing. One aliquot was further concentrated by adding 10% w/v polyethylene glycol 8000 (PEG), incubating overnight at 4˚C, and centrifuging at 13 000 x g for 50 min. The pellet was resuspended in Tris-EDTA buffer and incubated for 15 min with 0.7 volume of chloroform to lyse any remaining cellular contamination. Free DNA was removed by incubating with 10% DNAse I for 2h at 37˚C, then inactivated by adding EDTA to a final concentration of 0.02M. The QIAamp MinElute Virus Spin Kit (Qiagen) was used to extract the viral

DNA, which was not amplified for downstream sequencing. For further details regarding viral metagenome preparation and analysis, see Anderson *et al.* (2011a).

Metagenomic sequencing

The viral metagenome was generated on a Roche Genome Sequencer FLX (GSFLX) with GS FLX Titanium 454 sequencing protocols by the Broad Institute. For the cellular metagenome, libraries were created using the Nexterra transposon-mediated method (Epicentre) at the Josephine Bay Paul Center at the Marine Biological Laboratory, then sequenced using Roche Titanium 454 sequencing protocols on a GSFLX. Both metagenomes are publicly available on the MG-RAST database (Meyer *et al.* 2008), with accession numbers 4448187.3 for the viral metagenome and 4481541.3 for the cellular metagenome.

Tags were trimmed from the 5' end of each sequence in the cellular metagenome using TagCleaner (Schmieder *et al.*, 2010). Assembly of both the viral and cellular metagenomes was conducted in Geneious (Drummond *et al.*, 2009) using the "Medium Sensitivity" method, with a word length of 14, a maximum gap size of 2, maximum gaps per read of 15, and maximum mismatches of 2. To classify sequences using di, tri, and tetranucleotide analysis, a boutique database of bacterial and archaeal virus sequences (Table 4.1) was created as a training set to accompany the existing cellular dataset in PhylopythiaS (McHardy *et al.* 2007), which was used to identify archaea, bacteria, archaeal viruses and bacterial viruses. Metagenomes were assembled in Geneious prior to classification with PhylopythiaS; only contigs over 1000bp in length were used.

CRISPR analyses

We identified Clustered Regularly Interspaced Palindromic Repeats (CRISPRs) (Barrangou *et al.*, 2007; Brouns *et al.*, 2008; Sorek *et al.*, 2008; van der Oost, J. *et al.*, 2009; Horvath *et al.*, 2010; Labrie *et al.*, 2010; Marraffini and Sontheimer, 2010; Hale *et al.*, 2009; Jore *et al.*, 2011; Garneau *et al.*, 2010) in assembled cellular metagenome reads to evaluate the degree to which these viruses actively infect cells in this study site. We defined a "match" as a 100% identity over at least 20 base pairs (Anderson et al., 2011a).

Enrichment of prophages and mobile genetic elements

To identify the numbers of reads in each metagenome matching prophages, metagenomes were compared to a database of sequences from the "Prophages" category in the ACLAME database (Leplae *et al.*, 2010). To assess abundance of mobile genetic elements, metagenomes were compared to a dataset of Pfam seed sequences (Finn *et al.*, 2010) matching transposases, recombinases, resolvases and integrases. A full list of these elements is provided in Table 4.2. Analysis used tblastn with an e-value cutoff of 10^{-5}. The number of unique reads with a match to a sequence from the query sequence collection was tallied and normalized to the number of reads in the metagenome. Only metagenomes generated with 454 pyrosequencing were used for the analysis, so that all metagenomic reads had a length ranging from approximately 100 to 300 bp. Prophages were identified on bacterial and archaeal genomes using Prophage Finder (Bose and Barber, 2006) with default settings.

Assessing protein richness in metagenomes

To assess relative protein- encoding gene diversity of the metagenomes analyzed here, we used FragGeneScan (Rho *et al.*, 2010) to identify open reading frames (ORFs) for each metagenome and used 50% identity to form clusters of protein-encoding genes. For clustering, reads from each metagenome were sorted according to length in USEARCH 5.2.32 (Edgar, 2010) using a minimum length of 20, then reads were clustered to 50% identity with the following parameters: -id 0.50 - maxrejects 100 - maxaccepts 8 –minlen 20. USEARCH output was formatted to CD-HIT (Li and Godzik, 2006) output in USEARCH, then formatted to .list input for mothur (Schloss *et al.*, 2009) using in-house Python scripts. Rarefaction curves were generated in mothur.

Fragment recruitment

Cellular metagenomic reads were recruited to genomes of hydrothermal vent isolate using NUCmer, part of the MUMmer 3.0 package (Kurtz *et al.*, 2004), and the following parameters for the command line: -minmatch 10 -breaklen 1200 -maxgap 1000 -mincluster 50. Fragment recruitments were visualized using mummerplot in the MUMmer package. Coverage plots were created by using the show-coords command in

MUMmer, then in-house Python scripts were used to calculate coverage for each base pair position. Coverage plots were created with a convolution function in numpy, using a moving average window size of 50000.

Relative enrichment of gene categories

Gene categories were tallied by adding “abundance” counts for each functional category as defined by the KEGG Orthology database (Kanehisa *et al.* 2012) or the SEED Subsystems database (Overbeek *et al.*, 2005) in MG-RAST, using an e-value cutoff of 10^{-5}. For the combined analysis of 20 cellular metagenomes and 23 viral metagenomes, all abundance counts for either viral or cellular metagenomes were tallied together. To determine significance, abundances were entered into Xipe-Totec (Rodriguez-Brito *et al.*, 2006), a nonparametric method of statistical analysis using a difference of medians analysis. For this analysis, we used a confidence level of 95% and a sample size of 5000 to determine significance.

Calculation of dN/dS

Prior to calculation of dN/dS ratios for genes mapped by each metagenome, metagenomes were subjected to stringent error filtering using Prinseq (Schmieder & Edwards 2011) with the following parameters: minimum sequence length of 60bp; minimum mean quality score of 30; maximum number of allowed Ns per sequence of 4; and low-complexity threshold of 70 (using Entropy). The dN/dS ratio measures selection pressures by calculating whether the number of non-synonymous substitutions (dN) in a gene is greater or fewer than the number expected by chance compared to the number of synonymous substitutions (dS). A majority-rule consensus was calculated from the mapped reads; the number of possible synonymous or nonsynonymous substitutions was then tallied and compared to the number of actual synonymous and nonsynonymous substitutions.

Reads from both the viral and cellular metagenomes were mapped to the vent isolate genomes using CLC Genomics Workbench and the criteria of 80% identity and 80% coverage, as before (Tai *et al.*, 2011). Mapping results were exported in ACE format; dN/dS was calculated for each gene using the Python scripts described in Tai *et* *al.* (2011). Polymorphisms were only tallied for positions with a mapping depth of at least 5X; only genes with at least 100 nucleotides at 5X depth were included in the analysis. Redundant genes were deleted from the analysis; only the dN/dS value for the gene with higher coverage was retained. The files used to define gene coordinates were downloaded from JGI IMG, with the exception of *T. kodakarensis*, which was derived from the .gff file from NCBI. The 95% confidence interval was calculated for all genes mapped by the viral and cellular metagenomes with dN/dS less than 1 (subject to purifying selection) using alpha = 0.05.

Results and Discussion

General features of the metagenomes

Metagenomic sequencing of the cellular and viral fractions from the large sample of hydrothermal fluid yielded a total of 808,051 and 231,246 sequence reads, respectively. Of these, approximately 40% of the cellular metagenome and 64% of the viral metagenome (virome) sequences had no matches to the M5NR database (Wilke *et al.*, 2012). Matches of CRISPR spacers identified in the cellular metagenome with sequences in the viral metagenome suggest an active and relatively recent relationship between the two gene pools. Classification of the cellular and viral metagenomes showed that only 2% of the reads from the cellular fraction matched archaeal genes, whereas 4% of the viral metagenome reads matched archaeal genomes (Figure 4.1). Nucleotide signature matching of assembled contigs longer than 1000 bp using PhylopythiaS (McHardy *et al.* 2007) indicated that a disproportionate percentage of contigs in both metagenomes matched nucleotide compositional patterns of archaeal viruses, given that bacterial reads dominate both metagenomes (Figure 4.2). The percentage of contigs matching bacterial nucleotide compositional patterns was greater than the percentage of contigs with archaeal patterns by a ratio of 2.8 to 1 in the cellular metagenome. However, contigs matching bacterial virus patterns outnumbered contigs with archaeal virus patterns by only 1.8 to 1 in the viral metagenome. Taken together, these results suggest that archaeal viruses may be disproportionally abundant in the vent habitat compared to the relative abundance of bacteria to archaea.

Compared to the cellular metagenome, the viral metagenome contained a higher abundance of high-coverage (> 20X coverage) but short (shorter than 6kb) contigs (Figure 4.3). The short length and high coverage suggests that these contigs are derived from viral genomes, possibly suggesting that some very high frequency viral genomes dominate the entire assemblage. Many of the longer but low-coverage contigs in the viral metagenome appear to derive from cellular contamination, as evidenced by blast searches (Anderson *et al.* 2011a). In an attempt to reduce the number of reads likely derived from cellular contamination, we created a subset of reads based on contig coverage and annotation. Contig taxonomy was annotated according to the consensus taxonomy of reads within each contig according to the classification to the SEED database, with a maximum e-value cutoff of 10^{-5}. Reads contained within contigs with a coverage of 8 or greater, as well as reads within all contigs annotated as "unknown" or "viral," were included in what will be referred to as the "viral subset" for the remainder of this paper. While some of the contigs annotated as "unknown" may be derived from unknown regions on contaminating cellular DNA, a higher proportion of the viral metagenome was annotated as "unknown" compared to the cellular metagenome, suggesting that many of the unknowns were viral. The goal of this subset was not to generate a "pure" viral metagenome but rather to reduce the number of reads that may represent cellular contamination.

Assessing the potential for horizontal gene transfer

To assess the degree to which cells and viruses in hydrothermal ecosystems contain necessary machinery for horizontal gene transfer or integration of prophage, we determined the relative abundances of prophage genes (Table 4.3) and genes related to DNA transfer or mobilization (Table 4.4) in the hydrothermal vent cellular and viral metagenomes. We used the "Prophage" dataset in the ACLAME database (Leplae *et al.*, 2010) to identify prophage-related proteins in 22 pyrosequenced cellular metagenomes. These metagenomes were chosen to represent a range of aquatic and terrestrial environments, while controlling for sequencing method. Table 4.3 indicates that of the 22 cellular metagenomes, the hydrothermal vent cellular metagenome contained the second-highest percentage of reads (approximately 4%) that match prophage-coding regions.

This result provides compelling molecular evidence of abundant prophages in cellular genomes in vents and complements descriptions of high proportions of lysogenic cells in vents and the deep ocean based upon mitomycin C induction experiments (Weinbauer *et al.* 2003; Williamson *et al.* 2008). We also analyzed the relative abundance of mobile genetic elements in 40 viral and cellular metagenomes, also selected to represent a range of environments and controlled for sequencing method. Again, we found a relative enrichment of mobile genetic elements in both the viral and cellular gene pools at Hulk hydrothermal vent (Table 4.4). Abundant mobile elements in the form of transposases were previously observed at Lost City (Brazelton & Baross 2009) and Elsaied *et al.* (2007) have described prevalent integrases from the Suiyo Seamount and the Mariana Arc. On average, the genomes of archaea and bacteria in vent systems appear to contain more mobile elements than genomes native to other habitats, leading us to hypothesize that the dynamic vent environment selects for increased rates of horizontal gene transfer (HGT) mediated by lytic and lysogenic viruses.

Evidence of genomic plasticity in vent genomes

Having established the genetic potential for both lysogeny and gene transfer in the viral and cellular gene pools, we sought evidence for either prophage integration or gene transfer events in hydrothermal vent isolates. Of the 34 vent genomes investigated, 20 (59%) contained integrated prophages (Table 4.5). Most of these prophages encoded capsid genes or genes such as DNA ligases, which have been identified before as particularly abundant in the viral fraction (Anderson *et al.*, 2011b). However, identification of auxiliary metabolic genes (AMGs) in these prophage genomes is difficult, partly due to the high abundance of unknown genes and partly because the boundaries of the prophage genome are not always clearly delineated. To better identify regions that have been transferred in vent genomes, we compared genomes from bacterial or archaeal isolates with sequences sampled directly from the environment. This strategy can identify potential hypervariable regions, or "genomic islands," that display lower coverage than the rest of the genome (Coleman *et al.*, 2006; Cuadros-Orellana *et al.*, 2007; Rodriguez-Valera *et al.*, 2009). Previous work with *Haloquadratum walsbyi* DSM 16790 (Cuadros-Orellana *et al.*, 2007) and *Prochlorococcus* genomes (Coleman *et al.*,

2006) identified genomic islands that most likely represented regions of phage-mediated lateral gene transfer; in the case of *Prochlorococcus,* these genes are differentially expressed under light and nutrient stress (Coleman *et al.*, 2006).

We performed fragment recruitment of the hydrothermal vent cellular metagenome and viral subset against the genomes of all isolates of hydrothermal vent bacteria and archaea available in the NCBI database. In most cases, the metagenomes did not recruit to the isolate genomes with high enough coverage to yield useful data. *Nautilia profundicola* AmH successfully recruited reads at high coverage, but no metagenomic islands were found. However, recruitment of the cellular genome to the longest contig (ABCJ01000001) of the draft genome of *Caminibacter mediatlanticus* TB-2 yielded data of interest. *C. mediatlanticus* TB-2 is a chemolithotrophic, nitrate-ammonifying Epsilonproteobacterium that was isolated from the walls of a hydrothermal vent chimney on the Mid-Atlantic Ridge (Voordeckers *et al.*, 2005). Fragment recruitment yielded a number of regions with relatively low coverage. A series of CRISPR loci were detected between 150000 and 200000 bp, accompanied by a slight decrease in coverage in this region. Since they are dedicated to viral and plasmid immunity by effectively creating a library of previous infection, recruitment to CRISPR loci would naturally yield lower coverage, particularly for a metagenome sampled in a different geographic location than this isolate. Two distinct genomic islands were identified: one region of approximately 34 kbp, followed by a second low-coverage region of approximately 10 kbp (Figure 4.4), separated from each other by a 20 kbp region that includes a ribosome. The first genomic island coincides with a region with relatively high GC content, which suggests this region was transferred into the *C. mediatlanticus* genome. A phage integrase gene is located approximately 58 kbp downstream of the 3’ end of the first genomic island, though its presence there is not necessarily conclusive evidence that the region was introduced by a phage. The first genomic island begins near a tRNA gene, a common site for integration of horizontally transferred regions (Reiter *et al.*, 1989). Many of the genes in both the first and second genomic islands encode proteins that interact with the environment, including sugar and nitrate membrane transporters, and proteins related to energy metabolism, including hydrogenases (Table 4.6). Possession of a diverse suite of hydrogenases can enhance metabolic flexibility in variable redox conditions, and has been observed in other vent Epsilonproteobacteria (Campbell *et al.*, 2009).

The presence of this hypervariable region indicates that genomic islands from genomes in vent environments encode genes related to environmental interactions and energy metabolism. The genomic island shown here is unique to a vent isolate from the Mid-Atlantic Ridge. The *C. mediatlanticus* strains present in our sample on the Juan de Fuca Ridge most likely have genomic islands of their own, though these cannot be identified without a fully sequenced strain from the Juan de Fuca Ridge. None of the sequenced strains from the Juan de Fuca Ridge had high enough coverage with our metagenome to identify genomic islands. However, the genomic island on *C. mediatlanticus* provides evidence of horizontal transfer of genes that facilitate metabolic flexibility in an environment that is very similar to the Juan de Fuca Ridge. From this we can hypothesize that similar genes are transferred in our sample site.

What types of genes are enriched in the viral fraction?

If viruses are an important mediator of gene transfer, then the viral fraction may be enriched in genes with similar functionality to those identified in the metagenomic islands of *C. mediatlanticus*. As demonstrated above, they are likely to be genes that facilitate metabolic flexibility and thus enhance host fitness in the dynamic environment.

To determine relative enrichment of gene types in the cellular and viral fractions, we directly compared the relative abundance of genes in different functional categories between the cellular metagenome and the virome subset. Figure 4.5a depicts the percent of reads in each metagenome that match a certain functional category in the SEED Subsystems database (Overbeek *et al.*, 2005). Asterisks indicate selected functional categories that are significantly enriched in one metagenome over the other. The results are similar to those found in a study by Kristensen *et al.* (2009), who compared 42 viral and cellular metagenomes, annotated with the SEED Subsystems database, from a range of different environments (Kristensen *et al.*, 2009). Their results indicated a strong positive correlation between the gene distributions of different functional categories. Kristensen *et al.* (2009) point out that this correlation may be due in part to the choice of available functional categories, which encompass functions that are generally cellular rather than viral. However, significant differences in enrichment of certain gene types were detected in our study. The cellular metagenome was significantly enriched with genes related to the stress response, iron acquisition and metabolism, and metabolism of aromatic compounds, none of which play known roles in viral replication or packaging. As expected, the viral fraction contained many more genes related to phage, prophage, transposable elements, and plasmids relative to the cellular fraction. The large number of reads that classified to the "DNA metabolism" category most likely reflects the dependence of viral replication on DNA synthesis. As the viral metagenome subset had reduced gene richness relative to the cellular metagenome (Figure 4.6), the viral fraction likely has an overrepresentation of gene types necessary for viral function. Interestingly, the viral subset exhibited an abundance of genes related to cofactors, vitamins, prosthetic groups and pigments, yet there are few reports that viruses require these functions. This feature is not entirely unique to the vent viral gene pool we examined, as a similar enrichment was found in a combined analysis of 42 viral and cellular metagenomes (Kristensen *et al.*, 2009). This enrichment may reflect selection for genes to support cellular function while viruses are integrated as prophage. Kristensen *et al.* (2009) also posited that the abundance of certain cell-like genes enriched in the viral fraction may point to an abundance of gene transfer agents (GTAs), phage-like particles that transduce seemingly random assortments of genes from their cellular host.

Figure 4.5b reflects a similar analysis, but is categorized according to functional groups in the KEGG Orthology (KO) database (Kanehisa *et al.* 2012; Kanehisa & Goto 2000). We directly compared these results against an analysis of 20 cellular metagenomes and 23 viral metagenomes, all sequenced with pyrosequencing and annotated with the KO database, and all of which were sampled such that the cellular metagenomes had viral counterparts. These metagenomes are summarized in Table 4.7; the functional profiling analysis is depicted in Figure 4.7. Both the Hulk viral metagenome (Fig. 4.5) and the combined viral metagenomes (Fig. 4.7) were significantly enriched in genes related to nucleotide metabolism. The Hulk metagenome is also particularly enriched in genes related to replication and repair, which probably serve necessary functions in the synthesis and replication of viral DNA during the course of infection. The "replication and repair" category includes DNA ligases, which occur at very high abundance in the virome, as previously reported (Anderson *et al.* 2011b). The cellular metagenomes of other environments are significantly enriched in genes related to energy metabolism (Figure 4.7), which is generally expected as viruses are not known to metabolize independently of their host. However, in the case of the hydrothermal vent viral metagenome, the opposite is true: energy metabolism-related genes were significantly enriched in the viral fraction. NiFe hydrogenases, which are required for H_2 metabolism and constitute an important source of energy in hydrothermal systems, provided a key example of this trend. In the hydrothermal vent cellular metagenome, 107 reads matched a NiFe hydrogenase for every 100 Mbp in the metagenome, whereas the viral metagenome subset contained 163. One explanation is that these genes are auxiliary metabolic genes (AMDs), selected for retention in the viral gene pool to support the host during the course of infection, in analogy to photosynthesis genes encoded in cyanophage expressed during viral infection (Clokie & Mann 2006; Lindell *et al.* 2005; Sharon *et al.* 2009). Modeling work has indicated that these photosynthetic genes can enhance host fitness (Bragg and Chisholm, 2008; Hellweger, 2009). Energy metabolism genes in the vent viral fraction could be genes encoded by GTAs, or they could be viral genes expressed while integrated as prophage to boost host fitness by providing their hosts with new or supplemental means of surviving a challenging, dynamic environment.

Does selection operate differentially on viral and cellular genes?

An important question is how selection shapes the viral and cellular gene pools, and which genes are subject to stronger or weaker selection. Differing life strategies for cells and viruses, as well as disparate roles for functional genes within each of the respective gene pools, should leave different selective signatures on genes within each of these gene pools. To measure differential selection among viral and cellular genes, we calculated dN/dS ratios of genes encoded by the viral and cellular fractions. The challenge of calculating dN/dS ratios with shotgun metagenomic data is that the short sequences make it difficult to align long blocks of sequences to the same region of a gene. To circumvent this problem, we used the method developed by Tai *et al.* (2011) to calculate dN/dS ratios from metagenomic reads, in which sequencing reads are mapped to the genomes of previously sequenced isolates. Since the sequences used for analysis were likely derived from multiple taxa, and we do not know the specific phylogenetic relationship of these sequences to each other, this method cannot determine which polymorphisms have become "fixed" in the population, and instead provides an indicator of diversification within the environmental gene pool. Both metagenomes were mapped to pre-existing hydrothermal vent isolates as a high-throughput means to align reads to many genes at once. The mapping analyses and calculation of dN/dS ratios relied upon genomes from *Nautilia profundicola* AmH, *Thermococcus kodakarensis* KOD1, *Thermococcus onnurineus* NA1, *Caminibacter mediatlanticus* TB-2 (contig ABCJ01000001), and *Nitratiruptor* sp. SB155-2, which represent abundant strains in the vent environment. We attempted to map the virome to several existing viral genomes from various environments, but none exhibited sufficient depth of coverage to calculate dN/dS ratios, indicating that the genes encoded by viruses from this hydrothermal system are vastly different from those sequenced previously.

Overall, the cellular metagenome mapped to 831 bacterial and 32 archaeal genes, with an average dN/dS of 0.22. This result indicates that genes encoded by cells in the vent environment are subject to purifying selection. The viral metagenome mapped to 85 bacterial and 106 archaeal genes, with an average dN/dS of 0.15; the viral metagenome subset mapped to 39 bacterial and 25 archaeal genes, with an average dN/dS of 0.13 (Figure 4.8). These dN/dS values are significantly lower than the dN/dS of genes matching the cellular metagenome, within a confidence interval of 95%. This pattern was consistent for each of the genomes mapped (Fig. 4.9). The viral and cellular metagenomes mapped to different genes in each of the strains listed above, and so slightly different sets of genes were used to make this calculation. However, the difference in overall dN/dS is not due solely to differences in the types of genes to which each metagenome mapped: when we examined the dN/dS for only the genes to which both metagenomes mapped, the calculated dN/dS was consistently lower for the viral fraction compared to the cellular fraction (Figure 4.10). No clear patterns emerged when examining the dN/dS for different gene categories (Figure 4.11), except that hypothetical proteins matching the cellular metagenome had a slightly higher dN/dS than other gene types, consistent with previous findings (Tai *et al.*, 2011). This exception was not the case for hypothetical proteins matching the viral fraction. The viral hypothetical proteins may be prophage that are more weakly selected in the cellular fraction, but strongly selected in the viral fraction.

These results indicate that both the viral and cellular fractions are subject to purifying selection, but that the viral gene pool is under stronger purifying selection than the cellular gene pool. One possible explanation for the overall difference in dN/dS between the viral and cellular gene pools is that very little variation in viral genes is permitted. In this scenario, viral genes are under such strong selection that deviations from the consensus protein sequence produce enough of a fitness difference to eliminate the viral mutant. An alternative explanation posits that these viral genes undergo strong positive selection, resulting in selective sweeps of the viral population that eliminate variation from these sites (Kryazhimskiy and Plotkin, 2008). This scenario becomes more complicated as a result of the relationship between virus and host. If a virus were primarily lytic, then a selective sweep could act directly on the viral particles, reducing phenotypic variation (and therefore nonsynonymous polymorphisms). If, however, a virus were primarily lysogenic, a larger proportion of time would be spent integrated in the genome of the host. In this situation, the selective sweeps would act on both host and virus, and we might not expect to see a significant difference in the dN/dS ratios between viral and cellular genes. The difference observed here may indicate that selection acts most strongly on viruses while they are in the process of replicating or when they are in virion form (free in the environment), and can therefore be selected separately from the host.

The 80% identity cutoff we used should ensure that mapped reads were derived from the same population as the reference genome. Therefore, most of the genes mapped by the viral fraction must correspond to genes originally derived from the cellular fraction. It is unlikely that these genes were all derived from cellular contamination of the viral fraction, as that scenario would have resulted in identical dN/dS values. Instead, these genes are likely to be cellular genes that were horizontally transferred to a viral genome. If so, then the lower dN/dS ratio may indicate that these genes are under stronger selection in these small viral genomes, where there is little room for redundancy and extra genetic material is likely to be retained only if it provides a distinct fitness advantage. Therefore, these genes may encode the fitness factors maintained on prophage genomes as a means to enhance host fitness.

Conclusions

The dynamic, recirculating conditions of hydrothermal vent systems, combined with the vast diversity of archaea, bacteria, and their accompanying viruses, create an ecosystem with high potential for widespread sharing of the communal gene pool. Mobile elements were abundant in both the viral and cellular metagenomes we obtained, likely reflecting the abundance of lysogenic viruses, which require integrases for prophage genome insertion, as well as high potential for horizontal gene transfer. In the genomes of vent inhabitants, selection appears to have favored horizontal transfer of genes for environmental interaction and energy metabolism. Our results show that selection favored maintenance of genes related to energy metabolism in the viral gene pool, despite the fact that viruses themselves are not known to metabolize; any genes found in the viral gene pool must have some utility in order to be retained on small viral genomes. The abundance of lysogenic viruses and the strong selective signatures suggest that viruses in vents are selected to spend much of their time as integrated prophage rather than as free virions; the abundance of energy metabolism genes in the viral fraction suggests that viruses benefit their hosts while integrated as prophage. Prophage are selected to boost host fitness while the fitness of the virus and host are intertwined. In turn, host cells benefiting from prophage-encoded genes may gain an adaptive advantage through enhanced metabolic flexibility, which may favor selection for cells harboring prophage. This advantage complicates the symbiotic relationship between virus and host. While still capable of wreaking destruction upon the cells they infect, the viral evolutionary strategy in vents may occasionally transcend from a parasitic relationship into a mutualistic one, as both host and virus seek to survive the dynamic, extreme environment in which they coexist.

Table 4.1. List of viruses used to train PhylophythiaS for distinguishing between archaeal viruses and bacterial viruses.

Archaeal viruses	**Bacterial viruses**
Acidianus bottle-shaped virus	*Acinetobacter* phage 133
Acidianus spindle-shaped virus 1	Bacteriophage 11b
Haloarcula hispanica pleomorphic virus 1	Bacteriophage Aeh1
Halorubrum phage HF2	Bacteriophage T3 complete genome strain Luria
His1 virus	*Burkholderia* phage phi644-2
His2 virus	*Campylobacter* phage CP220
Hyperthermophilic Archaeal Virus 1	Deep-sea thermophilic phage D6E
Hyperthermophilic Archaeal Virus 2	*Methanothermobacter* prophage psiM100
Pyrococcus abyssi virus 1	*Ostreococcus tauri* virus 1
Sulfolobus islandicus rod-shaped virus 2	*Prochlorococcus* phage P-SSM2
Sulfolobus islandicus rudivirus 1 variant XX	*Pseudomonas* phage 201phi2-1
Sulfolobus turreted icosahedral virus 2	*Pseudomonas* phage gh-1
Sulfolobus virus Kamchatka 1	*Synechococcus* phage S-PM2
Thermoproteus tenax spherical virus 1	*Thermus* phage IN93
	Thermus phage P23-45
	Thermus phage P23-77
	Thermus phage P74-26
	Thermus phage phiYS40
	Vibrio phage ICP1_2004_A
	Vibriophage VP4

Table 4.2. Pfam domains included in search for genes associated with mobile genetic elements.

Pfam Accession	Description
PF00552	Integrase DNA binding domain
PF00665	rve Integrase
PF01609	Transposase DDE domain
PF01797	Transposase IS200 like
PF09299	Mu transposase C-terminal
PF00154	recA bacterial DNA recombination protein
PF00239	Resolvase
PF00589	phage integrase
PF00872	Transposase, Mutator family
PF01076	Plasmid recombination enzyme
PF01385	Probable transposase
PF01526	Tn3 transposase DDE domain
PF01527	Helix-turn-helix transposase
PF01548	Transposase
PF01610	Transposase
PF01710	Transposase helix-turn-helix
PF02022	integrase zinc binding domain
PF02281	Transposase Tn5 dimerisation domain
PF02316	Mu DNA-binding domain
PF02371	Transposase IS116/IS110/IS902 family
PF02534	Type IV secretory system conjugative DNA transfer
PF02646	RmuC family- DN recombination proteins
PF02899	phage integrase, N-terminal SAM-like domain
PF02914	Bacteriophage Mu transposase
PF02920	DNA binding domain of tn916 integrase
PF02945	Recombination endonucelase VII
PF03050	Transposase IS66 family
PF03400	IS1 transposase
PF03837	RecT family, involved in recombination
PF03838	Recombination protein U
PF03930	Recombinase Flp protein N-terminal domain
PF04404	ERF superfamily-- recombination proteins
PF04693	Archaeal putative transposase ISC1217
PF04740	LXG domain of WXG superfamily- not sure why this was included.
PF04754	Putative transposase, YhgA-like
PF04986	Putative transposase
PF05202	Recombinase Flp protein
PF05598	transposase domain
PF05717	IS66 Orf2 like protein (essential for transposition)
PF07508	Recombinase
PF07592	Rhodopiruellula transposase DDE domain
PF08423	Rad51-- DNA repair and recombination protein
PF09003	bacteriophage lambda integrase, N-terminal domain
PF09034	Excisionase from transposon Tn916
PF09124	T4 recombination endonuclease VII, dimerisation

PF09588	YqaJ-like viral recombinase domain
PF10136	Site-specific recombinase
PF10551	MULE transposase domain
PF12834	Phage integrase, N-terminal
PF12835	Integrase_1
PF12940	Recombination-activation protein 1 (RAG1)
PF13009	Putative phage integrase
PF13408	Recombinase zinc beta ribbon
PF13495	Phage integrase, N-terminal SAM-like domain
PF13542	Helix-turn-helix domain of transposase family ISL3
PF13683	Integrase core domain
PF13751	Transposase DDE domain

Table 4.3. Percent of reads in cellular metagenomes matching a protein in the "Prophage" grouping of the ACLAME database. Matches were found using tblastn with a minimum e-value of 10^{-5}. All metagenomes listed here were generated with shotgun pyrosequencing.

Metagenome	**Reads**	**ACLAME prophage hits**	**Percent reads**	**Biome**	**Sampling details**	**Reference**	**Accession number**
Monterey Bay	192162	9759	5.08	Open ocean	Monterey Bay, California, surface waters, October	--	4443713.3
Hulk hydrothermal vent	*808051*	*32595*	*4.03*	*Hydrothermal vent*	*Juan de Fuca Ridge, Northeast Pacific Ocean, 2198m depth*	*This study*	4481541.3
Glacial ice	1076539	40695	3.78	Glacial ice	Glacial ice of the Northern Schneeferner, Germany	Simon *et al.*, 2009	CAM_PROJ_IceMetagenome
North Atlantic Spring Bloom	257471	6913	2.68	Open ocean	Bermuda Atlantic Time-Series site	--	4443725.3
Human oral microbiota	339503	5596	1.65	Human	Dental plaque from 25 human volunteers	Belda-Ferre *et al.*, 2012	4447970.3
Healthy fish microbiota	51498	610	1.18	Fish	Healthy aquacultured fish, San Diego, CA	Angly *et al.*, 2009	4440055.3
Guaymas Basin	4970673	58638	1.18	Hydrothermal vent	Hydrothermal plumes from Guaymas Basin, CA	Baker *et al.* , 2012	CAM_P_0000545
Cow rumen	320471	3678	1.15	Cow	Fiber-adherent microbiome from cow rumen	Brulc *et al.*, 2009	4441681.3
Healthy fish microbiota	60580	541	0.89	Fish	Samples from aquacultured fish gut contents	Angly *et al.*, 2009	4440059.3
Medium salinity	108725	742	0.68	Salt water	Salinity 12-14% from	Rodriguez-Brito *et*	4440425.3

saltern					solar salterns, California	*al.*, 2010	
Salton Sea	161912	992	0.61	Sediments	Sulfidic, anoxic sediments of the Salton Sea	Swan *et al.*, 2010	4440329.3
Tilapia fish pond	344260	1712	0.50	Fresh water	Water samples from aquaculture facility raising striped bass	Rodriguez-Brito *et al.*, 2010	4440440.3
Peru Margin 1mbsf	100093	489	0.489	Deep biosphere-marine sediment	Peru Margin ODP Leg 201 Site 1229, 1 meter below seafloor	Biddle *et al.*, 2008	4440961.3
Peru Margin 50mbsf	63258	288	0.455	Deep biosphere-marine sediment	Peru Margin ODP Leg 201 Site 1229, 50 meters below seafloor	Biddle *et al.*, 2008	4459941.3
Peru Margin 32mbsf	135429	479	0.354	Deep biosphere-marine sediment	Peru Margin ODP Leg 201 Site 1229, 32 meters below seafloor	Biddle *et al.*, 2008	4459940.3
Low salinity saltern	31948	111	0.35	Salt water	Salinity 6-8% from solar salterns, California	Rodriguez-Brito *et al.*, 2010	4440426.3
Microbialites	257573	802	0.311	Micro-bialites	Highborne Cay, Bahamas	Desnues *et al.* 2008, Dinsdale *et al.*, 2008	4440061.3
High salinity saltern	33356	98	0.29	Salt water	Salinity 27-30% from solar salterns, California	Rodriguez-Brito *et al.*, 2010	4440419.3
Peru Margin 16mbsf	121414	191	0.157	Deep biosphere-marine sediment	Peru Margin ODP Leg 201 Site 1229, 16	Biddle *et al.*, 2008	4440973.3

					meters below seafloor		
Porites compressa coral	1053275	947	0.0900	Coral	Samples collected at the Hawaii Institute for Marine Biology	Vega Thurber *et al.*, 2009	CAM_PROJ_ CoralMetagenome
Soudan Mine	248038	193	0.0778	Deep biosphere-terrestrial mine	Water and sediments in mine, 714 m below surface, Soudan Mine, MN	Edwards *et al.*, 2006	4440282.3
Line Islands	178628	120	0.0672	Seawater	Water sampled near coral reefs, Christmas Island	Dinsdale *et al.*, 2008	4440041.3

Table 4.4. Percent of reads in cellular and viral metagenomes matching a mobile element. These include transposases, integrases, recombinases, and resolvases as defined by a keyword search in Pfam (database file included in supplementary material). Matches found using tblastn with a minimum e-value of 10^{-5}. All metagenomes listed here were generated with shotgun pyrosequencing.

Meta-genomes	**Cellular or viral**	**Reads**	**Mobile elements**	**% reads**	**Biome**	**Sampling details**	**Ref**	**Accession number**
Glacial ice	cellular	1076539	5598	0.52	Glacial ice	Glacial ice of the Northern Schnee-ferner, Germany	Simon *et al.*, 2009	CAM_PROJ_IceMetagenome
Hydro-thermal vent	**viral subset**	**64599**	252	0.39	*Hydro-thermal vent*	*Juan de Fuca Ridge, Northeast Pacific Ocean, 2198m depth*	*This study*	--
Healthy fish microbiota	cellular	51498	193	0.37	Fish	Healthy aquacultured fish, San Diego, CA	Dinsdale *et al.*, 2008	4440055.3
Healthy fish microbiota	cellular	60580	191	0.32	Fish	Samples from aquacultured fish gut contents	Angly *et al.*, 2009	4440059.3
Hydro-thermal vent	**cellular**	**808051**	2539	0.31	*Hydro-thermal vent*	*Juan de Fuca Ridge, Northeast Pacific Ocean, 2198m depth*	*This study*	4481541.3
Cow rumen	cellular	320471	976	0.30	Cow	Fiber-adherent microbiome from cow rumen	Brulc *et al.*, 2009	4441681.3
Hydro-thermal vent	**viral**	**231246**	579	0.25	*Hydro-thermal vent*	*Juan de Fuca Ridge, Northeast Pacific Ocean, 2198m depth*	*Anderson et al., 2011a*	*448187.3*
Antarctic Lake summer	viral	30515	66	0.22	Fresh water	Freshwater oligotrophic lake, Byers Peninsula, Antarctica (summer)	Lopez-Bueno *et al.*, 2009	4441558.3
Reclaimed water	viral	1531954	2330	0.15	Fresh water	Viral fraction of reclaimed water	Rosario *et al.*, 2009	CAM_PROJ_ReclaimedWaterViruses
Monterey Bay	cellular	192162	278	0.14	Seawater	Monterey Bay, California, October 2000, surface waters	-	4443713.3
Human oral	cellular	339503	450	0.13	Human	Dental plaque	Belda-	4447970.3

microbiota						from 25 human volunteers	Ferre *et al.*, 2012	
Healthy fish microbiota	viral	55690	66	0.12	Fish	Samples from aquacultured fish gut contents	Angly *et al.*, 2009	4440065.3
Tilapia Pond	cellular	344260	352	0.10	Fresh water	Water samples from aquaculture facility raising striped bass	Rodriguez-Brito *et al.*, 2010	4440440.3
High salinity saltern	cellular	33356	33	0.10	Salt water	Salinity 27-30% from solar salterns, California	Rodriguez-Brito *et al.*, 2010	4440419.3
Guaymas Basin	cellular	4970673	4728	0.10	Hydrothermal vent	Hydrothermal plumes from Guaymas Basin, CA	Baker *et al.*, 2012	CAM_P_0000545
Arctic Ocean	viral	688590	605	0.09	Seawater	10-3246m, Fall 2002, Arctic Ocean	Angly *et al.*, 2006	4441621.3
Medium salinity saltern	cellular	108725	95	0.09	Salt water	Salinity 12-14% from solar salterns, California	Rodriguez-Brito *et al.*, 2010	4440425.3
Bay of British Columbia	viral	138347	107	0.08	Seawater	0-245m, sampled over several dates, Bay of British Columbia	Angly *et al.*, 2006	4441623.3
North Atlantic Spring Bloom	cellular	257471	193	0.07	Seawater	Bermuda Atlantic Time-Series site	--	4443725.3
Peru Margin 1mbsf	cellular	100093	69	0.07	Deep biosphere-marine sediment	Peru Margin ODP Leg 201 Site 1229, 1 meter below seafloor	Biddle *et al.*, 2008	4440961.3
Salton Sea	cellular	161912	98	0.06	Sediments	Sulfidic, anoxic sediments of the Salton Sea	Swan *et al.*, 2010	4440329.3
Gulf of Mexico	viral	263908	153	0.06	Seawater	0-164m, sampled over several dates, Gulf of Mexico	Angly *et al.*, 2006	4441625.3
Microbialites	viral	621110	359	0.06	Microbialites	Pozas Azules, Mexico; Rio Mesquites, Mexico; Highborne Cay, Bahamas	Desnues *et al.*, 2008	4440320.3 4440321.3 4440323.3
Peru Margin 50mbsf	cellular	63258	28	0.04	Deep biosphere-marine sediment	Peru Margin ODP Leg 201 Site 1229, 50 meters below seafloor	Biddle *et al.*, 2008	4459941.3

Soudan Mine	cellular	248038	105	0.04	Deep biosphere-terrestrial mine	Water and sediments in mine, 714 m below surface, Soudan Mine, MN	Edwards *et al.*, 2006	4440282.3
Antarctic Lake spring	viral	31691	13	0.04	Fresh water	Freshwater oligotrophic lake, Byers Peninsula, Antarctica (spring)	Lopez-Bueno *et al.*, 2009	4441778.3
Coral	viral	36354	14	0.04	Coral	*Porites compressa* coral samples collected at the Hawaii Institute for Marine Biology	Vega Thurber *et al.*, 2009	4440374.3
Low salinity saltern	viral	56810	21	0.04	Salt water	Salinity 6-8% from solar salterns, California	Rodriguez-Brito *et al.*, 2010	4440420.3
Peru Margin 32mbsf	cellular	135429	49	0.04	Deep biosphere-marine sediment	Peru Margin ODP Leg 201 Site 1229, 32 meters below seafloor	Biddle *et al.*, 2008	4459940.3
Salton Sea	viral	27689	7	0.03	Sediments	Sulfidic, anoxic sediments of the Salton Sea	Swan *et al.*, 2010	4440328.3
Tilapia Pond	viral	231521	48	0.02	Fresh water	Water samples from aquaculture facility raising striped bass	Rodriguez-Brito *et al.*, 2010	4440439.3
Low salinity saltern	cellular	31948	6	0.02	Salt water	Salinity 6-8% from solar salterns, California	Rodriguez-Brito *et al.*, 2010	4440426.3
Medium salinity saltern	viral	33291	6	0.02	Salt water	Salinity 12-14% from solar salterns, California	Rodriguez-Brito *et al.*, 2010	4440427.3
Peru Margin 16mbsf	cellular	121414	18	0.01	Deep biosphere-marine sediment	Peru Margin ODP Leg 201 Site 1229, 16 meters below seafloor	Biddle *et al.*, 2008	4440973.3
Coral	cellular	1053275	144	0.01	Coral	*Porites compressa* coral samples collected at the Hawaii Institute for Marine Biology	Vega Thurber *et al.*, 2009	CAM_PROJ_CoralMetagenome
Tampa Bay	viral	257075	32	0.01	Fresh	Prophages	McDaniel	4440102.3

					wataer	induced with mitomycin C from Tampa Bay water samples	*et al.*, 2008	
High salinity saltern	viral	136564	13	0.01	Salt water	Salinity 27-30% from solar salterns, California	Rodriguez-Brito *et al.*, 2010	4440421.3
Microbialites	cellular	257573	20	0.01	Micro-bialites	Highborne Cay, Bahamas	Breitbart *et al.*, 2009	4440061.3
Sargasso Sea	viral	399343	22	0.01	Open ocean	80m, sampled June 2005, Sargasso Sea	Angly *et al.*, 2006	4441624.3

Table 4.5. Numbers of prophage identified in hydrothermal vent bacterial and archaeal genomes using Prophage Finder.

Organism	**Number of predicted prophage**
Aciduliprofundum boonei T469	2
Archaeoglobus profundus DSM 5631	2
Archaeoglobus fulgidus DSM 4304	1
Aquifex aeolicus VF5	1
Caminibacter mediatlanticus TB2	0
Deferribacter desulfuricans SSM	1
Ferroglobus placidus DSM 10642	1
Hyperthermus butylicus DSM 5456	0
Ignicoccus hospitalis KIN4/I	0
Methanocaldococcus sp. FS406-22	0
Methanocaldococcus fervens AG86	0
Methanocaldococcus jannaschii DSM 2661	0
Methanocaldococcus vulcanius M7	0
Methanopyrus kandleri AV19	0
Nanoarchaeum equitans Kin4-M	0
Nautilia profundicola AmH	1
Nitratiruptor sp. SB155-2	4
Persephonella marina EX-H1	4
Pyrobaculum aerophilum str. IM2	0
Pyrococcus abyssi GE5	1
Pyrococcus furiosus DSM 3638	1
Pyrococcus horikoshii OT3	2
Pyrococcus sp NA2	2
Pyrococcus sp. ST04	0
Rhodothermus marinus DSM 4252	1
Staphylothermus marinus F1	0
Thermus thermophilus HB8	2
Thermococcus gammatolerans EJ3	0
Thermococcus kodakarensis KOD1	0
Thermococcus onnurineus NA1	2
Thermococcus sibricus MM_739	1
Thermococcus sp. CL1	2
Thermotoga neapolitana DSM 4359	1
Thiomicrospira crunognea XCL 2	4

Table 4.6. Annotation and best hit of reads within the low-coverage region of fragment recruitment from the Hulk cellular metagenome to the longest contig in the *Caminibacter mediatlanticus* TB-2 draft genome. See methods for details of fragment recruitment. Annotations are as listed by the draft annotation file released by the JCVI. Organism best hit was determined by using blastn against the nr database.

Basepairs	**Annotation**	**Organism best hit**
358816-386118	ABC-type sugar transport system, permease component	*Nautilia profundicola* AmH
386196-387833	ATPase involved in DNA repair	*Nautilia profundicola* AmH
387830-388561	Chromosome segregation ATPases	*Nautilia profundicola* AmH
3888827-389372	Anaerobic dehydrogenases, typical selenocysteine-containing	*Nautilia profundicola* AmH
389421-391646	Anaerobic dehyrogenases, typical selenocysteine-containing	*Nautilia profundicola* AmH
391657-392250	Fe-S-cluster-containing hydrogenase components I	*Arcobacter nitrofigilis* DSM7299
392228-393193	Cytochrome b subunit of formate dehydrogenase	*Nautilia profundicola* AmH
393295-394212	ABC-type molybdate transport system, periplasmic component	*Nautilia profundicola AmH*
394212-395165	ABC-type sugar transport systems, ATPase components	*Deferribacter desulfuricans* SSM1
395174-395899	ABC-type sulfate transport system, permease component	*Nautilia profundicola* AmH
396139-397005	Formate/nitrite family of transporters	*Desulfurobacterium thermolithotrophicum* DSM 11699
397286-397708	Ni,Fe-hydrogenase maturation factor	No closely related hits
397705-398049	ABC-type cobalt transport system, ATPase component	No closely related hits
398042-398884	Ni,Fe-hydrogenase III small subunit	*Arcobacter nitrofigilis* DSM7299
398881-399420	Formate hydrogenlyase subunit 6/NADH:ubiquinone oxidoreductase	*Arcobacter nitrofigilis* DSM7299
399430-401169	Ni,Fe-hydrogenase III component G	*Arcobacter nitrofigilis* DSM7299
401180-402661	Formate hydrogenlyase subunit 3/Multisubunit Na+/H+ antiporter, MhhD subunit	*Arcobacter nitrofigilis* DSM7299
402665-403279	Hydrogenase 4 membrane component	*Arcobacter nitrofigilis* DSM7299
403320-404246	Formate hydrogenlyase subunit 4	*Arcobacter nitrofigilis*

		DSM7299
404257-406224	Formate hydrogenlyase subunit 3/Multisubunit Na+/H+ antiporter, MhhD subunit	*Arcobacter nitrofigilis* DSM7299
406235-406798	Fe-S-cluster-containing hydrogenase components 2	*Arcobacter nitrofigilis* DSM7299
406801-408564	Anaerobic dehydrogenase, typically selenocysteine-containing	*Desulfurobacterium thermolithotrophicum* DSM 11699
408613-409032	Anaerobic dehydrogenases, typically selenocysteine-containing	*Desulfurobacterium thermolithotrophicum* DSM 11699
409198-409839	camp-binding proteins- catabolite gene activator and regulatory subunit of camp-dependent protein kinases	No close matches.
409895-411214	Selenocysteine synthase [seryl-tRNASer selenium transferase)	*Arcobacter butzleri* RM4018
411211-413034	Selenocysteine-specific translation elongation factor	*Nautilia profundicola* AmH
413355-413939	Predicted redox protein, regulator of disulfide bond formation	*Nautilia profundicola* AmH/*Campylobacter lari* RM2100
413949-415064	Transglutaminase-like enzymes, putative cysteine proteases	*Nautilia profundicola* AmH
415068-415655	Predicted redox protein, regulator of disulfide bond formation	*Nautilia profundicola* AmH
415657-415929	Cation transport ATPase	No closely related hits
415926-416282	Uncharacterized conserved protein involved in intracellular sulfur reduction	*Nautilia profundicola* AmH
416348-417325	Selenophosphate synthase	*Nautilia profundicola* AmH
417332-420496	Predicted phosphohydrolases	*Thermodesulfatator indicus* DSM 15286
High coverage region between genomic islands 1 and 2		
440514-442073	Transcriptional regulator	No close matches
442082-443608	Glycosyltransferases, probably involved in cell wall biogenesis	*Nitratiruptor* sp. SB155-2
443626-444942	Predicted UDP-glucose 6-dehydrogenase	*Nitratiruptor* sp. SB155-2

Table 4.7. List of viral and cellular metagenomes used for functional profiling of viral and cellular metagenomes using the KEGG Orthology database. Metagenomes obtained from the MG-RAST database were first analyzed by Dinsdale *et al.* (2009).

Metagenome name	**Accession number**	**Viral or Cellular**	**Type of biome**
Fish slime	4440059.3	Cellular	Fish
Fish slime	4440065.3	Viral	Fish
Healthy fish pond	4440413.3	Cellular	Freshwater
Healthy fish pond	4440412.3	Viral	Freshwater
High salinity salterns, west California	4440419.3	Cellular	Salt water
High salinity salterns, west California	4440145.4	Viral	Salt water
High salinity salterns, west California	4440144.4	Viral	Salt water
High salinity salterns, west California	4440421.3	Viral	Salt water
Highborne Cay	4440061.3	Cellular	Salt water
Highborne Cay	4440323.3	Viral	Salt water
Line Islands, Christmas Island	4440041.3	Cellular	Seawater
Line Islands, Christmas Island	4440038.3	Viral	Seawater
Line Islands, Kingman Island	4440037.3	Cellular	Seawater
Line Islands, Kingman Island	4440036.3	Viral	Seawater
Line Islands, Palmyra Island	4440039.3	Cellular	Seawater
Line Islands, Palmyra Island	4440040.3	Viral	Seawater
Line Islands, Tabuaeran	4440279.3	Cellular	Seawater
Line Islands, Tabuaeran	4440280.3	Viral	Seawater
Low salinity salterns, San Diego	4440437.3	Cellular	Salt water
Low salinity salterns, San Diego	4440436.3	Viral	Salt water
Low salinity salterns, San Diego	4440432.3	Viral	Salt water
Low salinity salterns, west California	4440426.3	Cellular	Salt water
Low salinity salterns, west California	4440420.3	Viral	Salt water
Medium salinity salterns, San Diego	4440434.3	Cellular	Salt water
Medium salinity salterns, San Diego	4440435.3	Cellular	Salt water
Medium salinity salterns, west California	4440416.3	Cellular	Salt water
Medium salinity salterns, west California	4440425.3	Cellular	Salt water
Medium salinity salterns, west California	4440428.3	Viral	Salt water
Medium salinity salterns, west California	4440431.3	Viral	Salt water
Medium salinity salterns, west California	4440417.3	Viral	Salt water
Medium salinity salterns, west California	4440427.3	Viral	Salt water
Porites compressa coral	4440378.3	Cellular	Coral
Porites compressa coral	4440374.3	Viral	Coral

Pozas Azules	4440067.3	Cellular	Microbialite
Pozas Azules	4440320.3	Viral	Microbialite
Rio Mesquites	4440060.4	Cellular	Microbialite
Rio Mesquites	4440321.3	Viral	Microbialite
Salton Sea	4440329.3	Cellular	Sediments
Salton Sea	4440327.3	Viral	Sediments
Salton Sea	4440328.3	Viral	Sediments
Tilapia Pond	4440422.3	Cellular	Freshwater
Tilapia Pond	4440440.3	Cellular	Freshwater
Tilapia Pond	4440424.3	Viral	Freshwater
Tilapia Pond	4440439.3	Viral	Freshwater

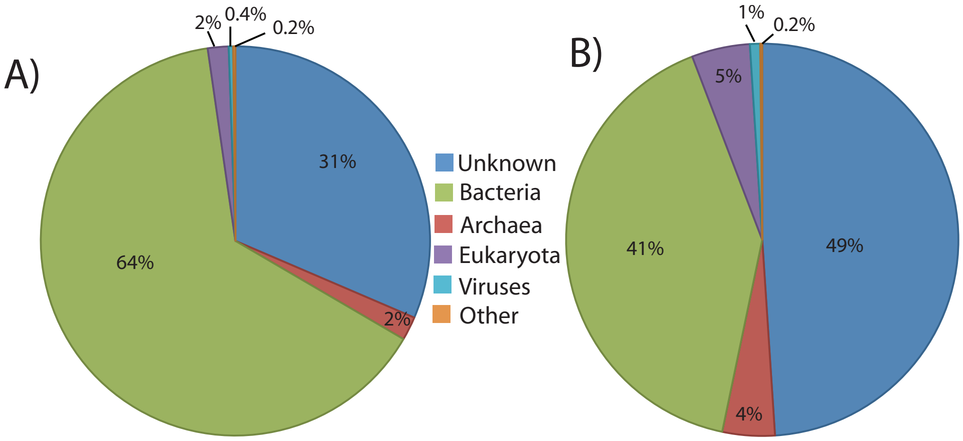


Figure 4.1. Pie charts showing breakdown of read classification as categorized by MG-RAST for the cellular metagenome (A) and the viral metagenome (B).

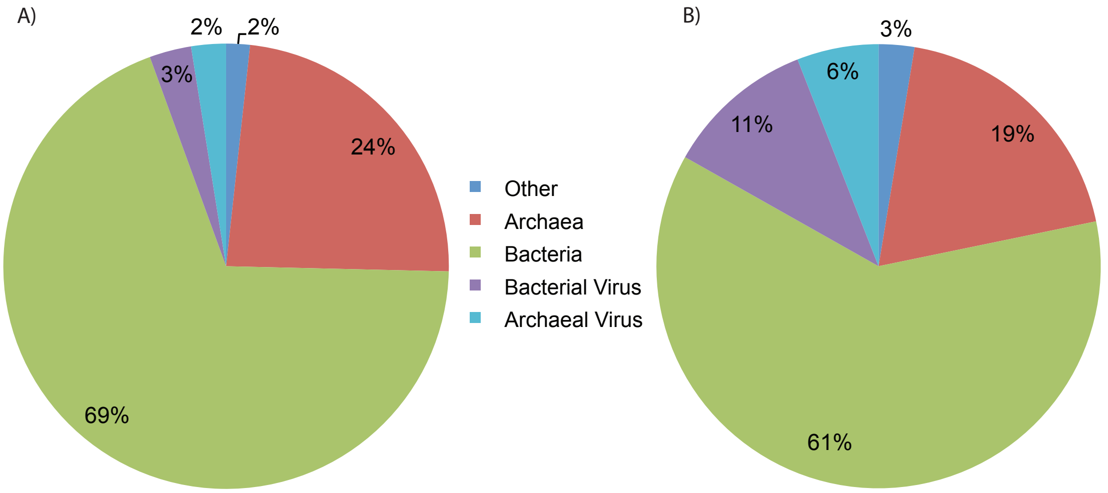


Figure 4.2. Assignment of metagenomic contigs for the cellular metagenome (A) and the viral metagenome (B), based on di-, tri-, and tetranucleotide abundance determined by PhylopythiaS. Boutique PhylopythiaS training datasets were created to classify contigs in the cellular and viral metagenomes as archaeal, bacterial, archaeal virus or bacterial virus.

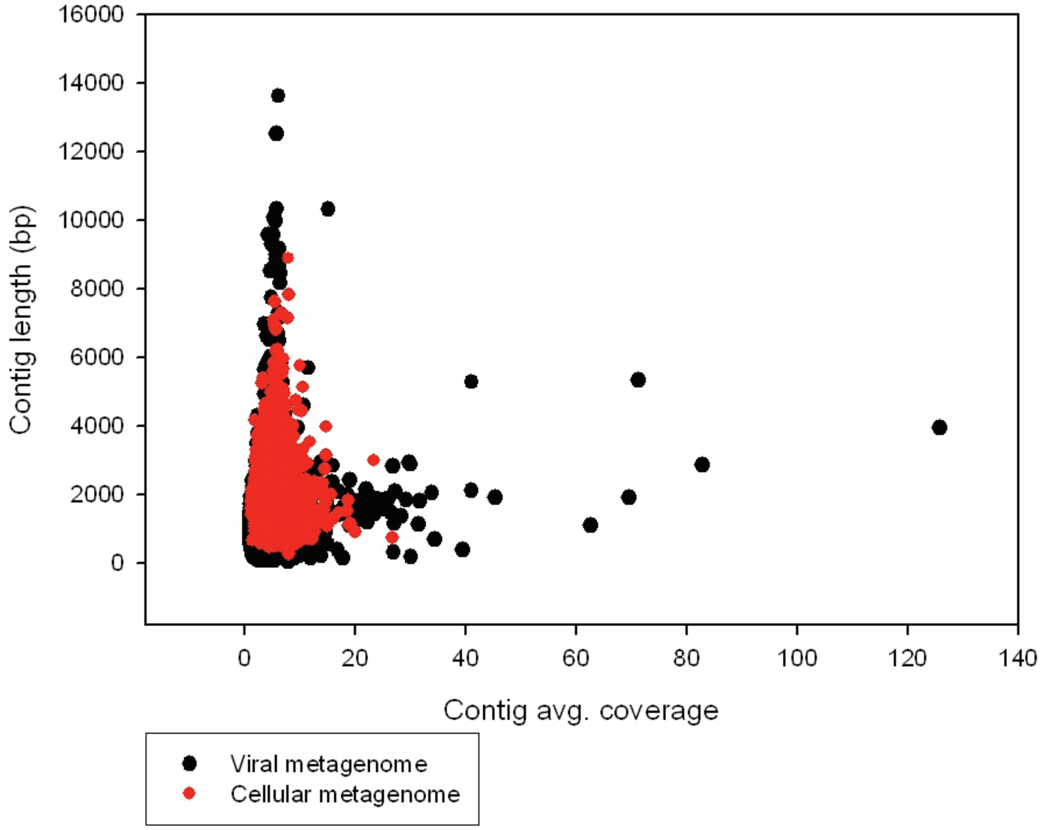


Figure 4.3. Coverage and length of assembled contigs in the viral and cellular metagenomes. Average coverage per base pair across the entire contig is shown on the x-axis; contig length is shown on the y-axis. Cellular metagenome contigs are shown in red; viral metagenome contigs are shown in black.

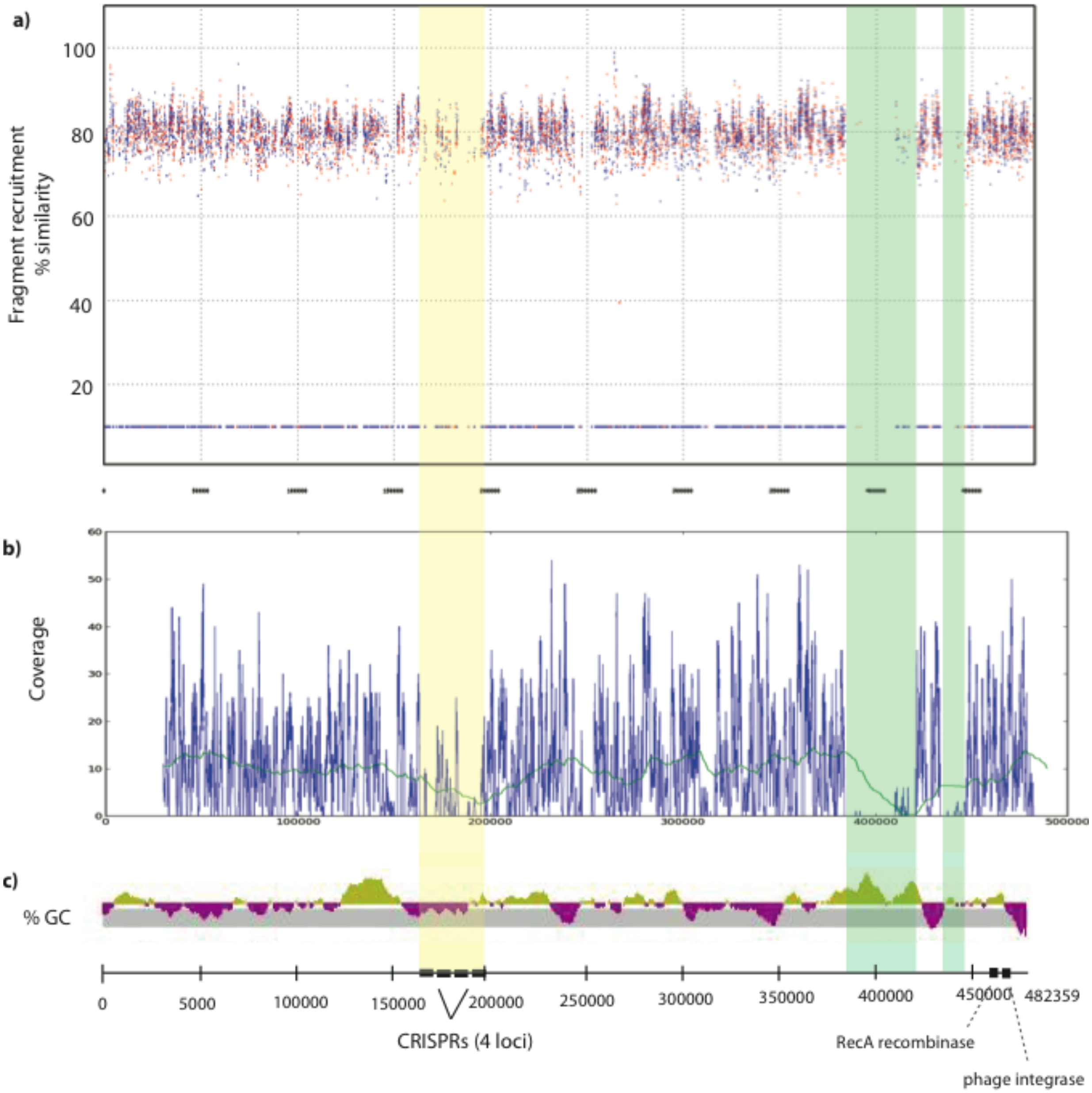


Figure 4.4. Recruitment plot of metagenomic reads to *Caminibacter mediatlanticus* TB-2. Cellular metagenomic reads were mapped to the longest contig of the draft genome of *C. mediatlanticus* TB-2, with percent similarity on the y-axis and base pair numbers on the x-axis (A). Coverage plot of read recruitment is shown per basepair, with blue line showing actual coverage and green line showing a convolution function of the coverage plot using a weighting of 50000 (B). Percent GC plot for the same contig is shown on the same scale, with base pair numbers marked below (C), and are annotated with CRISPR loci and recombinases or integrases found on the contig. Orange shading shows the location of CRISPR loci on the genome; green shading shows the location of two metagenomic islands.

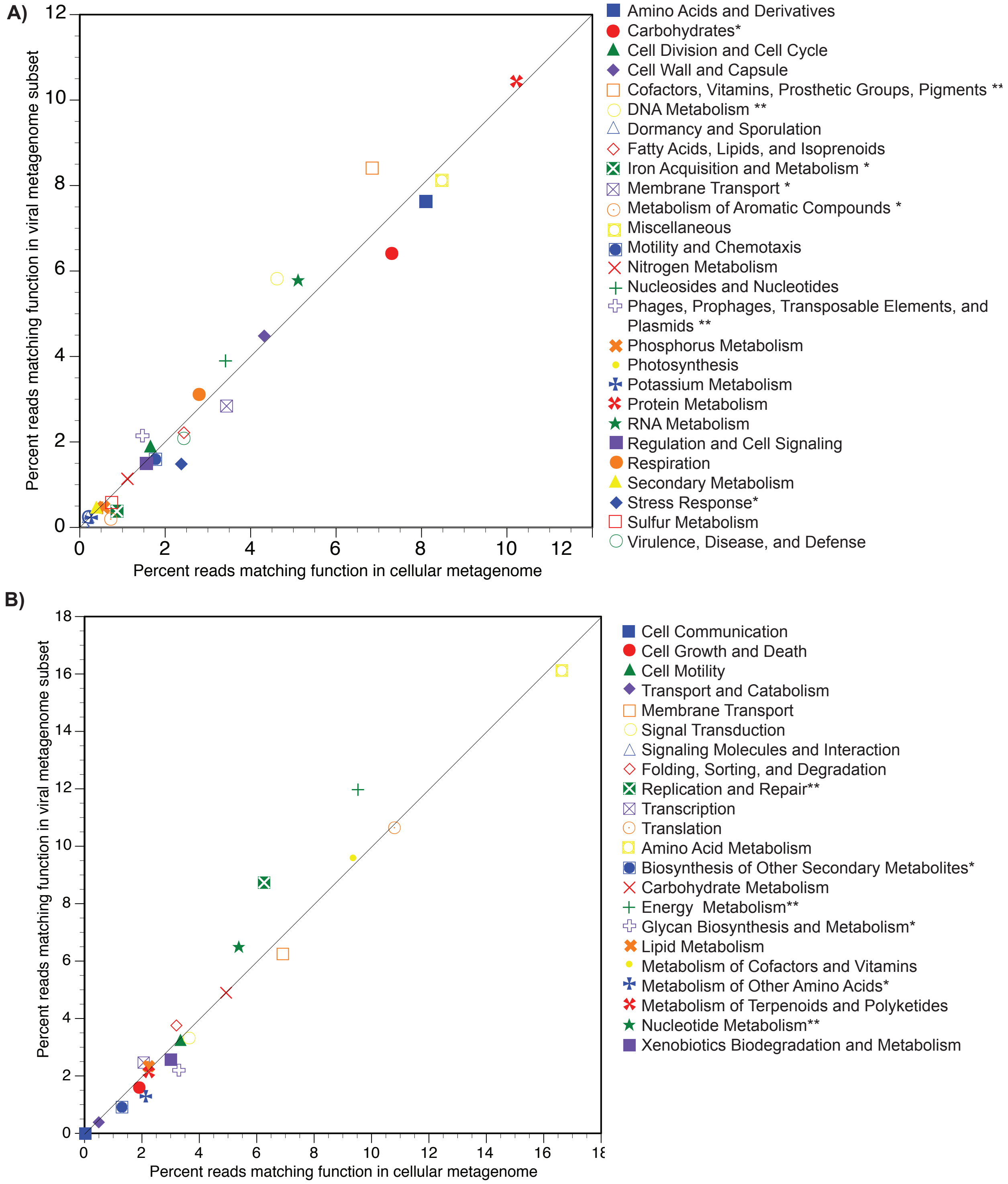


Figure 4.5. Percent of the cellular metagenome and virome subset matching a functional gene category, with cutoff of 1e-05. Annotation used the SEED database (A), where clustering-based subsystems at 15.95% (cellular) and 16.70% (viral) are not shown, and the KEGG Orthology database (B). Lines represent a 1-to-1 ratio. In the legends, one asterisk denotes significant enrichment in the cellular metagenome and two asterisks denote significant enrichment in the virome subset.

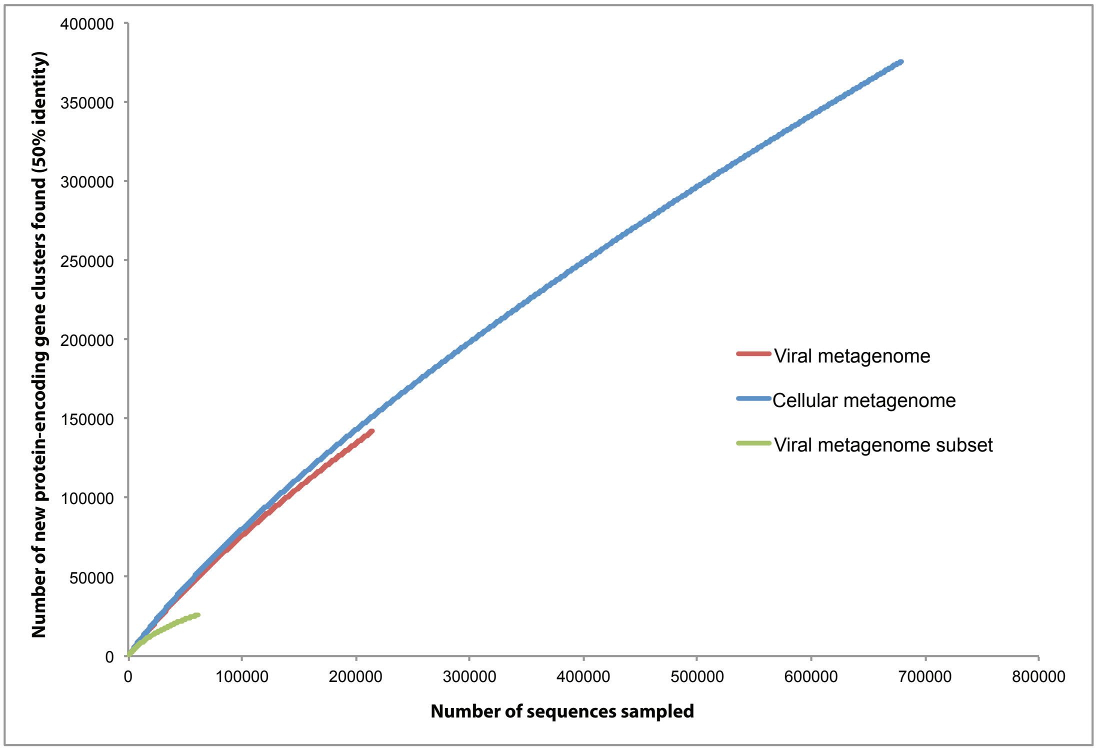


Figure 4.6. Rarefaction curves of metagenomic read clusters in the cellular and viral metagenomes and the viral subset. Open reading frames were identified in reads using FragGeneScan in MG-RAST and clustered at 50% identity in USEARCH using UCLUST.

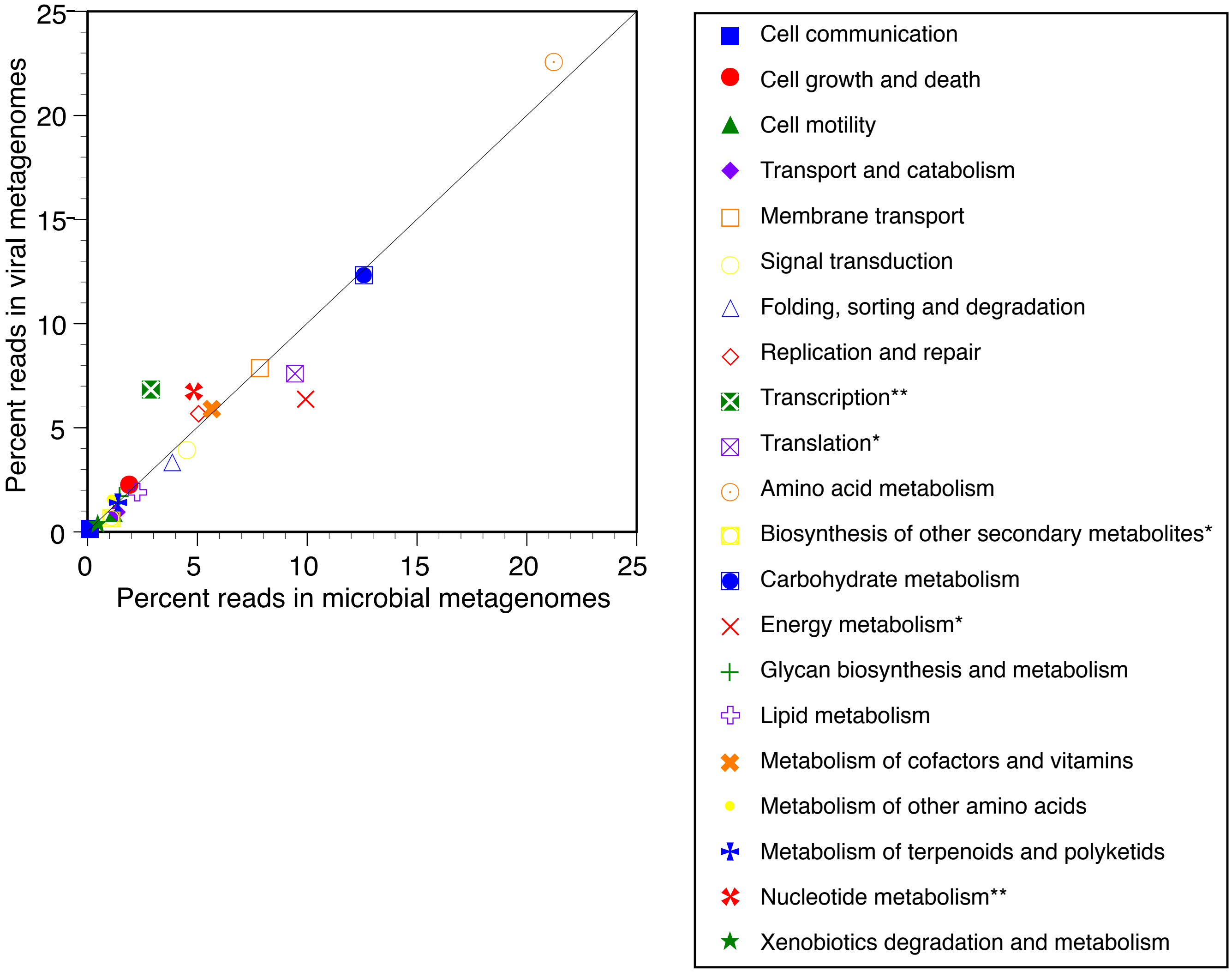


Figure 4.7. Functional comparisons of 20 microbial metagenomes and 23 viral metagenomes according to the KEGG Orthology annotation system. Metagenomes were annotated in MG-RAST with a minimum e-value of 1e-05 and a minimum identity cutoff of 60%; they derive from studies that directly compared viral and microbial metagenomes. Data are from Dinsdale *et al.* (2008); analysis follows Kristensen *et al.* (2009). In the legend, one asterisk denotes significant enrichment in the cellular metagenomes; two asterisks denote significant enrichment in the viral metagenomes, as determined by Xipe-Totec.

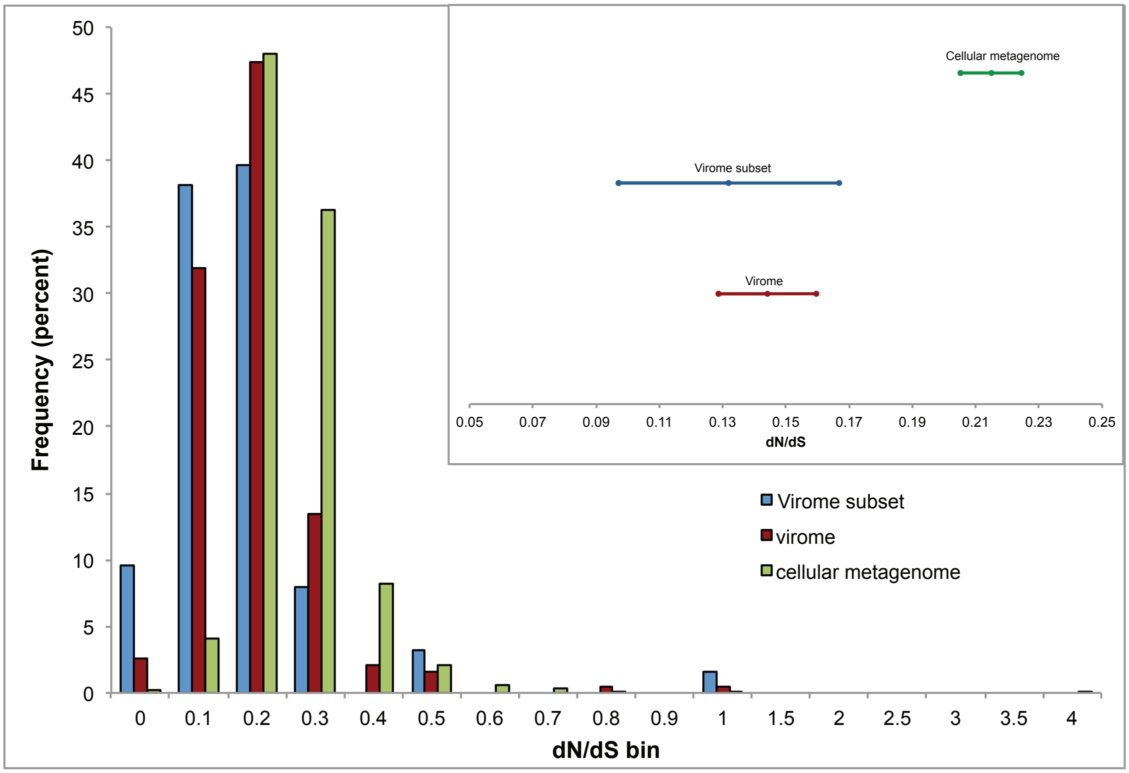


Figure 4.8. Histogram of dN/dS ratios for each metagenome. A total of 863 genes were included for the cellular metagenome calculation, 190 for the viral metagenome and 63 for the viral subset. Values are shown only for genes that had a minimum depth coverage of 5 and minimum nucleotide coverage of 100. Frequency values are normalized by percent. Bins are scaled in increments of 0.1 until 1, and then in increments of 0.5. Inset shows mean and 95% confidence intervals for calculated dN/dS for all three data sets, indicating that the average cellular dN/dS is significantly greater than the average dN/dS for both the virome and the virome subset.

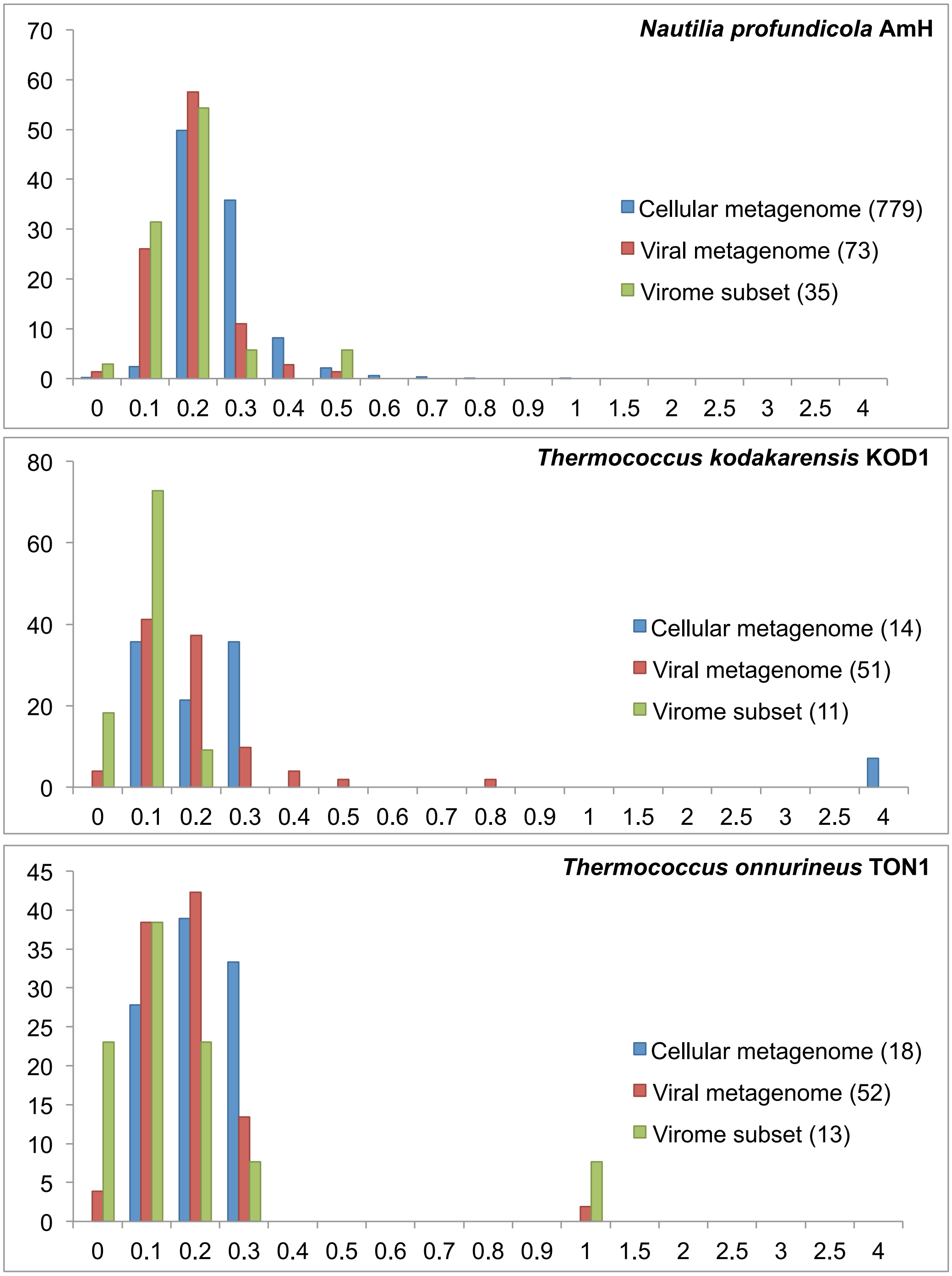


Figure 4.9. Histograms of dN/dS for genes in three different genomes mapped by the cellular metagenome, virome, and virome subset. *Caminibacter mediatlanticus* TB-2 and *Nitratiruptor* sp. SB155-2 are not shown because the virome subset mapped to only four and zero genes in the genome, respectively. Number of genes included in each histogram is indicated in parentheses.

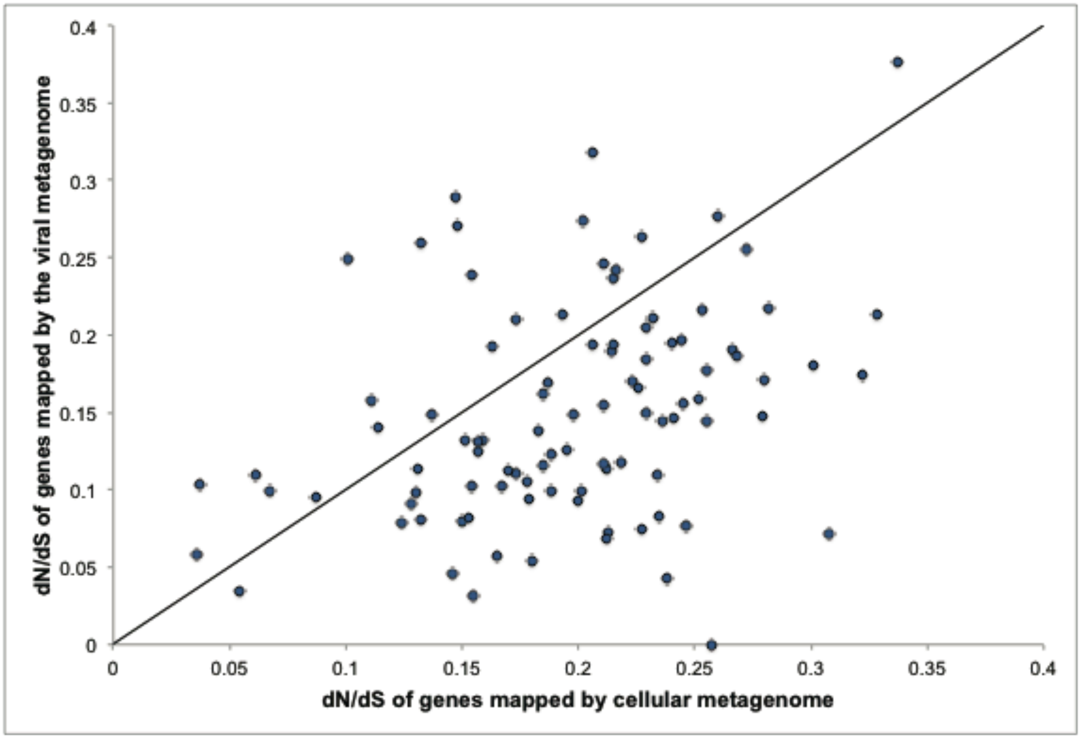


Figure 4.10. Values of dN/dS for genes mapped by the virome versus dN/dS for the same genes mapped by cellular metagenome. The line has a slope of 1.

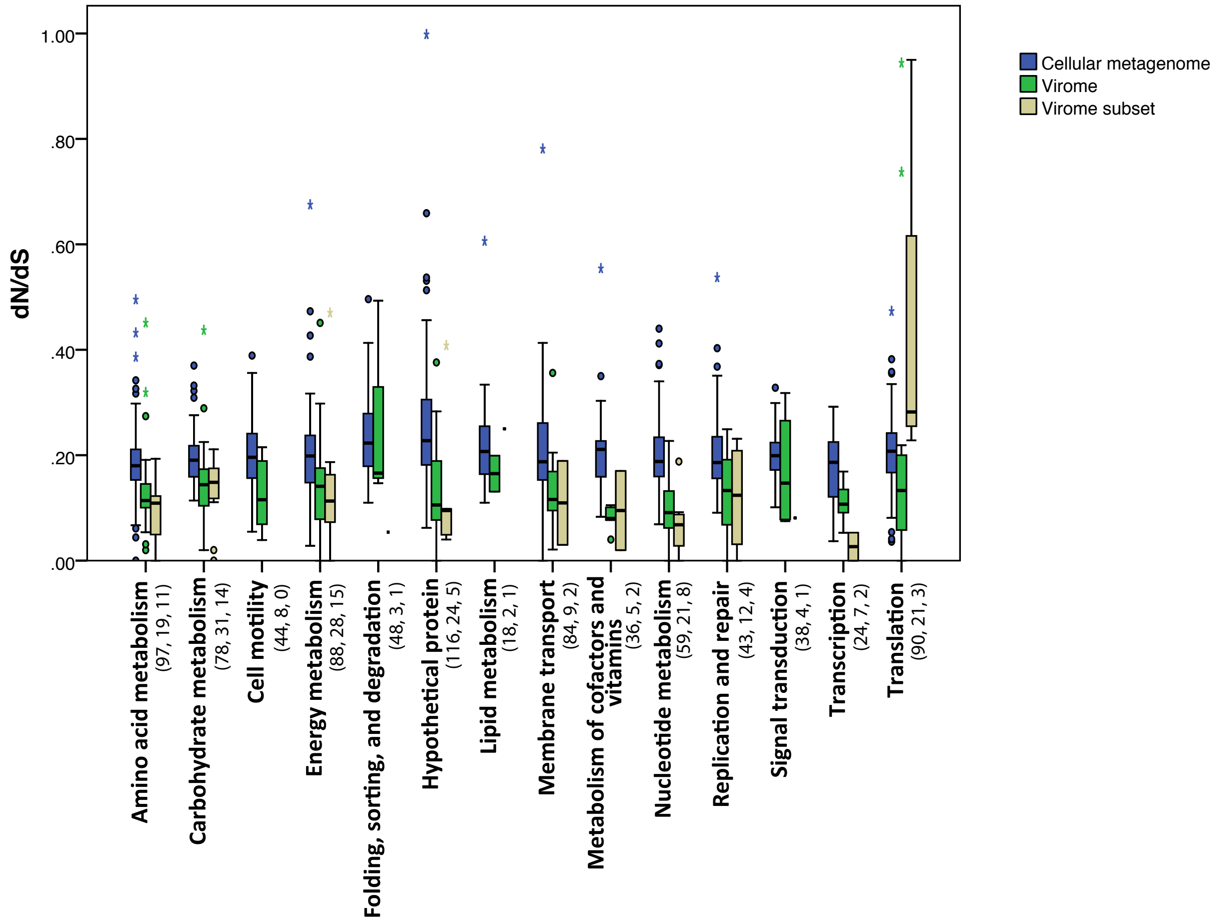


Figure 4.11. Box-and-whisker plots of dN/dS values for genes mapped by the cellular metagenome, the virome and the viral subset, Genes are categorized according to KO categorization. A dotted line indicates where dN/dS = 1 and selection is neutral. Boxes indicate upper and lower quartiles; whiskers denote 1.5 times the interquartile range. Numbers below gene categories indicate the number of genes included for that category for the cellular metagenome, virome, and virome subset, respectively.

References

Anderson, R.E., Beltrán, M.T., Hallam, S.J., and Baross, J.A. (2013) Microbial community structure across fluid gradients in the Juan de Fuca Ridge hydrothermal system. FEMS Microbiol Ecol **83**: 324–339.

Anderson, R.E., Brazelton, W.J., and Baross, J. A. (2011) Is the genetic landscape of the deep subsurface biosphere affected by viruses? Front Extrem Microbiol **2**:

Anderson, R.E., Brazelton, W.J., and Baross, J.A. (2013) The deep viriosphere: Assessing the viral impact on microbial community dynamics in the deep subsurface. Rev Mineral Geochemistry **75**: 649–675.

Anderson, R.E., Brazelton, W.J., and Baross, J. A (2011) Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. FEMS Microbiol Ecol **77**: 120–133.

Baker, B.J., Lesniewski, R.A., and Dick, G.J. (2012) Genome-enabled transcriptomics reveals archaeal populations that drive nitrification in a deep-sea hydrothermal plume. ISME J **6**: 2269–2279.

Baross, J.A. and Hoffman, S.E. (1985) Submarine hydrothermal vents and associated gradient environments as sites for the origin and evolution of life. Orig Life Evol Biosph **15**: 327–345.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., *et al.* (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science (80-) **315**: 1709–1712.

Beiko, R.G., Harlow, T.J., and Ragan, M.A. (2005) Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U S A **102**: 14332 –14337.

Belda-Ferre, P., Alcaraz, L.D., Cabrera-Rubio, R., Romero, H., Simón-Soro, A., Pignatelli, M., and Mira, A. (2012) The oral metagenome in health and disease. ISME J **6**: 46–56.

Biddle, J.F., Fitz-Gibbon, S., Schuster, S.C., Brenchley, J.E., and House, C.H. (2008) Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. Proc Natl Acad Sci U S A **105**: 10583 –10588.

Bose, M. and Barber, R.D. (2006) Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. In Silico Biol **6**: 223–7.

Bragg, J.G. and Chisholm, S.W. (2008) Modeling the fitness consequences of a cyanophage-encoded photosynthesis gene. PLoS One **3**: e3550.

Brazelton, W.J. and Baross, J.A. (2009) Abundant transposases encoded by the metagenome of a hydrothermal chimney biofilm. ISME J **3**: 1420–1424.

Breitbart, M. (2012) Marine Viruses: Truth or Dare. Ann Rev Mar Sci **4**: 425–448.

Breitbart, M., Thompson, L.R., Suttle, C.A., and Sullivan, M.B. (2007) Exploring the vast diversity of marine viruses. Oceanography **20**: 135–139.

Brouns, S.J.. J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.. H., Snijders, A.P.. L., *et al.* (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. Science (80-) **321**: 960 –964.

Campbell, B.J., Smith, J.L., Hanson, T.E., Klotz, M.G., Stein, L.Y., Lee, C.K., *et al.* (2009) Adaptations to submarine hydrothermal environments exemplified by the genome of Nautilia profundicola. PLoS Genet **5**: e1000362.

Clokie, M.R.J. and Mann, N.H. (2006) Marine cyanophages and light. Environ Microbiol **8**: 2074–82.

Clokie, M.R.J., Shan, J., Bailey, S., Jia, Y., Krisch, H.M., West, S., and Mann, N.H. (2006) Transcription of a "photosynthetic" T4-type phage during infection of a marine cyanobacterium. Environ Microbiol **8**: 827–35.

Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., Delong, E.F., and Chisholm, S.W. (2006) Genomic islands and the ecology and evolution of Prochlorococcus. Science (80-) **311**: 1768–70.

Cuadros-Orellana, S., Martin-Cuadrado, A.-B., Legault, B., D'Auria, G., Zhaxybayeva, O., Papke, R.T., and Rodriguez-Valera, F. (2007) Genomic plasticity in prokaryotes: the case of the square haloarchaeon. ISME J **1**: 235–45.

Dinsdale, E., Pantos, O., and Smriga, S. (2008) Microbial ecology of four coral atolls in the Northern Line Islands. PLoS One **3**: e1584.

Drummond, A., Ashton, B., Cheung, M., Heled, J., Kearse, M., Moir, R., *et al.* (2009) Geneious v4. 7. Biomatters Ltd.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26**: 2460–1.

Edwards, R.A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D.M., *et al.* (2006) Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics **7**: 57.

Elsaied, H., Stokes, H.W., Nakamura, T., Kitamura, K., Fuse, H., and Maruyama, A. (2007) Novel and diverse integron integrase genes and integron-like gene cassettes are prevalent in deep-sea hydrothermal vents. Environ Microbiol **9**: 2298–312.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., *et al.* (2010) The Pfam protein families database. Nucleic Acids Res **38**: D211–22.

Garneau, J.E., Dupuis, M.È., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., *et al.* (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature **468**: 67–71.

Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., *et al.* (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. Cell **139**: 945–956.

Hellweger, F.L. (2009) Carrying photosynthesis genes increases ecological fitness of cyanophage in silico. Environ Microbiol **11**: 1386–94.

Holden, J.F., Summit, M., and Baross, J.A. (1998) Thermophilic and hyperthermophilic microorganisms in 3-30 ^{o}C hydrothermal fluids following a deep-sea volcanic eruption. FEMS Microbiol Ecol **25**: 33–41.

Horvath, P., Barrangou, R., and Hovarth, P. (2010) CRISPR/Cas, the immune system of bacteria and archaea. Science (80-) **327**: 167–170.

Huber, J.A., Butterfield, D.A., and Baross, J.A. (2003) Bacterial diversity in a subseafloor habitat following a deep-sea volcanic eruption. FEMS Microbiol Ecol **43**: 393–409.

Huber, J.A., Butterfield, D.A., and Baross, J.A. (2002) Temporal changes in archaeal diversity and chemistry in a mid-ocean ridge subseafloor habitat. Appl Environ Microbiol **68**: 1585–1594.

Huber, J.A., Mark Welch, D.B., Morrison, H.G., Huse, S.M., Neal, P.R., Butterfield, D.A., and Sogin, M.L. (2007) Microbial population structures in the deep marine biosphere. Science (80-) **318**: 97–100.

Jiang, S.C. and Paul, J.H. (1998) Gene transfer by transduction in the marine environment. Appl Environ Microbiol **64**: 2780–2787.

Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., *et al.* (2011) Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat Struct Mol Biol **18**: 529–536.

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res **28**: 27–30.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res **40**: D109–14.

Koonin, E. V., Makarova, K.S., and Aravind, L. (2001) Horizontal gene transfer in prokaryotes: Quantification and classification. Annu Rev Microbiol **55**: 709–742.

Kristensen, D.M., Mushegian, A.R., Dolja, V. V, and Koonin, E. V (2009) New dimensions of the virus world discovered through metagenomics. Trends Microbiol **18**: 11–19.

Kryazhimskiy, S. and Plotkin, J.B. (2008) The population genetics of dN/dS. PLoS Genet **4**: e1000304.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. Genome Biol **5**: R12.

Labrie, S.J., Samson, J.E., and Moineau, S. (2010) Bacteriophage resistance mechanisms. Nat Rev Microbiol **8**: 317–327.

Leplae, R., Lima-Mendez, G., and Toussaint, A. (2010) ACLAME: a CLAssification of Mobile genetic Elements, update 2010. Nucleic Acids Res **38**: D57–61.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**: 1658–9.

Lindell, D., Jaffe, J.D., Coleman, M.L., Futschik, M.E., Axmann, I.M., Rector, T., *et al.* (2007) Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. Nature **449**: 83–6.

Lindell, D., Jaffe, J.D., Johnson, Z.I., Church, G.M., and Chisholm, S.W. (2005) Photosynthesis genes in marine viruses yield proteins during host infection. Nature **438**: 86–89.

Marraffini, L.A. and Sontheimer, E.J. (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat Rev Genet **11**: 181–190.

McDaniel, L.D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K.B., and Paul, J.H. (2010) High frequency of horizontal gene transfer in the oceans. Science (80-) **330**: 50.

McHardy, A.C., Martín, H.G., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2007) Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods **4**: 63–72.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., *et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics **9**: 386.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., *et al.* (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima. Nature **399**: 323–329.

Van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M., Brouns, S.J.J., van der Oost, J., *et al.* (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem Sci **34**: 401–407.

Ortmann, A.C. and Suttle, C.A. (2005) High abundances of viruses in a deep-sea hydrothermal vent system indicates viral mediated microbial mortality. Deep Sea Res Part I Oceanogr Res Pap **52**: 1515–1527.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J. V, Chuang, H.-Y., Cohoon, M., *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res **33**: 5691–702.

Paul, J.H. (2008) Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? ISME J **2**: 579–589.

Reiter, W.-D., Palm, P., and Yeats, S. (1989) Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. Nucleic Acids Res **17**: 1907–1914.

Rho, M., Tang, H., and Ye, Y. (2010) FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res **38**: e191.

Rodriguez-Brito, B., Rohwer, F., and Edwards, R.A. (2006) An application of statistics to comparative metagenomics. BMC Bioinformatics **7**: 162.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B.B., Rodriguez-Brito, B., Pasic, L., Thingstad, T.F., Rohwer, F., *et al.* (2009) Explaining microbial population genomics through phage predation. Nat Rev Microbiol **7**: 828–36.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol **75**: 7537–41.

Schmieder, R. and Edwards, R. (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics **27**: 863–4.

Schmieder, R., Lim, Y.W., Rohwer, F., and Edwards, R. (2010) TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. BMC Bioinformatics **11**: 341.

Schrenk, M.O., Kelley, D.S., Delaney, J.R., and Baross, J.A. (2003) Incidence and diversity of microorganisms within the walls of an active deep-sea sulfide chimney. Appl Environ Microbiol **69**: 3580–3592.

Sharon, I., Alperovitch, A., Rohwer, F., Haynes, M., Glaser, F., Atamna-Ismaeel, N., *et al.* (2009) Photosystem I gene cassettes are present in marine virus genomes. Nature **461**: 258–262.

Simon, C., Wiezer, A., Strittmatter, A.W., and Daniel, R. (2009) Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome. Appl Environ Microbiol **75**: 7519–26.

Sorek, R., Kunin, V., and Hugenholtz, P. (2008) CRISPR — a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat Rev Microbiol **6**: 181–186.

Summit, M. and Baross, J.A. (2001) A novel microbial habitat in the mid-ocean ridge subseafloor. Proc Natl Acad Sci U S A **98**: 2158–63.

Tai, V., Poon, A.F.Y., Paulsen, I.T., and Palenik, B. (2011) Selection in coastal Synechococcus (cyanobacteria) populations evaluated from environmental metagenomes. PLoS One **6**: e24249.

Vega Thurber, R., Willner-Hall, D., Rodriguez-Mueller, B., Desnues, C., Edwards, R.A., Angly, F., *et al.* (2009) Metagenomic analysis of stressed coral holobionts. Environ Microbiol **11**: 2148–63.

Voordeckers, J.W., Starovoytov, V., and Vetriani, C. (2005) *Caminibacter mediatlanticus* sp. nov., a thermophilic, chemolithoautotrophic, nitrate-ammonifying bacterium isolated from a deep-sea hydrothermal vent on the Mid-Atlantic Ridge. Int J Syst Evol Microbiol **55**: 773–9.

Waldor, M.K. and Mekalanos, J.J. (1996) Lysogenic conversion by a filamentous phage encoding cholera toxin. Science (80-) **272**: 1910 –1914.

Weinbauer, M.G., Brettar, I., and Hölfe, M.G. (2003) Lysogeny and virus-induced mortality of bacterioplankton in surface, deep, and anoxic marine waters. Limnol Oceanogr **48**: 1457–1465.

Wilke, A., Harrison, T., Wilkening, J., Field, D., Glass, E.M., Kyrpides, N., *et al.* (2012) The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. BMC Bioinformatics **13**: 141.

Williamson, S.J., Cary, S.C., Williamson, K.E., Helton, R.R., Bench, S.R., Winget, D., and Wommack, K.E. (2008) Lysogenic virus–host interactions predominate at deep-sea diffuse-flow hydrothermal vents. ISME J **2**: 1112–1121.