The Macroecology of

Microorganisms:

From Pattern to Process

David Robert Clark

A thesis submitted for the degree of Doctor of Philosophy in Microbiology

School of Biological Sciences,

University of Essex

December 2017

“If I could do it all over again, and relive my vision in the twenty-first century, I would be a microbial ecologist. … Into that world I would go with the aid of modern microscopy and molecular analysis. I would cut my way through clonal forests sprawled across grains of sand, travel in an imagined submarine through drops of water proportionately the size of lakes, and track predators and prey in order to discover new life ways and alien food webs”

- A quote from E. O. Wilson, co-author of “The Theory of Island Biogeography” (1967). I fear too much of my time as a PhD student has been spent in front of a computer, and not enough piloting imagined submarines...

Abstract

Microorganisms are the most ubiquitous, diverse, and functionally important organisms on Earth, yet their ecological patterns, and the underlying causative processes that determine their distributions over large spatial scales, remain poorly understood. Therefore, I test for macroecological patterns and processes within microbial communities. I use a combination of data generating approaches including meta-analysis, published sequence datasets and databases, and high-throughput sequencing, coupled with modern statistical methods. Firstly, I show that metagenomic sequencing, is superior to amplicon sequencing as a method of surveying microbial biodiversity, as it detected more diversity at all taxonomic levels. However, cost analysis shows that metagenomics is prohibitively expensive for macroecological studies, where many samples are required. I find that classic macroecological patterns, such as the distance-decay of similarity, are context-dependent and vary according to ecological context, and methodological differences. I therefore make recommendations for future analyses of spatial analyses of microbial communities. Furthermore, I show that whilst microbial communities may exhibit distance-decay relationships, they do not necessarily form biogeographic regions, highlighting a difference in the macroecology of micro- and macroorganisms. I build on this finding by showing that different regional microbial communities can show considerably different responses to the same environmental gradient, hinting that regional communities play an important role in determining microbial community composition at local scales. Finally, I investigate whether regional-scale climatic variables determine the distributions of microorganisms. I show that the climatic drivers and influence of these drivers varies strongly between microbial taxa, suggesting that different microbial taxa have very different macroecologies. I conclude that macroecological patterns in microorganisms may not be as general as in “macroorganisms”, and can be highly context-dependent, varying with taxon, regional metacommunity dynamics, or methodological choices. Careful consideration of these factors is therefore required when attempting to understand how microbial communities will respond to future environmental changes.

Acknowledgements

To pick out specific people who have contributed to my completion of this thesis is difficult, as there are many. I must begin by thanking my supervisors, Alex Dumbrell, Terry McGenity, and Graham Underwood, for their unwavering guidance, and for allowing me the autonomy to develop my own ideas and skills. Thanks to Leonie Mourot and Mégane Mathieu for their assistance with the molecular analyses for Chapter 4, and to Laurent Dufossé, for providing many of the samples for this chapter. Thanks also to Nikolai Friberg and Guy Woodward for providing samples for Chapter 5. I also thank Joseph Chipperfield for his valuable guidance on the statistical analyses in Chapter 6. Many thanks also to the academic and technical staff associated with the Environmental Microbiology and Ecology group, for their general interest, support, and feedback on my work.

On a personal level I thank my friends, Tom Jones, Alessandro Moret, Dan Baxter, Leila Tavaleili, Maddy Giles, Nataša Šibanc, Dani Harris, Hannah Prentice, and Scott Warren, simply for putting up with me. I'd like to give special mention to my friend and peer Aisha Coggan, who sadly passed away before finishing her PhD, but who would have made a brilliant academic and is someone I am grateful I had the chance to work alongside.

Finally, I owe everything to my family for their continuing support. Thank you to my partner, Jo, and my daughter, Emmy, for motivating me to succeed in

Academia. I hope this thesis will be the first in a long line of achievements to make them proud.

Declaration

I declare that this thesis is the result of my own work, and was written entirely by myself. Chapters 2 and 6 contain data from previously published sources, and are appropriately cited where applicable. Chapter 4 has been published as:

Clark DR, Mathieu M, Mourot L, Dufossé L, Underwood GJC, Dumbrell AJ, McGenity TJ (2017) Biogeography at the Limits of Life: Do Extremophilic Microbial Communities Show Biogeographic Regionalisation? *Global Ecology and Biogeography,* **26:**1435–1446*.*

David Clark

University of Essex

2017

Contents

List of Figures

Index of Tables

Chapter 1

Introduction

Macroecology

Macroecology - Seeing The Wood For The Trees

Ecology is the scientific study of the distribution, abundance, and activity of organisms in time and space. Brown and Maurer (1989) are credited with popularising the term “macroecology” in their synthesis paper. The authors recognised a need for a new approach, expressing their dissatisfaction at the ability of (then) contemporary ecological approaches to answer “big” questions, and to draw useful generalisations on the ecology of species. Conservation ecology was typically concerned with relatively few species or populations, and often over relatively small spatial scales, thus making it difficult to extrapolate conclusions to other species, populations, or geographic regions. Experimental ecology on the other hand is too simplistic and unable to replicate the complexity of real ecological communities. Therefore, a new approach was required, that considered multiple species, populations, and (spatial or temporal) scales (Keith *et al.*, 2012), and so macroecology was born.

Macroecology is perhaps easier to define through examples than via any colloquial definition. For instance, Brown (1995) cites the example of determining the extinction risk of species under climate change. A typical ecological approach might be based around experimental manipulations to determine the effects of climatic change on individual populations. However, such an approach would require excessive extrapolation of the results to other populations in order to determine the global effect of climate change on a given species. In contrast, a macroecological approach might utilise data on many populations, and use general ecological laws such as the species-area relationship to determine the risk of extinction for a given species. Brown, (1995) argues that macroecology, is not a distinct sub-discipline of ecology, but a philosophical approach to addressing research questions that were previously beyond the scope of other ecological disciplines. Therefore, macroecology is not solely focused on any one aspect of ecology. Instead, it is an approach to addressing ecological hypotheses or questions, often utilising large datasets in order to elucidate general patterns and processes in ecology.

Is Macroecology Biogeography?

Another similar (to macroecology) field is biogeography. Biogeography is a discipline that studies the distribution of organisms in space and time, by combining ecology with phylogeny (Brown and Lomolino, 1998). Compared with biogeography, macroecology is a relatively recently developed field (Nee, 2002). Macroecology undoubtedly shares traits with biogeography in that both are concerned with studying species' distributions, considering large spatiotemporal scales, and using large datasets. As a consequence, the similarities with biogeography invoked suggestions that macroecology was merely a rebranding of biogeography (Fisher, 2002). However, such suggestions were rejected on the basis that macroecology seeks a more mechanistic, process-based understanding of the general statistical patterns observed in nature (Blackburn and Gaston, 2002a), and that macroecology is not solely concerned with “large” scales, but “large” questions (Marquet, 2002).

Arguably, the semantics of whether these two research fields are distinct, is perhaps irrelevant. What is important, is that the development of macroecology as a discipline has stimulated a great volume of research searching for general patterns in ecology at a variety of spatiotemporal scales in a variety of organisms (Brown, 1999). The body of research that has arisen, as a result of the stimulus provided by macroecology, is a better validation of macroecology than any paper dealing with semantics and definitions of the word (Riddle, 2005; Keith *et al.*, 2012).

Methods and Data in Macroecology

From its origins, macroecology has been a heavily quantitative research programme, relying on the examination of statistical patterns to illuminate underlying ecological mechanisms (Ruggiero and Hawkins, 2006; Beck *et al.*, 2012; Blackburn and Gaston, 2006). Primarily, the reliance of macroecology on statistical patterns is largely due to the same reasons that macroecology as a field was developed; that incorporating multiple scales, and the complexity of natural systems are both necessary to understand ecological mechanisms. For these reasons, the statistical methods utilised by macroecologists must be able to detect ecological signal amongst the noise (unexplained variation) of the natural world.

Macroecological datasets may vary in two main properties, which determine their appropriateness to addressing specific hypotheses or questions. These properties are grain and extent (Fig. 1.1). Extent describes the size of the study system (either spatially or temporally) over which measurements are taken, whilst grain describes the size of the space represented by an individual measurement, or the resolution of the data. For example, in a survey of plant biodiversity in a field, the extent would be the area over which quadrats are placed, whilst the grain could be considered as the quadrat size. Macroecologists typically make use of large extent, and large grain data. That is, the data represent a large spatiotemporal scale, and individual measurements may represent large spatial areas or temporal periods within this (Beck *et al.*, 2012). Both grain and extent have been shown to influence the nature and detection of macroecological patterns (Rahbek, 2005; Hulme, 2008; Steinbauer *et al.*, 2012). Therefore, careful consideration of these properties is required to ensure the data are suitable to test the hypothesis in question (Blackburn and Gaston, 2002b).

Whilst differences in grain and extent may influence ecological perception, other biases may also be present that may ultimately effect ecological inference. Imperfect detection of species may lead to false absences, in which the species is erroneously recorded as absent from a sample (Royle *et al.*, 2005; Guillera-Arroita, 2017), or false positives, in which a species is mistakenly recorded as positive (Pillay *et al.*, 2014; Chambert *et al.*, 2015). Such errors can be caused by suboptimal sampling techniques, or identification errors. Spatial biases may be present, in which the data or sampling do not adequately represent the study extent (Yang *et al.*, 2013), perhaps due to under-, or over-sampling in certain areas.

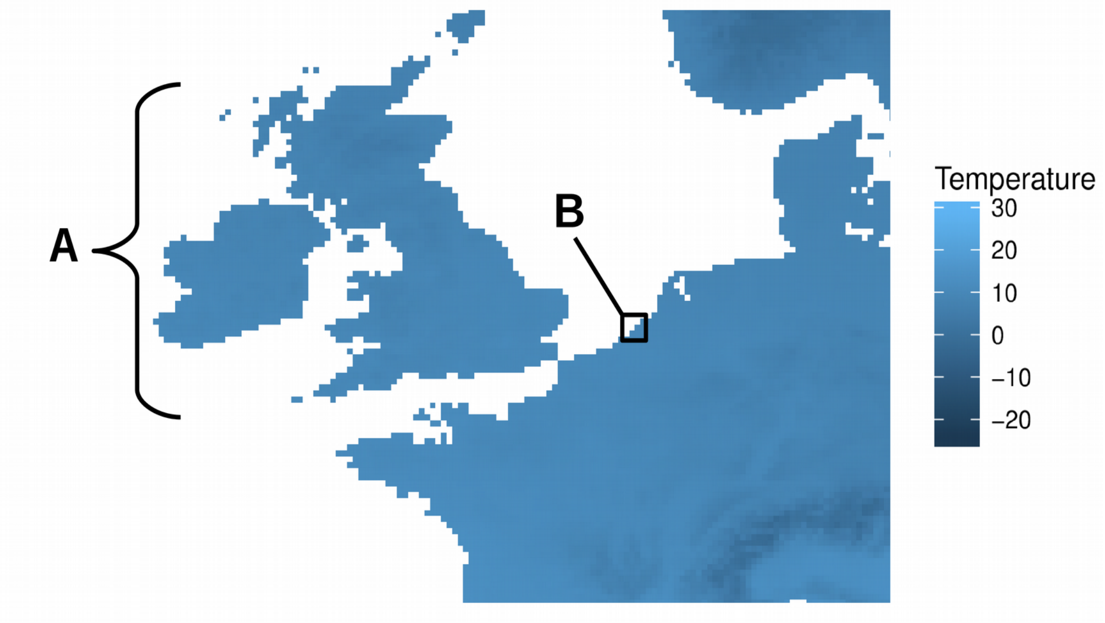


Figure 1.1 Climate data for the United Kingdom and NW Europe. Study extent (A) indicates the spatial scale of the dataset, whilst grain (B) is the resolution of each measurement, or in this case, the pixel size.

Furthermore, macroecological datasets often exhibit statistical properties that effect the inference of patterns. For instance, environmental predictors may be co-linear, potentially making it difficult to decipher the true causative relationships driving the pattern of interest (Dormann *et al.*, 2013). Data may also exhibit spatial (or temporal) autocorrelation, whereby measurements taken close together (in space or time) are likely to influence each other (Segurado *et al.*, 2006; Dormann, 2007), violating the assumption of many statistical tests that observations are independent from each other.

The macroecological approach to addressing hypotheses in ecology therefore requires highly robust statistical methods in order to address potentially confounding biases, and account for the complexity present in nature (McGill, 2003; Blackburn, 2004; Beck *et al.*, 2012).

Microorganisms

Ubiquitous Microorganisms

Fossil evidence dating back at least 3.5 billion years suggests that microorganisms were the first inhabitants of Earth. It is widely postulated that Bacteria represent the oldest domain, with Archaea evolving from Bacteria at least 3.5 billion years ago (Shen *et al.*, 2001). The Archaea are thought to be the origins of Eukaryotic life on Earth (Woese *et al.*, 1990), with the recently described “Asgard” superphylum, a potential link between these domains

(Zaremba-Niedzwiedzka *et al.*, 2017). Whilst some consider viruses to be part of the tree of life, I do not consider them within the scope of this thesis (Moreira and López-García, 2009).

Microorganisms have evolved to occupy every available niche on Earth, and are truly ubiquitous in the environment, defining the known limits of life (Hallsworth *et al.*, 2007). Microorganisms are also recognised as phylogenetically and functionally highly diverse. Their ubiquity, activity, and functional capabilities mean they are drivers of Earth's biogeochemical cycles, and thus contribute greatly to global ecosystem processes (Falkowski *et al.*, 2008) as well as being mediators of human health and disease. Thus, the study of microbial ecology has widespread implications for our understanding of disease, ecosystem functioning and processes, and the origins of life on Earth.

The Problematic Study of Microorganisms

Until recently, microbial ecology has lagged far behind other ecological fields. Principally, this was due to problematic methodologies associated with the study of microorganisms. Being microscopic, simple tasks such as enumerating and identifying them requires advanced methods. Van Leeuwanhoek was the first to observe microbial life through a self-constructed microscope, and he dubbed the organisms “wee animalcules” (Lane, 2015). However, whilst microscopy remains a valuable technique in microbial ecology, it is not an effective way of enumerating or identifying the microorganisms in environmental samples. This is because microorganisms can have extremely high population densities and display cryptic speciation, a phenomenon in which organisms resemble each other morphologically, yet are phylogenetically distinct (Martiny *et al.*, 2006). For this reason, microbial ecologists must typically use molecular methods to directly study the genetic material present in microbial cells in order to enumerate the diversity and abundance of microorganisms in the environment.

Molecular Methods in Microbial Ecology

The molecular methods typically utilised in microbial ecology studies are based on the genetic material present in microorganisms, DNA or RNA. The genetic material must first be extracted from the microbial cells present in the environmental sample(s) of interest. The aim of DNA (or RNA) extraction is to obtain a pure and unbiased DNA sample, free from environmental compounds that might inhibit downstream analyses (Martin-Laurent *et al.*, 2001). Typically, a single phylogenetically or, functionally informative marker gene is amplified from the original sample using the polymerase chain reaction (PCR). Primers are chosen to amplify this gene, that should ideally target all of the organisms of interest, and exclude those not of interest. In practise, this is challenging and therefore a great emphasis is placed upon optimising primer designs for various applications in microbial ecology (Wang and Qian, 2009; Klindworth *et al.*, 2013; Hugerth *et al.*, 2014). The resulting "amplicon" (pool of amplified genes), can then be subjected to a variety of different analyses, dependent on the goal of the research and the available resources.

The simplest approaches to studying microbial diversity are those referred to as "fingerprinting" methods. In these methods, mixed amplicons are separated out in a gel based on their melting or denaturing properties (as in DGGE), or are sheared into fragments of differing sizes (as in TRFLP). The amplicon of each taxon will have different denaturing properties, or will be sheared at specific points, resulting in distinctly sized fragments. These fragments are then separated out on a gel (Green *et al.*, 2014; Tebbe *et al.*, 2015). The resulting bands within the gel can then be interpreted as the presence of a given taxon, and the composition of microbial communities can be compared across samples. This data can be used to compare the α- (within sample) and composition (β-diversity) of microbial communities. It is also possible to infer the relative abundance of the taxa based on the "intensity" of bands. However, the identity of the taxa present can not be gained from fingerprinting approaches alone, and so fingerprinting is of limited use.

Therefore, to identify the taxa present, sequencing must be undertaken. Early sequencing approaches involved the creation of a clone library, followed by Sanger sequencing. After PCR amplification of the marker gene, the mixed amplicon is then separated by inserting into a plasmid (small section of DNA). A bacterial host is then transformed with the plasmid, and within the bacterial cells, the plasmid is amplified. Positive clones are then selected at random for sequencing (Leigh *et al.*, 2015). By comparing the resulting sequence to carefully curated databases, the identities of the organisms present can be revealed (DeSantis *et al.*, 2006; Wang *et al.*, 2007). Clone library sequencing enables analyses of both α-, and β-diversity, in addition to quantifying the relative abundance of microbial taxa. However, the workflow involved in creating a clone library is relatively slow, and Sanger sequencing is costly and low-throughput, as only single DNA sequences can be obtained at a time. As a result, clone library analysis is not suitable for diverse microbial communities, where many sequences are needed.

The "Omics" Revolution

Driven by the need for more sequencing depth to adequately sample microbial diversity, the development of "omics" methods has allowed microbial ecologists to delve ever deeper into microbial communities. Whilst Sanger sequencing require separation of the amplicon, high-throughput amplicon sequencing (also known as metagenetic sequencing or metabarcoding) using a next-generation sequencing platform does not. These platforms are able to sequence many millions of amplicons at a time, thus generating far larger datasets more rapidly, and at a far cheaper price (Loman *et al.*, 2012; Quail *et al.*, 2012). Whilst the cost of reagents is higher, "omics" methods allow far greater sequencing coverage, and have yielded insight into the enormous taxonomic and functional diversity present in microbial communities (Roesch *et al.*, 2007; Caporaso *et al.*, 2012).

High-throughput amplicon sequencing is able to characterise the diversity of microbial communities based on a single amplified gene. However, metagenomic sequencing is able to characterise taxonomic, and functional diversity. Unlike amplicon sequencing, metagenomics does not require amplification of a particular gene. Instead, extracted DNA is randomly sheared into fragments. These fragments represent not only taxonomic genes, but also house-keeping, and functional genes. The fragmented DNA is then sequenced on a next-generation platform, yielding insight into the taxonomic and metabolic diversity within microbial communities (Dinsdale *et al.*, 2008).

Bioinformatics

After sequencing, the resulting sequence dataset requires careful quality control in order to yield biologically meaningful conclusions. Therefore, rigorous bioinformatic analyses have become a standard part of modern molecular microbial ecology workflows (Dumbrell *et al.*, 2016). Next-generation sequencing platforms are known to have higher error rates than traditional Sanger sequencing approaches, and may produce lower quality sequences. It is therefore necessary to remove low quality sequences from the dataset, or to trim sequences to remove low quality base calls. These processes have been shown to improve the consistency of sequence datasets (Bokulich *et al.*, 2013; Schirmer *et al.*, 2015).

If the sequences are from an amplicon sequencing run, sequences can then be grouped into operational taxonomic units (OTUs). OTUs act as a pseudo-species concept in molecular microbial ecological analyses, and help to reduce the enormous complexity of next-generation amplicon sequencing datasets into data that is more easily analysable. An enormous amount of research has focussed on developing and improving algorithms to cluster sequences into OTUs (Edgar, 2010; Schloss, 2013; Mahé *et al.*, 2014; Rognes *et al.*, 2016). The choice of OTU clustering algorithm has been shown to have significant impacts on the ecological conclusions gained in microbial ecology studies, thus careful consideration is required when choosing how to cluster OTUs (Lekberg *et al.*, 2014; He *et al.*, 2015; Kopylova *et al.*, 2016). Recently, there has been a movement towards using exact sequence variants (ESVs) as an alternative to OTUs (Callahan *et al.*, 2017). ESVs do not group similar sequences, instead they represent groups of identical (or near identical) sequences. ESVs supposedly offer improved taxonomic resolution to OTUs as they “split”, rather than “lump”, closely related taxa. However, ESVs are only valid if the dataset is completely free of sequencing errors (otherwise, erroneous ESVs are created), an assumption that is rarely verified, even after quality control of sequences (Schirmer *et al.*, 2015).

Furthermore, it is unclear whether ESVs represent a taxonomic entity any more ecologically meaningful than OTUs (Berry *et al.*, 2017).

After OTUs have been created, taxonomy can be assigned. Various algorithms have been created to assign taxonomy to marker gene sequences (Altschul *et al.*, 1990; Wang *et al.*, 2007; Edgar, 2016; Bokulich *et al.*, 2017), whilst the availability of high-quality and well curated sequence databases largely varies between marker genes and taxa (DeSantis *et al.*, 2006; Cole *et al.*, 2013; Guillou *et al.*, 2013; Quast *et al.*, 2013; Deshpande *et al.*, 2015). For example, the bacterial (and archaeal) 16S rRNA gene has several large, and curated databases available for taxonomy assignment, whilst other marker genes are less comprehensively covered by taxonomic databases.

The workflow described above is typical for an amplicon sequencing approach in microbial ecology (Dumbrell *et al.*, 2016). However, it is also possible to assess taxonomic diversity from metagenomic analyses. Within the metagenome are fragments of the commonly studied taxonomic marker genes used in amplicon sequencing (Guo *et al.*, 2016). Various algorithms have been developed to identify and extract commonly studied marker gene sequences from metagenomes (Sharpton *et al.*, 2011; Bengtsson-Palme *et al.*, 2015). In theory, such an approach should be beneficial over amplicon sequencing, as the biases associated with primer selection and PCR amplification are avoided. However, only a small fraction of the sequences originate from marker genes and therefore, much greater sequencing depth is required for metagenomics to be a viable method for surveying microbial diversity. The compromise between reducing bias at the cost of extra sequencing depth remains to be determined.

Microbial Macroecology

“Everything Is Everywhere” (EiE)

The first hypothesis relating to microbial macroecology can be attributed to Dutch microbiologist Lourens Baas Becking. Having microscopically studied the microorganisms found in hypersaline environments, Baas Becking concluded that “Everything is everywhere but, the environment selects” (Baas Becking, 1934). Within this eloquent statement, Baas Becking made two predictions about the macroecology of micoorganisms. Firstly, that microorganisms are dispersed ubiquitously. He postulated that their small size and often high population densities would allow them to passively disperse over large geographic distances e.g. by wind or oceanic currents. Secondly, EiE states that the environment selects (de Wit and Bouvier, 2006). This suggests that whilst micoorganisms may be dispersed ubiquitously, we observe distinct microbial communities in different habitats because of environmental selection.

Community Assembly in Microbial Communities

The widespread uptake of molecular methods to study microbial communities allowed microbial ecologists to test the EiE hypothesis in a variety of microbial habitats. The two predictions made by EiE are compatible with two competing, but not mutually exclusive, theories to describe how communities assemble, neutral theory and niche theory.

Niche theory suggests that the species in a community are only able to coexist if they have different environmental preferences, or resource uses, known as niche differentiation. There are many different interpretations of what a niche actually is. Perhaps the most general definition of the niche is the Hutchinsonian definition, which states that the niche of an organism can be thought of as an “n-dimensional hypervolume”, each axis of which represents an aspect of the environment (both biotic and abiotic) for which the species has limits. Within the limits of all axes, exists the niche of the species, or the set of environmental conditions that allow it to survive and reproduce indefinitely (Hutchinson, 1957). More specific definitions include the Grinnellian niche which describes the niche as the set of (often abiotic) non-interactive environmental conditions required by the species, whereas the Eltonian niche defines the niche as the resources required by a species (Soberón, 2007). If a microbial community assembles under niche theory, there will be a close relationship between environmental conditions, and the composition and diversity of the community.

Whilst niche theory focuses on the differences between species in terms of their interactions with their environment, neutral theory's central tenet is that all species (within a trophic level) are ecologically identical (Hubbell, 2001). Neutral theory therefore suggests that species arrive in, and are lost from communities as a result of stochastic processes such as migration, immigration, birth and death. Under neutral assembly, relationships between microbial communities and the environment are not predicted, instead neutral theory predicts that communities will decrease in compositional similarity with increasing distance, as a result of decreased dispersal between them. Neutral theory has arguably been one of the most controversial ecological theories in recent history, particularly because it suggests that species are ecologically equal, whilst most ecologists focus on the difference between species (Hubbell, 2006). Yet, neutral theory has been successfully invoked to explain high diversity maintained in some communities including tropical forests, coral reefs, and to a lesser extent, microbial communities (Hubbell, 1997; Condit *et al.*, 2002; Dumbrell *et al.*, 2010).

Niche and neutral theory are not the only community assembly theories, and various attempts have been made to integrate these seemingly polar ideologies (e.g. Tilman, 2004, Vellend 2010). In reality, it is likely that, to some extent, most communities are subjected to both niche and neutral processes, and various studies suggest that the relative roles of each may be dependent on the spatial scale at which we observe them (Chase, 2014).

Environmental Filtering in Microbial Communities

Under niche based community assembly, and consistent with EiE, microbial communities are expected to show close relationships with environmental factors. The process by which mal-adapted taxa are excluded from the community is referred to as “environmental filtering”. Environmental filtering can be thought of as a stack of sieves, in which the the least restrictive environmental variables are the upper most sieves with the largest mesh, and therefore exclude the fewest taxa. Lower “sieves” represent more restrictive environmental variables, and thus the community composition may show stronger correlation with these variables. The concept of environmental filtering has been questioned in the wider field of ecology, as it may be indistinguishable from demographic variables or competition (Cadotte and Tucker, 2017).

The range of environmental filters that can act on a microbial community represent the environment at a range of spatial scales. The spatial scale at which microbes perceive, and interact with, their environment is incredibly small, often at the micron scale. Therefore, most microbial ecology studies focus on environmental factors that act over comparably small scales, although this is somewhat limited by the resolution at which we are able to quantify the environment (Fierer, 2008). In soils, factors such as pH (Lauber *et al.*, 2009; Dumbrell *et al.*, 2010) have frequently been shown to be important, whereas in aquatic environments, factors such as salinity correlate well with microbial community structure (Logares *et al.*, 2013).

The link between small-scale, local environmental variables and microbial community structure represents something of a spatial paradox in microbial macroecology in that, the grain of studies is very small, yet the extent can be incredibly large. Furthermore, a result of the focus on small-scale environmental factors is that the role of larger scale environmental factors have remained relatively understudied in microbial macroecology. In particular the role of climate in determining the distributions of microorganisms is poorly understood, and warrants further research (Pajunen *et al.*, 2016; Kivlin *et al.*, 2017).

Whilst many macroecological studies on microorganisms focus on abundant habitats such as soils, locally extreme systems represent unique and interesting systems in which to study environmental filtering of microbial communities (Maček *et al.*, 2016). Extreme systems are often characterised by the presence of one or more that are significantly different from the surrounding *millieu.* For example, in hypersaline systems (such as coastal solar salterns) the salinity of water can be an order of magnitude higher than the surrounding seawater (Antón *et al.,* 2000). Such systems enable microbial ecologists to examine the effects of individual environmental variables on microbial communities whilst minimising the influence of other, potentially confounding variables. These properties have made locally extreme environments well studied systems to examine the effects of single environmental variables on microbial communities (e.g. Dillon *et al*., 2013). However, the macroecology of these systems is still poorly understood in that, the generality of community-environment relationships between such systems has not been well addressed (Sharp *et al.*, 2014). By studying spatially replicated, locally-extreme environmental gradients such as geothermally warmed stream systems, the generality of patterns in microbial community could be illuminated, helping us to understand the unifying macroecological processes acting on microbial communities.

Dispersal and Distance-Decay of Similarity

Due to the EiE hypothesis, the potential for dispersal limitation and neutral community assembly in microbial communities has received great attention. EiE suggests that microbes have extremely high dispersal probabilities, and are therefore able to access all suitable habitat, leading to cosmopolitan distributions. Early work on some microbial taxa supported the idea of high dispersal and cosmopolitanism in microorganisms. A study of ciliates in freshwater ponds yielded surprisingly high local:global species richness ratios, leading to the conclusion that geographically distinct communities should not differ much in composition (Finlay, 2002). However, the conclusions were largely rejected by microbial ecologists because the method of identification (microscopy) did not adequately address the possibility of cryptic species within ciliates, and therefore likely overestimated the local:global species ratio (Katz *et al.*, 2005). Additionally, because of the uncertainty around estimates of global microbial richness, the validity of local:global richness ratios is highly questionable (Foissner, 2006). Other studies provide evidence for widespread cosmopolitanism by illustrating the potential for global dispersal in microorganisms, for example via wind (Herbold *et al.*, 2014).

Despite this, many studies have reported evidence for dispersal limitation in microorganisms, in contrast to EiE. In particular, a multitude of studies have reported distance-decay relationships (Whitaker *et al.,* 2003; Dumbrell *et al.*, 2010; Martiny *et al.,* 2011; Wetzel *et al.*, 2012; Bahram *et al.*, 2013; Milici *et al.*, 2016). This relationship describes the way in which the similarity in composition of communities decreases with increasing geographic distance between them (Fig. 1.2; Nekola and White, 1999). Distance-decay curves are often interpreted as evidence of dispersal limitation, as microorganisms may only disperse to nearby communities, thus leading to more compositionally similar communities. Though, it is worth noting that distance-decay relationships may also arise due to spatial structure in the environment (e.g. where the environment is spatially autocorrelated). Studies have reported distance-decay relationships in a variety of microbial taxa, over a variety of spatial scales. However, many studies have failed to observe such relationships (e.g. Queloz *et al*., 2011). The generality of this macroecological relationship within microbial communities is therefore unclear, and requires further study (Soininen *et al.*, 2007).

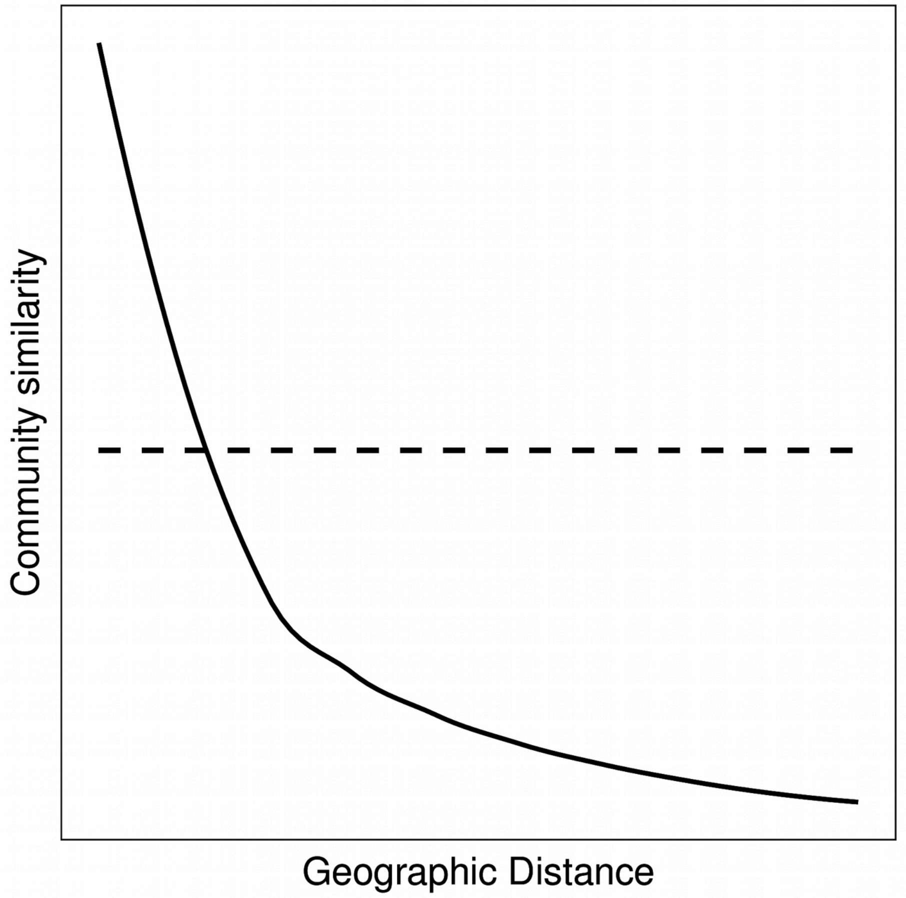


Figure 1.2 Two potential relationships between community similarity and geographic distance. A typical distance-decay curve (solid) in which community similarity declines with increasing distance due to dispersal limitation or spatially autocorrelated environmental conditions. Alternatively, community similarity may be unrelated to geographic distance (dashed) if species are dispersal unlimited.

Thesis Rationale

Microorganisms are the most ubiquitous, numerous, and diverse organisms on Earth, playing key roles in mediating ecosystem level processes, that in turn, benefit human society (Falkowski *et al.*, 2008). The widespread uptake of high-throughput sequencing technologies to study microbial communities in great detail has revolutionised our understanding of microbial ecology. The past two decades of research in microbial ecology has shifted the view of microbial communities from being entirely structured by the environment (Finlay, 2002), to being highly complex, with a myriad of ecological processes and interactions shaping them (Martiny *et al.*, 2006; Hanson *et al.*, 2012; Barberán *et al.*, 2014). However, few universal patterns have emerged from this research and therefore, the generality of various relationships in microbial ecology remain unclear. As a result, the application of ecological theory to understand the assembly, diversity, and activity of microbial communities is limited to a “one size fits all” approach, which may radically underestimate the complexity of microbial macroecology.

Spatial processes (processes that influence the emigration and immigration of organisms into an environment e.g. dispersal) have provided a stimulus for much research in microbial ecology (Green *et al.*, 2004), principally due to the provocative lack of importance placed on spatial processes by EiE. Yet, few general principles as to when spatial processes are, or are not important, have emerged. Furthermore, it is unclear whether the spatial processes operating on microbial communities are capable of producing regional patterns, similar to those observed for “macroorganisms”. Similarly, the role of small scale environmental factors in structuring microbial communities has been extensively studied. But again, the generality of environmental drivers across systems has been understudied, and therefore the consistency with which spatially distinct microbial communities respond to environmental change is unknown (Telford *et al.*, 2006). Finally, microbial ecologists have tended to focus on environmental drivers that operate over small spatial scales, such as the physicochemical environment. This has resulted in the influence of environmental factors that operate over larger spatial scales, such as regional climate, being largely ignored. The vulnerability of microbial communities to large scale environmental changes, such as climate change, is therefore unknown. Consequently, the general aim of this thesis is to take a “macroecological” approach to examining the factors that structure environmental microbial communities, and to test the generality of these factors across systems and microbial taxa.

Thesis Structure

- In Chapter 2, I test which method of sequencing (amplicon sequencing or metagenomic sequencing) detects the most diversity from microbial communities, and whether the difference in cost is favourable.
- In Chapter 3, I examine the generality of a macroecological relationship, the distance-decay of similarity, in microbial communities.

I test whether this relationship varies by biological context, or methodological differences between studies.

- In Chapter 4, I test the effects of spatial processes on extremophilic microbial communities. Specifically, I examine whether microbial communities may form biogeographic regions, as observed for “macroorganisms”. Additionally, I test whether individual microbial taxa show biogeographically structured distributions.
- Within Chapter 5, I investigate the generality of environment-diversity relationships in microbial communities, using a set of spatially replicated thermal gradients around the Arctic circle.
- In Chapter 6, I investigate the extent to which climate controls the distribution of microbial taxa over global scales, and whether this varies between and within microbial taxonomic groups.

References

Altschul SF, Gish W, Miller WT, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Antón J, Rosselló-Mora R, Rodríguez-Valera F, Amann R (2000) Extremely halophilic bacteria in crystallizer ponds from solar salterns. *Applied and Environmental Microbiology*, **66**, 3052–3057.

Baas Becking LGM (1934) *Geobiologie of inleiding tot de milieukunde*. W.P. Van Stockum & Zoon, The Hague, Netherlands.

Bahram M, Kõljalg U, Courty PE, *et al*. (2013) The distance decay of similarity in communities of ectomycorrhizal fungi in different ecosystems and scales. *Journal of Ecology*, **101**, 1335–1344.

Barberán A, Casamayor EO, Fierer N (2014) The microbial contribution to macroecology. *Frontiers in Microbiology*, **5**, 203.

Beck J, Ballesteros-Mejia L, Buchmann CM, *et al*. (2012) What's on the horizon for macroecology? *Ecography*, **35**, 673–683.

Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, Nilsson RH (2015) METAXA2: Improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Molecular Ecology Resources*, **15**, 1403–1414.

Berry MA, White JD, Davis TW, *et al*. (2017) Are oligotypes meaningful ecological and phylogenetic units? A case study of *Microcystis* in Freshwater lakes. *Frontiers in Microbiology*, **8**, 365.

Blackburn TM (2004) Method in macroecology. *Basic and Applied Ecology*, **5**, 401–412.

Blackburn TM, Gaston KJ (2002a) Macroecology is distinct from biogeography. *Nature*, **418**, 723.

Blackburn TM, Gaston KJ (2002b) Scale in macroecology. *Global Ecology and Biogeography*, **11**, 185–189.

Blackburn TM, Gaston KJ (2006) There's more to macroecology than meets the eye. *Global Ecology and Biogeography*, **15**, 537–540.

Bokulich NA, Subramanian S, Faith JJ, *et al*. (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods*, **10**, 57–59.

Bokulich NA, Kaehler BD, Rideout JR, *et al*. (2017) Optimizing taxonomic classification of marker gene sequences. *PeerJ*, **5,** e3208v1.

Brown JH (1995) *Macroecology*. University of Chicago Press, Chicago, IL, USA.

Brown JH (1999) Macroecology: progress and prospect. *Oikos*, **87**, 3–14.

Brown JH, Lomolino M V. (1998) *Biogeography*. Sinauer, Sunderland, MA, USA.

Brown JH, Maurer BA (1989) Macroecology: the division of food and space among species on continents. *Science*, **243**, 1145–1150.

Cadotte MW, Tucker CM (2017) Should Environmental Filtering be Abandoned? *Trends in Ecology and Evolution*, **32**, 429–437.

Callahan BJ, McMurdie PJ, Holmes SP (2017) Exact sequence variants should replace operational taxonomic units in marker gene data analysis. *bioRxiv*, 113597.

Caporaso JG, Lauber CL, Walters WA, *et al*. (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal*, **6**, 1621–1624.

Chambert T, Miller DAW, Nichols JD (2015) Modeling false positive detections in species occurrence data under different study designs. *Ecology*, **96**, 332–339.

Chase JM (2014) Spatial scale resolves the niche versus neutral theory debate. *Journal of Vegetation Science*, **25**, 319–322.

Cole JR, Wang Q, Fish JA, *et al*. (2013) Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, **42**, D633–D642.

Condit R, Pitman N, Leigh EG, *et al*. (2002) Beta-diversity in tropical forest trees. *Science*, **295**, 666–669.

DeSantis TZ, Hugenholtz P, Larsen N, *et al*. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, **72**, 5069–5072.

Deshpande V, Wang Q, Greenfield P, *et al*. (2015) Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences. *Mycologia*, **108**, 1–5.

Dillon JG, Carlin M, Gutierrez A, Nguyen V, McLain N (2013) Patterns of microbial diversity along a salinity gradient in the Guerrero Negro solar saltern, Baja CA Sur, Mexico. *Frontiers in Microbiology,* **4**, 399.

Dinsdale EEA, Edwards RAR, Hall D, *et al*. (2008) Functional metagenomic profiling of nine biomes. *Nature*, **452**, 629–632.

Dormann CF (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, **16**, 129–138.

Dormann CF, Elith J, Bacher S, *et al*. (2013) Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 27–46.

Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH (2010) Relative roles of niche and neutral processes in structuring a soil microbial community. *The ISME Journal*, **4**, 337–345.

Dumbrell AJ, Ferguson RMW, Clark DR (2016) Microbial Community Analysis by Single-Amplicon High-Throughput Next Generation Sequencing: Data Analysis -- From Raw Output to Ecology. In: *Hydrocarbon and Lipid Microbiology Protocols: Microbial Quantitation, Community Profiling and Array Approaches* (eds McGenity TJ, Timmis KN, Nogales B), pp. 155–206. Springer, Heidelberg, Germany.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Edgar R (2016) SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*, 74161.

Falkowski PG, Fenchel T, Delong EF (2008) The Microbial Engines That Drive Earth’s Biogeochemical Cycles. *Science*, **320**, 1034–1039.

Fierer N (2008) Microbial biogeography: patterns in microbial diversity across space and time. In: *Accessing Uncultivated Microorganisms: from the Environment to Organisms and Genomes and Back* (ed Zengler, K), pp. 95–115. The American Society of Microbiology, Washington, DC, USA.

Finlay BJ (2002) Global dispersal of free-living microbial eukaryote species. *Science*, **296**, 1061–1063.

Fisher HJ (2002) Macroecology: new or biogeography revisited? *Nature*, **417**, 787.

Foissner W (2006) Biogeography and dispersal of micro-organisms: a review emphasizing protists. *Acta Protozoologica*, **45**, 111–136.

Green JL, Holmes AJ, Westoby M, *et al*. (2004) Spatial scaling of microbial eukaryote diversity. *Nature*, **432**, 747–750.

Green SJ, Leigh MB, Neufeld JD (2014) Denaturing Gradient Gel Electrophoresis (DGGE) for Microbial Community Analysis. In: *Hydrocarbon and Lipid Microbiology Protocols: Microbial Quantitation, Community Profiling and Array Approaches* (eds McGenity TJ, Timmis KN, Nogales B), pp. 77–99. Springer, Berlin, Heidelberg.

Guillera-Arroita G (2017) Modelling of species distributions, range dynamics and communities under imperfect detection: advances, challenges and opportunities. *Ecography*, **40**, 281–295.

Guillou L, Bachar D, Audic S, *et al*. (2013) The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, **41**, D597-D604.

Guo J, Cole JR, Zhang Q, Brown CT, Tiedje JM (2016) Microbial Community Analysis with Ribosomal Gene Fragments from Shotgun Metagenomes. *Applied and Environmental Microbiology*, **82**, 157–166.

Hallsworth JE, Yakimov MM, Golyshin PN, *et al*. (2007) Limits of life in MgCl2-containing environments: Chaotropicity defines the window. *Environmental Microbiology*, **9**, 801–813.

Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH (2012) Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews Microbiology*, **10**, 497–506.

He Y, Caporaso JG, Jiang XT, *et al*. (2015) Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome*, **3**, 20.

Herbold CW, Lee CK, McDonald IR, Cary SC (2014) Evidence of global-scale aeolian dispersal and endemism in isolated geothermal microbial communities of Antarctica. *Nature Communications*, **5**, 3875.

Hubbell SP (1997) A unified theory of biogeography and relative species abundance and its application to tropical rain forests and coral reefs. *Coral Reefs*, **16**, S9–S21.

Hubbell SP (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton, NJ, USA.

Hubbell SP (2006) Neutral theory and the evolution of ecological equivalence. *Ecology*, **87**, 1387–1398.

Hugerth LW, Muller EEL, Hu YOO, *et al*. (2014) Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoS ONE*, **9**, e95567.

Hulme PE (2008) Contrasting alien and native plant species-area relationships: The importance of spatial grain and extent. *Global Ecology and Biogeography*, **17**, 641–647.

Hutchinson GE (1957) Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, **22,** 415–427.

Katz LA, McManus GB, Snoeyenbos-West OLO, Griffin A, Pirog K, Costas B, Foissner W (2005) Reframing the “everything is everywhere” debate: Evidence for high gene flow and diversity in ciliate morphospecies. *Aquatic Microbial Ecology*, **41**, 55–65.

Keith SA, Webb TJ, Böhning-Gaese K, *et al*. (2012) What is macroecology? *Biology Letters*, **8**, 904–906.

Kivlin SN, Muscarella R, Hawkes CV, Treseder KK (2017) The Predictive Power of Ecological Niche Modeling for Global Arbuscular Mycorrhizal Fungal Biogeography. In: *Biogeography of Mycorrhizal Symbiosis* (ed Tedersoo L), pp. 143–158. Springer, Heidelberg, Germany.

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, **41**, e1.

Kopylova E, Navas-Molina JA, Mercier C, *et al*. (2016) Open-Source Sequence Clustering Methods Improve the State Of the Art. *mSystems*, **1**, e00003-15.

Lane N (2015) The unseen world: reflections on Leeuwenhoek (1677) “Concerning little animals.” *Philosophical Transactions of the Royal Society of London B*, **370**, 20140344.

Lauber CL, Hamady M, Knight R, Fierer N (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Applied and Environmental Microbiology*, **75**, 5111–5120.

Leigh MB, Taylor L, Neufeld JD (2017) Clone Libraries of Ribosomal RNA Gene Sequences for Characterization of Microbial Communities. In: *Hydrocarbon and Lipid Microbiology Protocols: Microbial Quantitation, Community Profiling and Array Approaches* (eds McGenity TJ, Timmis KN, Nogales B), pp. 127–154. Springer, Berlin, Heidelberg.

Lekberg Y, Gibbons SM, Rosendahl S (2014) Will different OTU delineation methods change interpretation of arbuscular mycorrhizal fungal community patterns? *New Phytologist*, **202**, 1101–1104.

Logares R, Lindström ES, Langenheder S, *et al*. (2013) Biogeography of bacterial communities exposed to progressive long-term environmental change. *The ISME Journal*, **7**, 937–948.

Loman NJ, Misra R V, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, **30**, 434–439.

Maček I, Vodnik D, Pfanz H, Low-Décarie E, Dumbrell AJ (2016) Locally Extreme Environments as Natural Long-Term Experiments in Ecology. *Advances in Ecological Research,* **55**, 283–323.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M (2014) Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, **2**, e593.

Marquet PA (2002) The search for general principles in ecology. *Nature*, **418**, 723.

Martin-Laurent F, Philippot L, Hallet S, Chaussod R, Soulas G, Catroux G (2001) DNA Extraction from Soils: Old Bias for New Microbial Diversity Analysis Methods. *Applied and Environmental Microbiology*, **67**, 2354–2359.

Martiny JBH, Bohannan BJM, Brown JH, *et al*. (2006) Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology*, **4**, 102–112.

Martiny JB, Eisen JA, Penn K, Allison SD, Horner-Devine MC (2011) Drivers of bacterial β-diversity depend on spatial scale. *Proceedings of the National Academy of Sciences USA*, **108**, 7850-7854.

McGill B (2003) Strong and weak tests of macroecological theory. *Oikos*, **102**, 679–685.

Milici M, Tomasch J, Wos-Oxley ML, *et al*. (2016) Bacterioplankton biogeography of the Atlantic ocean: A case study of the distance-decay relationship. *Frontiers in Microbiology*, **7**, 590.

Moreira D, López-García P (2009) Ten reasons to exclude viruses from the tree of life. *Nature Reviews Microbiology*, **7**, 306–311.

Nee S (2002) Biodiversity: thinking big in ecology. *Nature*, **417**, 229–30.

Nekola JC, White PS (1999) The distance decay of similarity in biogeography and ecology. *Journal of Biogeography*, **26**, 867–878.

Pajunen V, Luoto M, Soininen J (2016) Climate is an important driver for stream diatom distributions. *Global Ecology and Biogeography*, **25**, 198–206.

Pillay R, Miller DAW, Hines JE, Joshi AA, Madhusudan MD (2014) Accounting for false positives improves estimates of occupancy from key informant interviews. *Diversity and Distributions*, **20**, 223–235.

Quail M, Smith ME, Coupland P, *et al*. (2012) A tale of three next generation sequencing platforms: comparison of Ion torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.

Quast C, Pruesse E, Yilmaz P, *et al*. (2013) The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, **41**, 590–596.

Queloz V, Sieber TN, Holdenrieder O, McDonald BA, Grünig CR (2011) No biogeographical pattern for a root-associated fungal species complex. *Global Ecology and Biogeography*, **20**, 160–169.

Rahbek C (2005) The role of spatial scale and the perception of large-scale species-richness patterns. *Ecology Letters*, **8**, 224–239.

Riddle BR (2005) Is biogeography emerging from its identity crisis? *Journal of Biogeography*, **32**, 185–186.

Roesch LFW, Fulthorpe RR, Riva A, *et al*. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME Journal*, **1**, 283–290.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2409v1.

Royle JA, Nichols JD, Kéry M (2005) Modelling occurrence and abundance of species when detection is imperfect. *Oikos*, **110**, 353–359.

Ruggiero A, Hawkins BA (2006) Mapping macroecology. *Global Ecology and Biogeography*, **15**, 433–437.

Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, **43**, e37.

Schloss PD (2013) Secondary structure improves OTU assignments of 16S rRNA gene sequences. *The ISME Journal*, **7**, 457–460.

Segurado P, Araújo MB, Kunin WE (2006) Consequences of spatial autocorrelation for niche-based models. *Journal of Applied Ecology*, **43**, 433–444.

Sharp CE, Brady AL, Sharp GH, Grasby SE, Stott MB, Dunfield PF (2014) Humboldt's spa: microbial diversity is controlled by temperature in geothermal environments. *The ISME Journal*, **8**, 1166–1174.

Sharpton TJ, Riesenfeld SJ, Kembel SW, *et al*. (2011) PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS Computational Biology*, **7**, e1001061.

Shen Y, Buick R, Canfield DE (2001) Isotopic evidence for microbial sulphate reduction in the early Archaean era. *Nature*, **410**, 77–81.

Soberón J (2007) Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters*, **10**, 1115–1123.

Soininen J, McDonald R, Hillebrand H (2007) The distance decay of similarity in ecological communities. *Ecography*, **30**, 3–12.

Steinbauer MJ, Dolos K, Reineking B, Beierkuhnlein C (2012) Current measures for distance decay in similarity of species composition are influenced by study extent and grain size. *Global Ecology and Biogeography*, **21**, 1203–1212.

Tebbe CC, Dohrmann AB, Hemkemeyer M, Näther A (2015) Microbial Community Profiling: SSCP and T-RFLP Techniques. In: *Hydrocarbon and Lipid Microbiology Protocols: Microbial Quantitation, Community Profiling and Array Approaches* (eds McGenity TJ, Timmis KN, Nogales B), pp. 101–126. Springer, Heidelberg, Germany.

Telford RJ, Vandvik V, Birks HJB (2006) Dispersal Limitations Matter for Microbial Morphospecies. *Science*, **312**, 1015–1015.

Tilman D (2004) Niche tradeoffs, neutrality, and community structure: A stochastic theory of resource competition, invasion, and community assembly. *Proceedings of the National Academy of Sciences USA*, **101**, 10854–10861.

Vellend M (2010) Conceptual synthesis in community ecology. *The Quarterly Review of Biology*, **85**, 183–206.

Wang Y, Qian PY (2009) Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS ONE*, **4**, e7401.

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naiive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–5267.

Wetzel CE, de Bicudo DC, Ector L, Lobo EA, Soininen J, Landeiro VL, Bini LM (2012) Distance Decay of Similarity in Neotropical Diatom Communities. *PLoS ONE*, **7**, e45071.

Whitaker RJ, Grogan DW, Taylor JW (2003) Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science*, **301**, 976-978.

de Wit R, Bouvier T (2006) "Everything is everywhere, but, the environment selects"; what did Baas Becking and Beijerinck really say? *Environmental Microbiology*, **8**, 755–758.

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences USA*, **87**, 4576–4579.

Yang W, Ma K, Kreft H (2013) Geographical sampling bias in a large distributional database and its effects on species richness-environment models. *Journal of Biogeography*, **40**, 1415–1426.

Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, *et al*. (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, **541**, 353–358.

Chapter 2

Shotgun Metagenomics Detects More Diversity than Amplicon Sequencing; But at What Cost?

Acknowledgements: Graham JC Underwood, Terry J McGenity, and Alex J Dumbrell

Abstract

Choosing the optimal method of quantifying biodiversity within microbial communities is of central importance to all microbial ecologists. Two methods frequently utilised to quantify the number of microbial taxa in an environment are amplicon-, and metagenomic sequencing. Amplicon sequencing relies on amplification of a marker gene, whilst metagenomics allows “shotgun” sequencing of DNA. Here, I sought to test which method is better able to quantify the number of species, both at the operational taxonomic unit level (OTU), and phylogenetic level. To do this, I assembled a dataset of paired metagenomic and amplicon sequence datasets, consisting of over one billion sequences, and covering a variety of biomes and sequencing platforms. I extracted putative 16S rRNA gene fragments from metagenomic data, and clustered them along with 16S rRNA gene amplicon sequences against a non-redundant (clustered at 99% sequence similarity) custom database of 16S rRNA gene sequences. At the OTU level, diversity was greater in the metagenomic datasets, and the difference between metagenomic and

amplicon datasets was larger for more diverse communities. Furthermore, phylogenetic diversity was also significantly greater in metagenomic datasets, at every rank from phylum to genus. Trends in α-diversity observed in amplicon datasets were accurately reflected in metagenomic datasets, showing that the same ecological conclusions are reached with either approach. I also determined the cost associated with producing appropriately sized metagenomic sequence datasets, and found that in most cases, the cost was an order of magnitude higher than amplicon sequencing. Overall, whilst metagenomic sequencing is able to detect more taxa and OTUs than amplicon sequencing, the cost of metagenomics as a method of surveying microbial diversity remains prohibitively expensive, especially so when the microbial communities are highly diverse and a large number of samples are required. In most cases, amplicon sequencing therefore remains a favourable option in terms of the balance between cost and ability to detect microbial diversity.

Introduction

Accurately quantifying the diversity of microbial communities remains one of the central challenges to microbial ecologists (Locey & Lennon, 2016; Schloss *et al*., 2016). The vast majority of microbial life is not (easily) cultivable (Staley & Konopka, 1985) and thus, molecular methods have become the dominant approach to address this problem. Rapid technological progress in molecular microbial ecology, driven by next-generation sequencing (NGS) developments over the last two decades, has made this possible. These sequencing platforms allow the high-throughput sequencing of many millions of DNAs in parallel, shedding light on the enormous diversity of microbial communities (Sogin *et al*., 2006; Roesch *et al*., 2007, Aslam *et al*., 2017, Clark *et al.*, 2017).

Generally, high-throughput sequencing approaches can be categorised as amplicon-based or metagenomic. In both approaches, DNA is initially extracted from the sample of interest. In amplicon sequencing (also referred to as metagenetic sequencing or metabarcoding), a single phylogenetically informative marker gene is chosen to be sequenced. An ideal marker gene should be conserved (present) across all of the taxa of interest, providing a genetic target that, with an appropriate primer set, can be used to study the entire range of organisms of interest in a given environmental sample. Additionally, a marker gene should show some variability between taxonomic groups, as this will allow the delineation and identification of different taxonomic groups present in the community. Frequently studied marker genes include the 16S rRNA gene to study Bacteria and Archaea (e.g. Roesch *et al*., 2007), 18S rRNA or cytochrome oxidase I genes to study Eukarya (e.g. Bik *et al*., 2012; Fayle *et al*., 2015), the ITS region to study fungal communities (Schoch *et al*., 2012), or specific functional genes associated with specific lineages (e.g. the hydrazine oxidoreductase gene to study anammox bacteria, Lansdown *et al*., 2016). The gene is then amplified via the PCR reaction, with specific primers, before being sequenced on a next-generation sequencing platform. This is by far the most common approach used in microbial ecology and has dramatically contributed to our understanding of microbial diversity (Schloss *et al*., 2016).

However, despite its widespread use, amplicon sequencing is flawed. Several steps throughout the workflow of preparing samples for amplicon sequencing have been shown to introduce bias to the results, with potential effects on estimates of microbial diversity. These biases are predominantly introduced by the necessity for PCR amplification of marker genes. Initially, the selection of suboptimal primers can influence the results even before PCR. Rarely do primer sets ever target 100% of the target organisms, and this can lead to entire taxonomic lineages remaining unamplified, and therefore missing from the sequence dataset (Jeon *et al.*, 2008; Hong *et al.*, 2009). Furthermore, the different properties of individual marker gene sequences can influence the efficiency of amplification during PCR. The relative content of guaninecytosine (G-C) base pairs in a sequence can affect how easily the DNA denatures during denaturation steps in PCR. G-C base pairs form three hydrogen bonds, as opposed to two hydrogen bonds formed between adenine and thymine (A-T), meaning that marker gene sequences with many G-C bases may resist denaturation, and therefore amplify less efficiently (Aird *et al*., 2011). Furthermore, if marker gene sequences are of variable length, shorter sequences are often preferentially amplified over long sequences, resulting in taxa with longer marker genes being under-represented in the sequence dataset (Huber *et al*., 2009; Bellemain *et al*., 2010). Formation of chimeric sequences, artificial hybrid sequences produced as a result of incomplete primer extension, have the potential to inflate estimates of microbial diversity, and may be mistakenly interpreted as novel diversity (Ashelford *et al*., 2005; Pinto & Raskin, 2012), although chimeric sequences can be removed during bioinformatic analyses (e.g. Edgar *et al*., 2011). Finally, other aspects of the PCR procedure have also been demonstrated to potentially bias sequencing datasets if not appropriately controlled such as, the number of PCR cycles and use of different polymerase enzymes (Fonseca *et al*., 2012; Pinto & Raskin, 2012). Combined, these biases can lead to erroneous conclusions about the ecology of microbial communities (Jumpponen, 2007; Bergmann *et al*., 2011; Fredriksson *et al*., 2013), although this is not universal (Cotton *et al.,* 2014).

An alternative approach to amplicon sequencing is to use metagenomic sequencing (also referred to as “shotgun sequencing”). Here, instead of amplifying a gene of interest, total DNA is sheared into smaller fragments that are easier to sequence. These fragments therefore represent not only phylogenetic marker genes, but also functional, and “housekeeping” genes associated with the microbiota, as well as extra-cellular and non-microbial DNA. The library of DNA fragments are then sequenced in a high-throughput manner. Short DNA sequence fragments can then be aligned and computationally assembled into longer fragments, referred to as contigs. Contigs can then be aligned to genomic databases in order to infer functional capacity and taxonomic composition (Sharpton, 2014), allowing identification of the organisms present, and their functional capabilities.

Additionally, fragments of taxonomic marker genes (see above) can be bioinformatically “mined”, allowing analyses of microbial diversity (Bengtsson-Palme *et al*., 2015; Guo *et al*., 2016). The major advantage of this workflow is that it circumnavigates PCR amplification, therefore negating the effects of primer- and PCR-associated biases. However, the compromise is that, because taxonomically informative marker genes (such as the 16S rRNA gene) often represent < 1% of the total sequences in the dataset (Guo *et al*., 2016), enormous sequencing depth may be required to obtain enough marker gene fragments to sufficiently sample the community, thereby limiting the number of samples that can be sequenced at once. For analyses of β-diversity (between site/sample diversity), where large numbers of samples are required, amplicon sequencing may be most appropriate method of quantifying the microbial community, as many samples can be multiplexed (sequenced on the same run) and any biases are likely to be consistent across samples. Instead, metagenomic sequencing is a promising technique for studies of α-diversity (within site/sample diversity), in which the taxonomic diversity of a smaller number of samples is the desired information.

However, the usefulness of metagenomics for quantifying α-diversity in microbial communities depends on its ability to detect extra diversity, and the financial cost of such an approach. Few studies have previously compared the ability of metagenomic and amplicon sequencing to detect microbial diversity, and none have quantified the potential difference in cost of these two approaches. Therefore, I sought to determine whether metagenomic or amplicon sequencing quantifies the most microbial diversity, at both the OTU level, and taxonomic level. Furthermore, I analysed the cost of these approaches to determine whether the difference in cost would promote the use of one approach over the other.

Methods

Data Sources

I initially downloaded 8 datasets comprising of paired metagenomic, and amplicon sequence datasets (Table 2.1) from the mg-RAST server (Meyer *et al.,* 2008), Sequence Read Archive database (Leinonen *et al.,* 2010), or other data repositories described in each manuscript. Whilst I did not perform an exhaustive literature search for suitable datasets, these studies represent a variety of biomes, sequencing platforms, and sequencing depths, allowing a robust and representative comparison between these two molecular workflows.

Bioinformatic Analyses

A schematic of all bioinformatic analyses is provided (Fig. 2.1). For sequence datasets where quality scores were provided, quality trimming was conducted using Sickle (Joshi & Fass, 2011), with a quality threshold of Q20 and discarded sequences shorter than 100 nucleotides, as filtering low quality and short reads has been shown to improve estimates of microbial diversity (Bokulich *et al*., 2013). Datasets from Illumina sequencing platforms (HiSeq or MiSeq) consisting of forward and reverse reads were pair-end aligned using the Pear algorithm implemented in the PandaSeq software (Masella *et al*., 2012; Zhang *et al*., 2014), as pair-end alignment reduces error rates in Illumina sequence datasets (Schirmer *et al*., 2015). To extract putative 16S rRNA gene sequences from the metagenomic sequence datasets, I used

Metaxa2 (Bengtsson-Palme *et al*., 2015), which identifies marker gene sequences using hidden Markov models.

In order to compare the diversity between (metagenomic and amplicon) datasets, it is not possible to directly compare sequences. Normally, in an amplicon sequencing approach, sequences are clustered against each other into operational taxonomic units (OTUs), at a given sequence similarity threshold (often 97%). However, because metagenomic 16S rRNA gene fragments represent discontinuous parts of the entire 16S rRNA gene, they can not be directly compared with the amplicon sequences, as it would be impossible to determine whether an amplicon sequence and a non-overlapping metagenomic 16S rRNA fragment belong to the same species or not. Therefore, in order to compare the diversity recovered by each method at the OTU level, a closed-reference OTU clustering strategy (also known as phylotyping) was chosen. Here, sequence reads are clustered against a reference database of near full length 16S rRNA gene sequences. Therefore, even non-overlapping fragments should theoretically map to the same database sequence. To do this, a non-redundant version of the RDP bacterial 16S rRNA database (Cole *et al*., 2013) was created. This involved using VSEARCH (Rognes *et al*., 2016) to de-replicate the original database to remove any duplicate sequences, sort the unique sequences by length, and finally clustering at 99% sequence similarity. This was done to remove highly similar sequences from the database that could cause metagenomic 16S fragments to map to different OTUs, artificially inflating our estimates of alpha diversity in metagenomic libraries (where fragments may not overlap). Having created the reference database, I mapped marker gene sequences (from metagenomic and amplicon datasets) against this database at a 97% similarity threshold using VSEARCH.

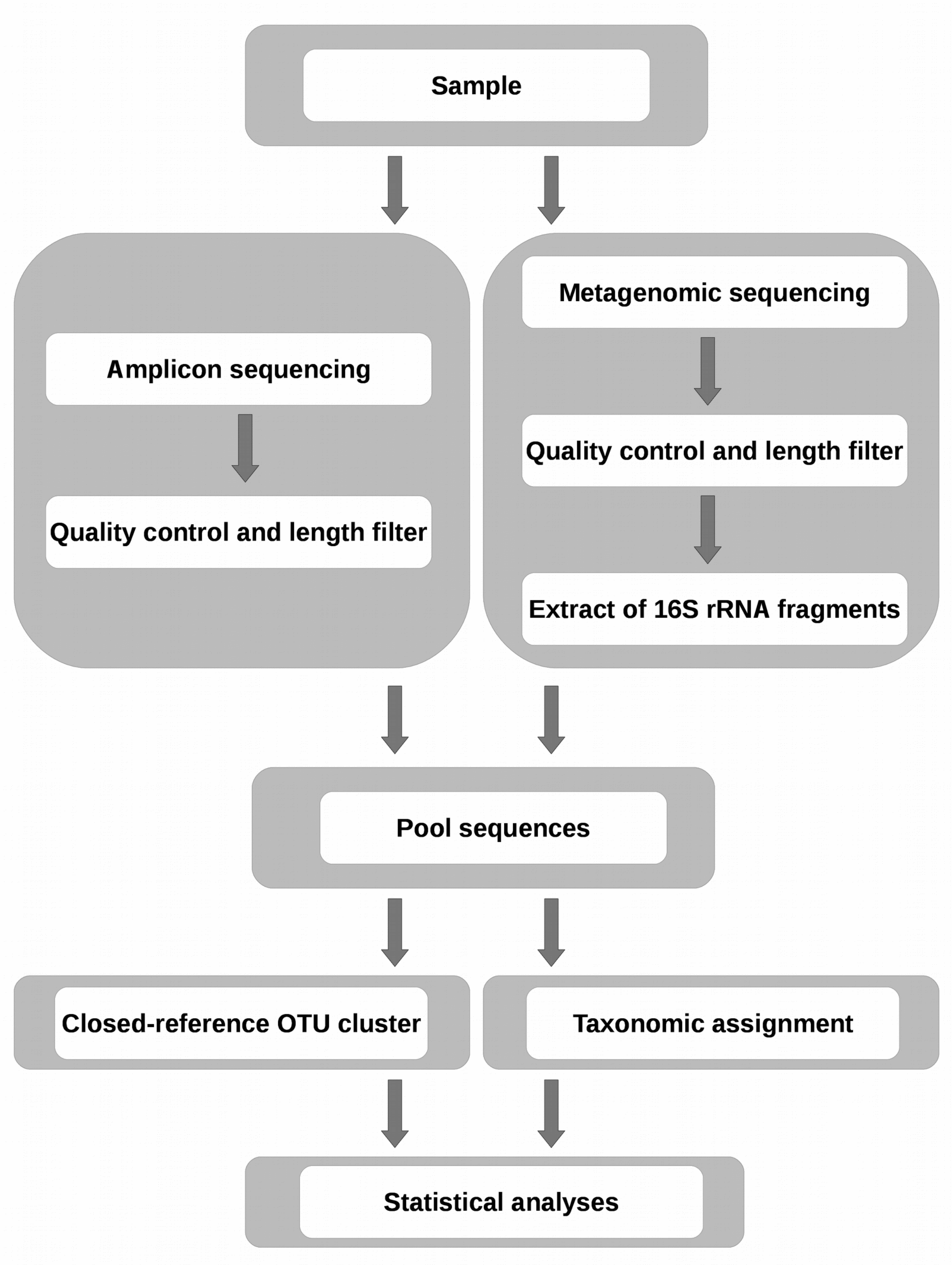

Sample
Amplicon sequencing
Quality control and length filter
Metagenomic sequencing
Quality control and length filter
Extract of 16S rRNA fragments
Pool sequences
Closed-reference OTU cluster
Taxonomic assignment
Statistical analyses

Figure 2.1 A schematic view of the bioinformatics workflow used to analyse and compare metagenomic and amplicon sequence datasets.

Table 2.1. Details of the sequence datasets used in this study.

Dataset	Biome	Sequencing platform	Number of paired samples[b]
Aslam *et al.* (2016)	Desert soil	Pyrosequencing (amplicon) Illumina MiSeq (Metagenome)	1
Chan *et al.* (2015)	Spring	Illumina HiSeq (Metagenome) Illumina MiSeq (amplicon)	1
Delforno *et al.* (2017)	Wastewater	Illumina HiSeq (Metagenome) Illumina MiSeq (amplicon)	1
Gibbons *et al.* (2014)	River sediment	Illumina HiSeq (Metagenome) Illumina MiSeq (amplicon)	14
Muegge *et al.* (2011)	Zoo animal faeces	Pyrosequencing	38
Navarrete *et al.* (2015)	Forest soil	Pyrosequencing (amplicon) Illumina HiSeq (metagenome)	15
Steven *et al.* (2012)	Desert soil	Pyrosequencing	6
Turnbaugh *et al.* (2009)	Human gut	Pyrosequencing	18 metagenome-V2 16S rRNA 18 metagenome-V6 16S rRNA

[a] In studies where different sequencing platforms were used to generate metagenomic and amplicon libraries, this is indicated in parentheses.

[b] Refers to the number of samples from each study for which sequence data could be obtained and paired sequence datasets could be verified.

In addition to comparing amplicon and metagenomic sequencing at the OTU level, I also compared them at the phylogenetic level. To do this, taxonomy was assigned to all amplicon and metagenomic 16S rRNA gene fragments using the RDP classifier (Wang *et al*., 2007), set to a minimum confidence threshold of 0.7. Any non-bacterial sequences (such as archaeal 16S rRNA fragements) were excluded from taxonomic analyses.

Statistical Analyses

To analyse whether metagenomic or amplicon sequencing recovers more OTUs or taxa, I first rarefied OTU tables and taxonomic tables for each pair (amplicon and metagenome) of samples, to whichever sample had the smallest number of 16S rRNA sequences. This was to ensure that unequal library sizes did not bias our results, as diversity is known to be strongly influenced by library size (Gihring *et al*., 2012). I then used negative binomial generalised linear mixed effects models (GLMMs) to test whether amplicon or metagenomic sequencing recover significantly different OTU or taxonomic richness (Bolker *et al*., 2009). These models allow modeling of count data, which is non-normally distributed (as it is integer and bound by 0), without transformation, improving interpretation (O'Hara and Kotze, 2010). I included a study specific intercept in all models to account for the expected differences in OTU or taxonomic richness between datasets from different biomes (e.g. soils are expected to be richer than gut microbiomes). For analyses of taxonomic diversity, I created separate models for each of the five taxonomic ranks provided by the RDP classifier (phylum, class, order, family, and genus).

To examine the cost of each approach I first obtained cost estimates for the three different sequencing platforms utilised by the studies featured here (Table 1). These were obtained from Quail *et al*. (2012) and Loman *et al*. (2012), and were calculated by dividing the cost per run by the expected output (to obtain a cost per base), and then multiplying by the average read length of each platform (to obtain a cost per sequence). These costs provide a useful estimate of the relative costs associated with each sequencing approach. I then estimated the cost of each sequence dataset. For metagenomes, I estimated the number of metagenomic sequences required to yield a number of 16S rRNA fragments equivalent to the corresponding amplicon dataset.

Results

Our initial dataset consisted of 1.16 billion metagenomic and amplicon sequences. Quality filtering reduced this total to 1.03 billion sequences, of which 1.02 billion were metagenomic sequences, and 2.86 million were from amplicon datasets. Quality filtered metagenomic datasets contained a mean of 9.14 million sequences (std. error = 2.27 million), whereas amplicon datasets contained a mean of 25,558 sequences (std. error = 4,567).

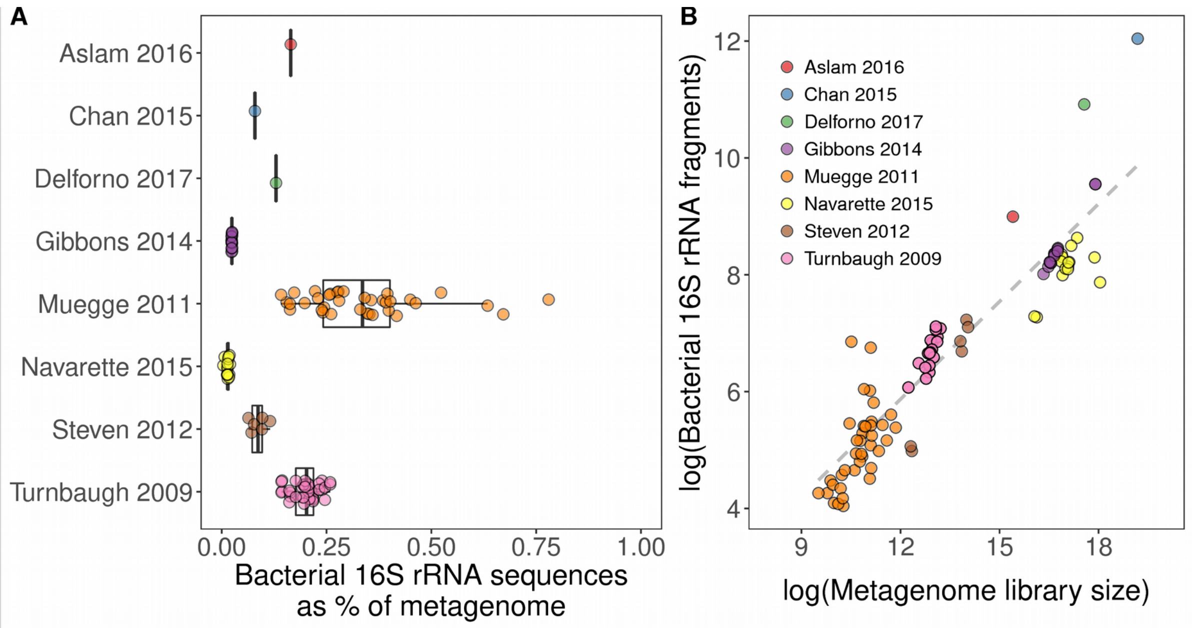


Figure 2.2 (A) The proportion of metagenomic sequences identified as putative bacterial 16S rRNA fragments within each dataset, and (B) the relationship between total metagenome library size and the number of 16S rRNA fragments. The grey dashed line represents the fit of a linear regression (slope = 0.56, *P* < 0.001, adj-R^2 = 0.86).

Using Metaxa2, putative bacterial 16S rRNA fragments were extracted from the metagenomic datasets. Between 57, and 169,553 fragments were extracted from metagenomic libraries, which represented on average 0.0022% (std. error = 0.00027) of the total library size (Fig. 2.2A). There was a strong, positive relationship between the (log) library size of a metagenome and the (log) number of bacterial 16S rRNA fragments detected (Fig. 2.2B; slope = 0.56, $P < 0.001$, adj-R^2 = 0.86).

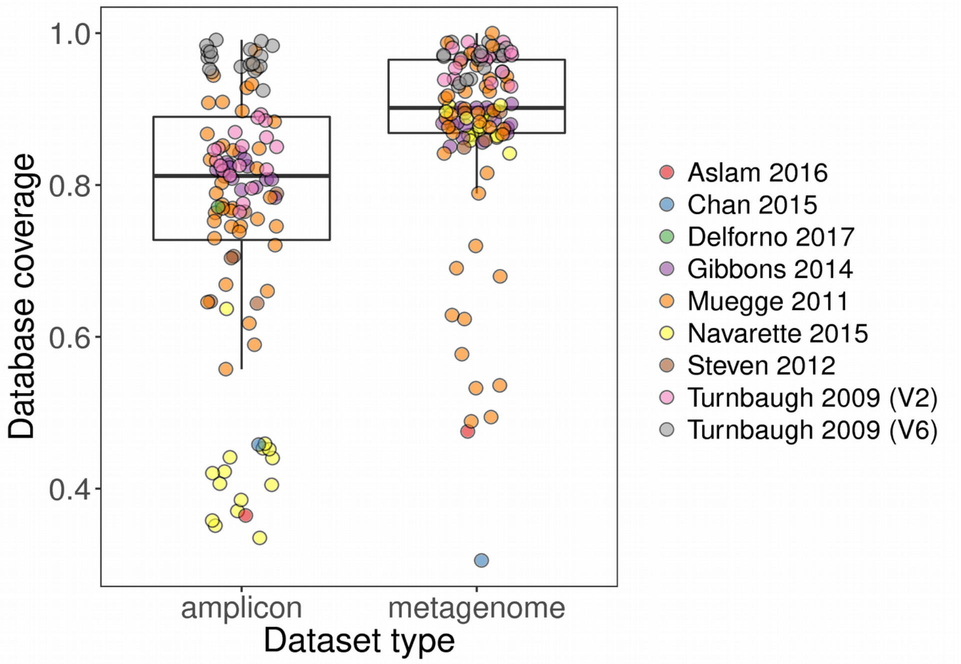


Figure 2.3 The proportion of metagenomic 16S rRNA fragments and amplicon sequences matching the non-redundant RDP database, at a similarity threshold of 97%.

Diversity Analyses

In order to compare the diversity obtained via metagenomic and amplicon sequencing approaches, 16S rRNA sequences and fragments were clustered against a non-redundant version of the RDP database. Mixed effects models revealed that overall, metagenomic datasets obtained significantly higher coverage in the reference database than amplicon datasets (Fig. 2.3; coefficient = 0.85, $z = 2.52$, $P < 0.05$).

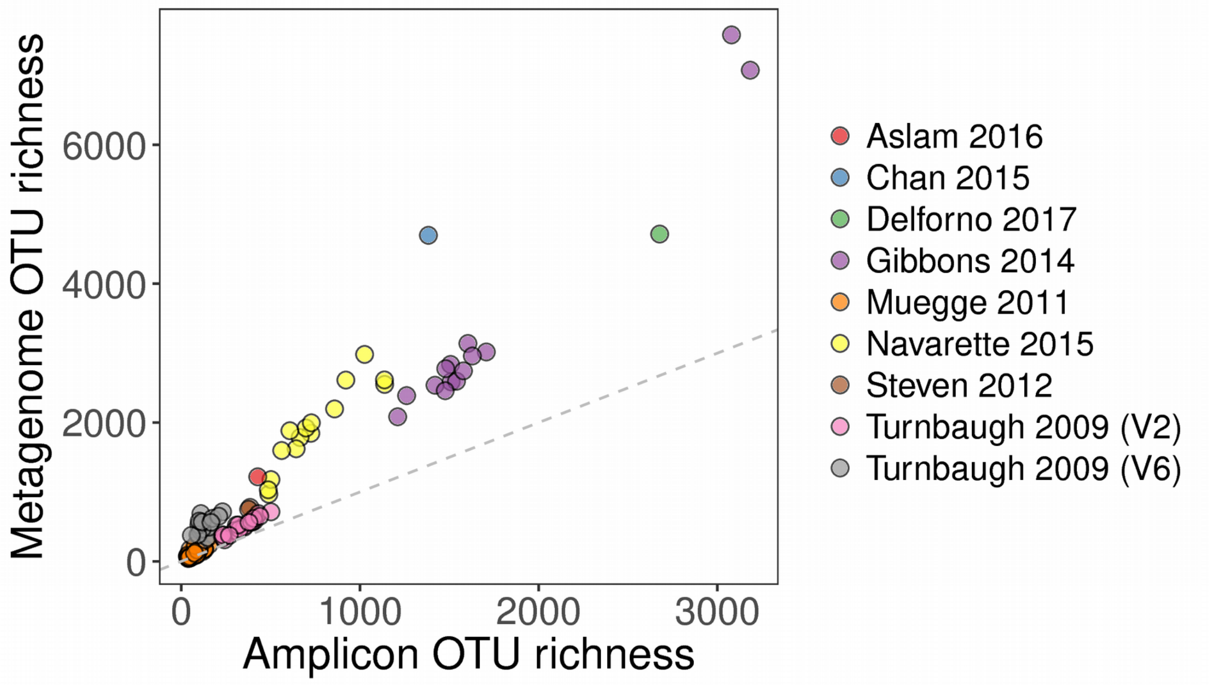


Figure 2.4 The relationship between OTU richness within amplicon and metagenomic datasets. The dashed line shows a 1:1 relationship, representing equal OTU richness in both datasets. OTU richness between the two dataset types were strongly and significantly correlated (Pearson's ρ = 0.97, $P < 0.001$).

Analysis of the OTU richness recovered by each method showed that alpha diversity patterns observed in amplicon datasets were strongly correlated with those in corresponding metagenomic dataset (Pearson's ρ = 0.97, $P < 0.001$). Furthermore, GLMM analysis showed that metagenomic sequence datasets recovered significantly more OTUs than amplicon datasets (Fig 2.4; coefficient = 0.77, *z*-value = 6.60, *P* < 0.001).

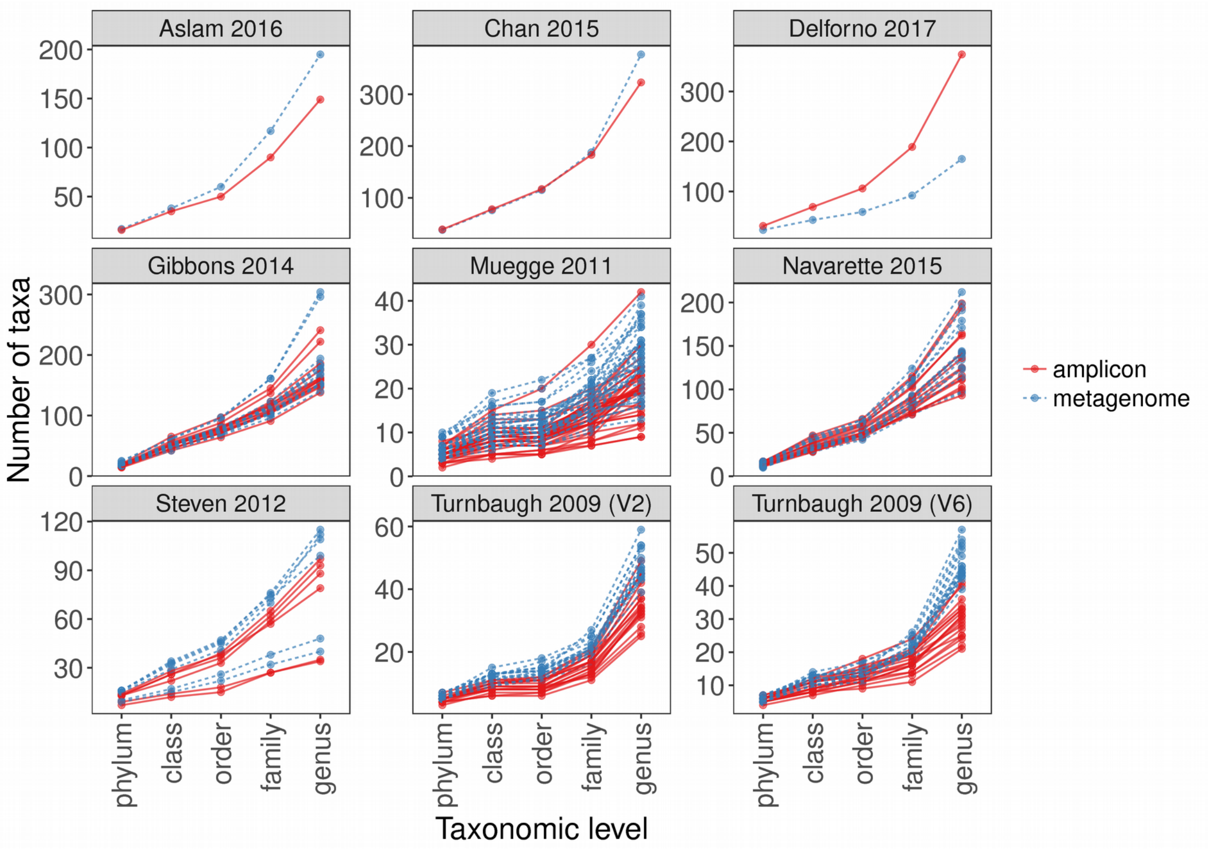


Figure 2.5 The taxonomic richness of each sample, at each taxonomic rank defined by the RDP classifier. Metagenomes tend to recover greater taxonomic richness, even at higher taxonomic levels, but especially at lower taxonomic levels (genus).

When taxonomic richness was analysed instead of OTU richness, similar patterns were observed. At all taxonomic ranks analysed, metagenomes tended to recover significantly greater taxonomic richness than corresponding amplicon data, and this effect was more evident at lower taxonomic ranks (e.g. genus) than at higher ranks (Fig. 2.5 and Table 2.2). As with OTUs, taxonomic alpha diversity was strongly correlated between amplicon and metagenomic datasets, for all taxonomic ranks (P < 0.001 for all taxonomic ranks; Fig. 2.6).

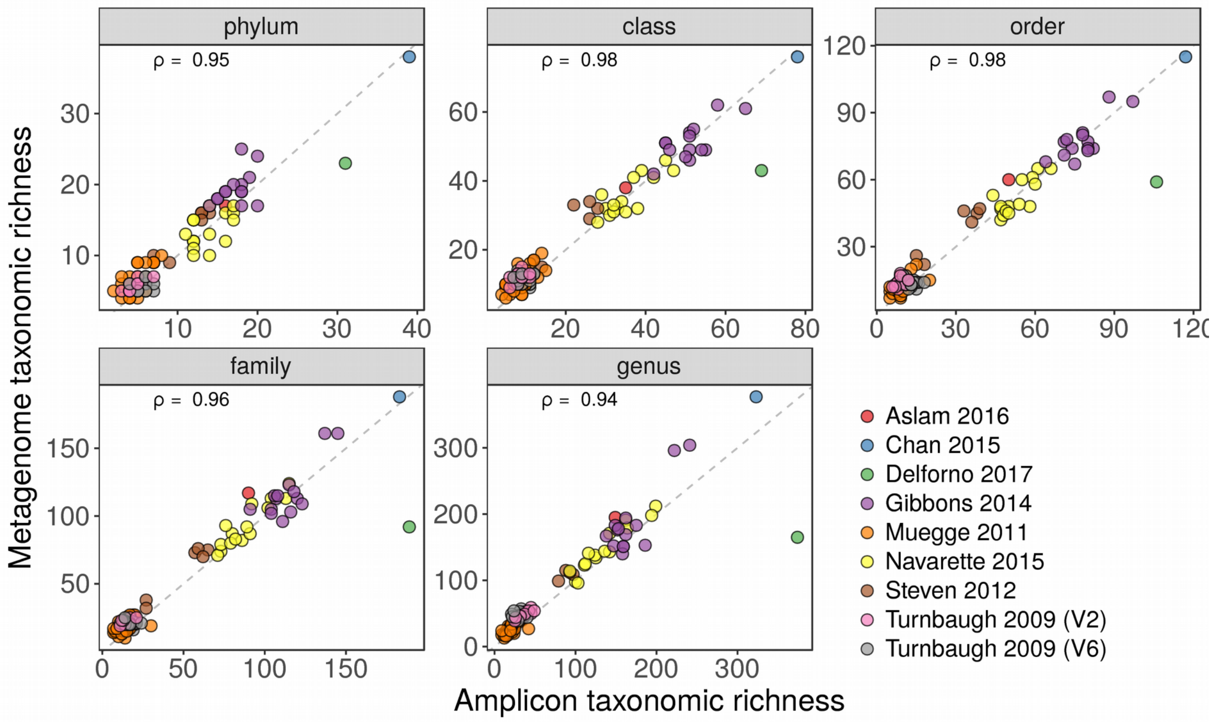


Figure 2.6 The relationship between the taxonomic richness from amplicon and metagenomic datasets. Panels represent each taxonomic level, and dotted lines represent a 1:1 relationship. Taxonomic richness was strongly and significantly correlated between datasets at every taxonomic level (*P* < 0.001 in all cases).

Table 2.2 Results from negative binomial GLMMs testing for differences in taxonomic richness between metagenomic and amplicon datasets. At all taxonomic ranks, metagenomes were found to be significantly richer, as indicated by the positive coefficient values.

Taxonomic rank	Estimated coefficient	Standard error	*Z*-value	*P*-value
Phylum	0.10	0.04	2.28	< 0.05
Class	0.08	0.03	2.74	< 0.01
Order	0.06	0.03	2.37	< 0.05
Family	0.09	0.02	4.15	< 0.001
Genus	0.20	0.02	8.06	< 0.001

Cost Analysis

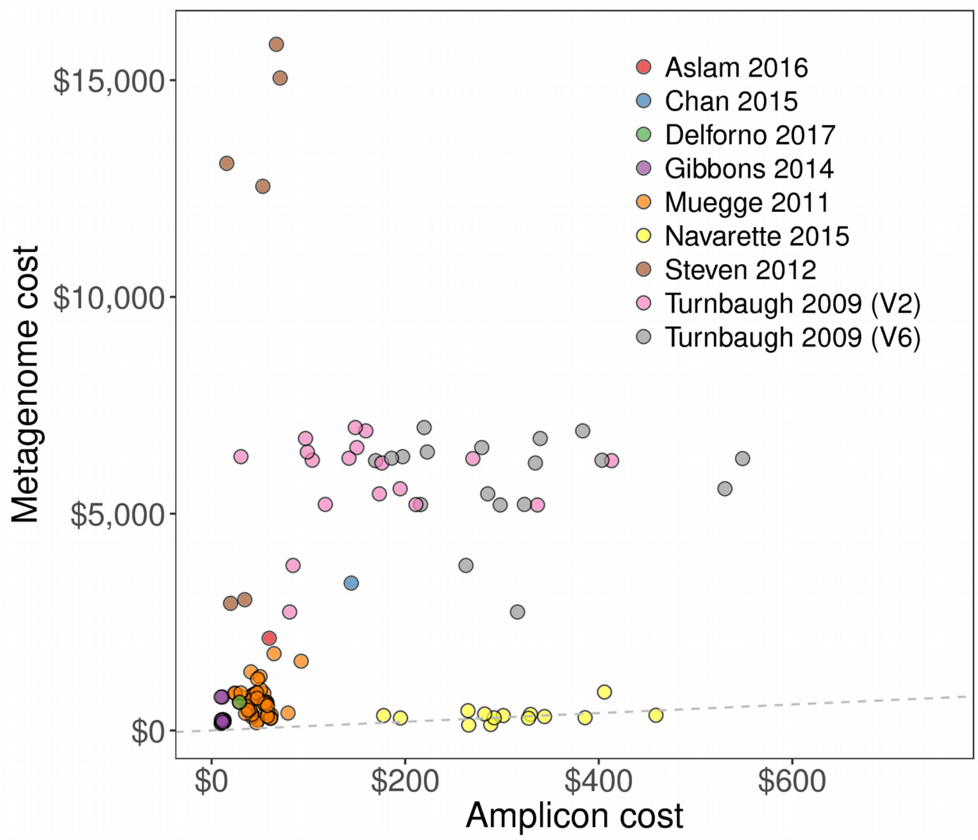


Figure 2.7 The estimated cost of each sample for both metagenomic and

amplicon sequencing datasets. The cost of the metagenomic datasets was calculated for the number of sequences needed to yield an equivalent number of 16S rRNA fragments as the corresponding amplicon dataset.

For each sample, I calculated the cost of sequencing a metagenome of sufficient size to produce an equivalent number of 16S rRNA fragments to the corresponding amplicon dataset. The difference in costs between amplicon and metagenomic approaches varied greatly between studies (Fig. 2.7), but metagenomes were on average 36.6 times more expensive than the equivalent amplicon cost. Only in the Navarette 2015 study was the total cost of metagenome sequencing not an order of magnitude greater than the cost of amplicon sequencing. This was due to expensive pyrosequencing being used for amplicon sequencing whilst the Illumina HiSeq, the cheapest of the three platforms, was used to sequence metagenomes, thus minimising the difference in cost.

Discussion

Within this study, I compared the ability of two common sequencing approaches, amplicon sequencing and metagenomic sequencing, to quantify α-diversity in bacterial communities. To do this, I assembled the largest dataset of paired metagenomic and amplicon samples to date consisting of over 1 billion sequences to date, and the first representing multiple biomes. I show that metagenomic sequencing recovers more diversity at both the OTU and taxonomic levels when differences in library size are accounted for. However, I also found that the costs associated with the two sequencing approaches can be vastly different. In all studies aside from one, the cost of producing a metagenome that yields an equivalent number of 16S sequences to the corresponding amplicon dataset was an order of magnitude higher than the cost of the amplicon dataset. The results clearly demonstrate the advantages of metagenomic approaches in terms of quantifying more taxa (and OTUs), but also highlight that this extra diversity comes at a large financial cost, thereby limiting the usefulness of metagenomics in multiple sample studies (Neufeld, 2017).

Against expectation, metagenomic sequencing recovered more taxonomic diversity than amplicon sequencing at all taxonomic levels including. Amplicon sequencing is known to exclude certain taxa due to suboptimal primer coverage (Hong *et al*., 2009). Primers are typically tested for coverage properties against databases of full length marker genes (Klindworth *et al*., 2013; Hugerth *et al*., 2014). However, this has the potential to lead to a cycle in which novel taxa may be missed by primers, and therefore are less likely to be discovered and added to databases, resulting in reduced discovery rates of novel organisms. In contrast, I have shown that metagenomics facilitates the discovery of greater taxonomic richness, even at more basal taxonomic ranks. This finding is in contrast to recent findings by Tessler *et al*. (2017), who found that amplicon sequencing recovered more families and phyla than did metagenomic sequencing. However, this is explained by the fact that the authors did not perform any normalisation between amplicon and metagenome library sizes. As I have shown, 16S rRNA fragments almost always represent < 1% of the total metagenome library size, meaning that in most studies, the number of metagenomic 16S sequences is fewer than the corresponding amplicon dataset. Therefore, amplicon datasets may recover more taxonomic diversity in an absolute sense, but once normalised, metagenomes recover more.

By mapping both metagenomic 16S rRNA fragments and amplicon sequences to a database of near full length 16S rRNA sequences, I was also able to compare the two sequencing approaches at the OTU level, unlike previous comparative studies (e.g. Tessler *et al*., 2017). However, the fact that metagenomic 16S rRNA fragments will likely map to different regions of the 16S gene represents something of a caveat to our study. This is because different regions of the 16S rRNA gene show different rates of evolution, and therefore differ in their ability to resolve closely related taxa. One approach to deal with this is to choose a region for which many metagenomic fragments overlap, as in Guo *et al*. (2016). However, this does not make good use of the data, as fragments that do not overlap the chosen region are discarded, even though they may still be informative. An alternative approach might be to apply a gene-region-specific sequence similarity threshold. This would allow a higher threshold to be used for fragments that map to well conserved regions, whilst a lower threshold could be used for fragments that map to hyper-variable regions. However, such an approach would not be trivial. A database of full length (or near full length) marker genes would be required to ensure that marker gene fragments are mapped to the correct gene region. Furthermore, robust evolutionary models would be required to ensure that hyper-variable and conserved gene regions are correctly characterised, and that appropriate sequence similarity thresholds are calculated. However, the use of locus specific sequence similarity thresholds could be a robust alternative to using static sequence similarity thresholds.

Whilst metagenomics is able to recover more diversity than an equivalently sized amplicon sequence dataset, the cost of metagenomics is still prohibitively high for its application as a method of surveying microbial diversity. Here, I illustrated that in all but one of the studies used, the cost to sequence a sufficiently a metagenome of equal depth to an amplicon, would be at least an order of magnitude higher than amplicon sequencing. In all of the studies used, Bacteria were the focal taxa, and it is known that they are reasonably abundant in all of the studied biomes. However, if a less abundant group of microbes were of interest (e.g. Archaea of Fungi), the need for greater sequencing depth would be exacerbated, increasing the relative cost even more. In contrast, amplicon sequencing carries the advantage that organisms which are very rare in the environment can still be targeted (Lansdown *et al*., 2016). Therefore, whilst metagenomics shows great promise as a method for recovering hitherto unknown microbial diversity (Neufeld, 2017), it is currently not a cost effective means of doing so, particularly in biogeographic studies where a large number of samples are often required to achieve sufficient spatial or temporal replication to test a given hypothesis.

Conclusions

By comparing the diversity within amplicon and metagenomic datasets, I have shown that metagenomic sequencing is able to quantify more diversity at both the taxonomic and OTU level. Patterns of alpha diversity were highly correlated between datasets, indicating that metagenomes are able to successfully mirror the ecological patterns observed within amplicon datasets. However, in order to extract an adequate number of 16S rRNA fragments from a metagenome, far larger sequencing depth is required, meaning that the difference in cost between metagenomic and amplicon sequence datasets often spans an order of magnitude. The difference in cost means that metagenomic sequencing remains prohibitively expensive as a method of surveying microbial diversity, particularly when many samples are required. Therefore, until the price of metagenomic sequencing decreases sufficiently, amplicon sequencing will remain the most frequent method of surveying microbial diversity.

References

Aird D, Ross MG, Chen WS, *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, **12**, R18.

Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology*, **71**, 7724–7736.

Aslam SN, Dumbrell AJ, Sabir JS, *et al.* (2016) Soil compartment is a major determinant of the impact of simulated rainfall on desert microbiota. *Environmental Microbiology*, **18**, 5048–5062.

Bellemain E, Carlsen T, Brochmann C, Coissac E, Taberlet P, Kauserud H (2010) ITS as an environmental DNA barcode for fungi: an in silico approach reveals potential PCR biases. *BMC Microbiology*, **10**, 189.

Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, Nilsson RH (2015) METAXA2: Improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Molecular Ecology Resources*, **15**, 1403–1414.

Bergmann GT, Bates ST, Eilers KG, *et al.* (2011) The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biology and Biochemistry*, **43**, 1450-1455.

Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology and Evolution*, **27**, 233–243.

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, **24**, 127–135.

Bokulich NA, Subramanian S, Faith JJ, *et al.* (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods*, **10**, 57–59.

Chan CS, Chan KG, Tay YL, Chua YH, Goh KM (2015) Diversity of thermophiles in a Malaysian hot spring determined using 16S rRNA and shotgun metagenome sequencing. *Frontiers in Microbiology*, **6**, 177.

Cole JR, Wang Q, Fish JA, *et al.* (2013) Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, **42**, D633–D642.

Cotton TEA, Dumbrell AJ, Helgason T (2014) What Goes in Must Come out: Testing for Biases in Molecular Analysis of Arbuscular Mycorrhizal Fungal Communities. *PloS ONE*, **9**, e109234.

Delforno TP, Lacerda Júnior GV, Noronha MF, Sakamoto IK, Varesche MBA, Oliveira VM (2017) Microbial diversity of a full-scale UASB reactor applied to poultry slaughterhouse wastewater treatment: integration of 16S rRNA gene amplicon and shotgun metagenomic sequencing. *MicrobiologyOpen*, **6**, e00443.

Dinsdale EEA, Edwards RAR, Hall D, *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature*, **452**, 629–632.

Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH (2010) Relative roles of niche and neutral processes in structuring a soil microbial community. *The ISME Journal*, **4**, 337–345.

Fayle TM, Scholtz O, Dumbrell AJ, Russell S, Segar ST, Eggleton P (2015) Detection of mitochondrial COII DNA sequences in ant guts as a method for assessing termite predation by ants. *PLoS ONE*, **10**, e0122533.

Fonseca VG, Nichols B, Lallias D, Quince C, Carvalho GR, Power DM, Creer S (2012) Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. *Nucleic Acids Research*, **40**, e66.

Fredriksson NJ, Hermansson M, Wilén BM, Sternberg C, Givskov M (2013) The Choice of PCR Primers Has Great Impact on Assessments of Bacterial Community Diversity and Dynamics in a Wastewater Treatment Plant. *PLoS ONE*, **8**, e76431.

Gibbons SM, Jones E, Bearquiver A, *et al.* (2014) Human and environmental impacts on river sediment microbial communities. *PloS ONE*, **9**, e97435.

Gihring TM, Green SJ, Schadt CW (2012) Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. *Environmental Microbiology*, **14**, 285–290.

Guo J, Cole JR, Zhang Q, Brown CT, Tiedje JM (2016) Microbial Community Analysis with Ribosomal Gene Fragments from Shotgun Metagenomes. *Applied and Environmental Microbiology*, **82**, 157–166.

Hong S, Bunge J, Leslin C, Jeon S, Epstein SS (2009) Polymerase chain reaction primers miss half of rRNA microbial diversity. *The ISME Journal*, **3**, 1365–1373.

Huber JA, Morrison HG, Huse SM, Neal PR, Sogin ML, Mark Welch DB (2009) Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. *Environmental Microbiology*, **11**, 1292–1302.

Hugerth LW, Muller EEL, Hu YOO, *et al.* (2014) Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PloS ONE*, **9**, e95567.

Jeon S, Bunge J, Leslin C, Stoeck T, Hong S, Epstein SS (2008) Environmental rRNA inventories miss over half of protistan diversity. *BMC Microbiology*, **8**, 222.

Joshi N, Fass J (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. *Available at https://github.com/najoshi/sickle*, 2011.

Jumpponen A (2007) Soil fungal communities underneath willow canopies on a primary successional glacier forefront: rDNA sequence results can be affected by primer selection and chimeric data. *Microbial Ecology*, **53**, 233–246.

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, **41**, e1.

Lansdown K, McKew BA, Whitby C, *et al.* (2016) Importance and controls of anaerobic ammonium oxidation influenced by riverbed geology. *Nature Geoscience*, **9**, 357–360.

Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database Collaboration (2010) The Sequence Read Archive. *Nucleic Acids Research*, **39**, D19-D21.

Locey KJ, Lennon JT (2016) Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences USA*, **113**, 5970–5975.

Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, **30**, 434–439.

Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD (2012) PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics*, **13**, 1–7.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J (2008) The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Muegge BD, Kuczynski J, Knights D, *et al.* (2011) Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans. *Science*, **332**, 970–974.

Navarrete AA, Tsai SM, Mendes LW, *et al.* (2015) Soil microbiome responses to the short-term effects of Amazonian deforestation. *Molecular Ecology,* **24**, 2433-2448.

Neufeld JD (2017) Migrating SSU rRNA gene surveys to the metagenomics era. *Environmental Microbiology Reports*, **9**, 23–24.

O’Hara RB, Kotze DJ (2010) Do not log-transform count data. *Methods in Ecology and Evolution*, **1**, 118–122.

Pinto AJ, Raskin L (2012) PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PloS ONE*, **7**, e43093.

Quail M, Smith ME, Coupland P, *et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.

Roesch LFW, Fulthorpe RR, Riva A, *et al.* (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME Journal*, **1**, 283–290.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2409v1.

Schirmer M, Ijaz UZ, D’Amore R, Hall N, Sloan WT, Quince C (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, **43**, e37.

Schloss PD, Girard R, Martin T, Edwards J, Thrash JC (2016) The status of the microbial census: an update. *Microbiology and Molecular Biology Reviews*, **68**, 686-691.

Schoch CL, Seifert KA, Huhndorf S, *et al* (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences USA*, **109**, 6241–6246.

Sharpton TJ (2014) An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, **5**, 209.

Sogin ML, Morrison HG, Huber JA, *et al.* (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere.” *Proceedings of the National Academy of Sciences USA*, **103**, 12115–12120.

Staley JT, Konopka A (1985) Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. *Annual Review of Microbiology*, **39**, 321–346.

Steven B, Gallegos-Graves LV, Starkenburg SR, Chain PS, Kuske CR (2012) Targeted and shotgun metagenomic approaches provide different descriptions of dryland soil microbial communities in a manipulated field study. *Environmental Microbiology Reports*, **4**, 248–256.

Tedersoo L, Bahram M, Polme S, *et al.* (2014) Global diversity and geography of soil fungi. *Science*, **346**, 1256688.

Tessler M, Neumann JS, Afshinnekoo E, *et al.* (2017) Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Scientific Reports*, **7**, 6589.

Turnbaugh PJ, Hamady M, Yatsunenko T, *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naiive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–5267.

Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, **30**, 614–620.

Chapter 3

The Spatial Scaling of β-Diversity is Context-Dependent for the Majority of Global Biodiversity

Acknowledgements: Graham JC Underwood, Terry J McGenity, and Alex J Dumbrell

Abstract

Ecological communities closer together in space and time, are generally more similar than those further apart, as defined by the distance-decay (d-d) of similarity relationship. Historically, microorganisms were assumed to defy this relationship due to their capacity for long distance, passive dispersal, and high population densities. Yet, recent studies have recorded highly variable d-d relationships in a range of microbial communities from disparate environments, using very different methods. The range of biological contexts incorporated by these studies could explain the differing distance-decay relationships reported as the dispersal of microorganisms may vary between different study systems, or spatial scales. Furthermore, methodological differences between studies will differ in their ability to detect rare species, thereby leading to contrasting estimates of compositional similarity between communities. Therefore, I sought to understand whether the variability in microbial d-d relationships is caused by different study methodologies, or biological contexts. To do this, I conducted an exhaustive meta-analysis and

gathered data on 287 microbial d-d relationships. Given that most studies statistically test for d-d relationships using the Mantel correlation test, I used the Mantel correlation coefficient as a measure of effect size. I found that d-d relationships were weakly but significantly related to measures of community coverage, whilst different community quantification methods (e.g. community fingerprinting, high-throughput sequencing, morphological) only effected statistically-significant d-d relationships. The use of phylogenetic community similarity indices resulted in significantly weaker d-d relationships than compositional similarity metrics (e.g. Jaccard’s or Bray-Curtis index). Distance-decay relationships were significantly weaker in soils than other study systems, but significantly stronger in host-associated systems, potentially reflecting the ecological properties of the host taxon. The strength of the d-d relationships was also positively related to the spatial scale of the study but, against expectation, did not vary between different study taxa. I conclude that the microbial d-d relationship is dependent on biological context, but that methodological choices by the researcher can also strongly influence the strength of this relationship. I provide suggestions for selecting methods that will minimise methodological noise, and enhance ecological signal.

Introduction

The distance-decay (d-d) of community similarity is one of the most commonly studied relationships in macroecology (Nekola & White 1999; Condit *et al.* 2002; Soininen *et al.* 2007). The relationship quantifies how the similarity of community composition decays with increasing geographic distance between communities, such that communities close together contain more similar species assemblages than those further apart. Distance-decay relationships are able to inform us about the dispersal abilities of organisms present in the community, the connectivity of communities, as well as the spatial configuration of the environment. Consequently, the relationship is of great importance in understanding the spatial configuration of global biodiversity, with potential implications for conservation efforts (Nekola & White, 1999). Accordingly, the relationship has been well studied across a wide range of organisms with varying dispersal abilities and ecological properties, revealing distance-decay relationships over a range of spatial scales.

One group of organisms previously thought to defy the d-d relationship are microorganisms. One of the earliest hypotheses about the potential for microbial distance-decay relationships was formulated by Dutch microbiologist, Baas-Becking. Within this hypothesis, Baas Becking postulated that “Everything is everywhere but, the environment selects” (Baas Becking 1934). The rationale behind this hypothesis was that microorganisms should be efficient dispersers, as their small size may facilitate long distance, passive dispersal (Wilkinson *et al*., 2012). Additionally, the high population densities often observed in environmental microbial communities, might facilitate dispersal through “mass effects”, whereby organisms disperse from areas of high density to less favourable habitat (Shmida & Wilson, 1985). Therefore, “Everything is everywhere” suggests that microbial distance-decay relationships are exclusively the result of spatially structured environmental factors (Finlay & Fenchel 2004, Hanson *et al.,* 2012). This interpretation of the microbial d-d relationship is compatible with niche theory, which posits that communities are assembled form as the result of interactions between species with the environment (Holt, 2009). Therefore, spatial processes play a lesser role in the formation of microbial d-d relationships.

However, modern molecular evidence suggests that the causes of d-d relationships are considerably more complex than “Everything is everywhere” (Hanson *et al.,* 2012). The rapid development of molecular methods to study microbial communities, coupled with the provocative nature of “Everything is everywhere”, has resulted in an explosion of studies testing the distance-decay relationship in microbial communities. These studies have yielded mixed results. A number of studies have found no correlation between microbial community composition and distance (Hazard *et al.* 2013; Kivlin *et al.* 2014), showing that communities separated by large geographic distances can be equally similar to those separated by small distances. However, many studies have reported relationships, varying in steepness, between microbial community composition and geographic distance for a range of spatial scales and taxa (Dumbrell *et al.* 2010; Martiny *et al.* 2011; Barreto *et al.* 2014), even when spatially autocorrelated environmental gradients have been accounted for (e.g. Green *et al*., 2004). These results suggests that microbial communities may be structured by spatial processes, and not solely by the environment, in disagreement with “Everything is everywhere” and niche theory. This finding is concordant with neutral theory, suggesting that stochastic processes, such as dispersal, can contribute to the composition of a community (Hubbell, 2001). The ability of ecologically very different processes (niche and neutral) to generate d-d relationships suggests that, variability in this relationship may be related to organisms’ dispersal abilities, connectivity and/or spatial distance between communities, and environmental heterogeneity. Therefore, biological context may explain the disparity in reported microbial d-d relationships.

Here, context could be considered to be the organisms studied (e.g. Bacteria, Archaea, Fungi, Protists etc.), the study system in question (soil, freshwater, extreme systems), or the spatial scale of the study. Distance-decay relationships may vary by taxonomic groups if dispersal is a trait-based process, for example varying cell sizes (Wilkinson *et al.* 2012; Soininen *et al.* 2013). Different study systems may also influence the rate of the d-d relationship as previously reported (Soininen *et al.* 2007). This may be because environments differ in connectivity, for example, host associated communities may be poorly connected due to the restricted dispersal/range size of the host, and therefore will exhibit steeper distance-decay relationships. Additionally, environments will differ according to the physicochemical gradients they are able to support. Stable, undisturbed habitats such as soils have been shown to support considerable pH gradients over relatively short distances (e.g. Dumbrell *et al*. 2010), resulting in stronger distance-decay relationships. In contrast, well mixed surface waters may support far more diffuse gradients, resulting in shallower distance-decay curves. Finally, the spatial extent of a study could influence the observed d-d relationship. Larger spatial scales may result in a decrease in dispersal between communities, and greater environmental heterogeneity, both of which should result in steeper d-d relationships (Martiny *et al.* 2011). In contrast, studies covering small spatia extents will likely sample more similar communities, that may be better connected by dispersal, leading to a shallow d-d relationship.

On the other hand, methodological differences between studies may contribute to variability in microbial distance-decay relationships. From serially sequencing clone libraries, through community fingerprinting methods, and most recently high-throughput sequencing, previous research into the microbial d-d relationship is based upon a plethora of methods with varying degrees of taxonomic resolution and community coverage (Muyzer 1999;

Glenn 2011). These methodologies differ markedly in their ability to quantify microbial communities, and in particular the rare species that form the majority of a microbial community. Methods that are only able to quantify the most common (and widespread) species (such as morphological, or community fingerprinting methods) are likely to miss the rare, restricted taxa. The result of this is that communities will appear artificially similar, resulting in a weaker distance-decay relationship. In contrast, methods that adequately quantify the “rare biosphere”, such as high-throughput sequencing, will be able to better detect the species that differ between communities, and therefore more accurately quantify the community similarity, resulting in stronger d-d relationships. In addition to the varying methods used to quantify microbial communities, there are now an array of indices available to quantify the (dis)similarity between microbial communities, including qualitative (based on presence/absence of species), quantitative (based on composition and abundance of species), and phylogenetic (based on relatedness of communities). Such indices have different properties in terms of how they weight rare or common species, and in how they are influenced by sample sizes or species richness (Baselga 2012; Beck *et al.* 2013), as well as what they quantify (e.g. phylogenetic similarity versus compositional similarity). The use of different indices could contribute to the strength of a distance-decay relationship. For example, phylogenetic indices may result in weaker distance-decay relationships because communities can be phylogenetically closely related, but may differ at the operational taxonomic unit (OTU) level (e.g. Bryant *et al*., 2008).

Given the variability in microbial d-d relationships reported in the literature, I sought to understand whether methodological or contextual differences between studies may influence reported d-d relationships in microbial communities. To do this, I conducted a meta-analysis to synthesise available data on the microbial distance-decay relationship, and test whether factors relating to methodological or contextual aspects of each study influence this relationship. Specifically, I test the following hypotheses:

- H1: Bacteria will show weaker d-d relationships than other microbial groups due to their smaller size and higher population densities.
- H2: Soils and host-associated study systems will show stronger d-d relationships than other systems due to their ability to maintain steep physicochemical gradients, or limited range size of the host taxon, respectively.
- H3: The spatial extent of a study will be positively related to the strength of a d-d relationship, as larger scale studies will incorporate greater environmental heterogeneity, and lower dispersal between the most distant communities.
- H4: Higher resolution community quantification methods, such as high-throughput sequencing, will yield stronger d-d relationships due to their ability to quantify rare taxa and resolve closely related taxa, and thus more accurately quantify community (dis)similarity.

- H5: Sampling depth (e.g. number of sequences, or number of individuals counted) will be positively related to the strength of d-d relationships, for the same reason as in H4.
- H6: The strength of d-d relationships will vary between similarity indices, and in phylogenetic methods will result in weaker d-d relationships than compositional metrics.

Methods

Meta-Analysis

In order to test the effects of ecological context and methodology on the microbial d-d relationship, I first conducted a systematic literature search using the Web of Science search portal. To do this, I designed five different search terms in order to maximise the size of the resulting dataset, whilst minimising irrelevant (e.g. studies of “macroorganisms”) studies (Table 3.1). All five searches were conducted on 08/06/2017, and all search results published between 1900-2017 were retained. I downloaded all search results from Web of Science and used the “metagear” package (version 0.4) in R (version 3.4.1) to manually screen abstracts for suitability for inclusion in our study (R Development Core Team 2016; Lajeunesse 2016). Suitable studies were defined as those that indicated a test of the relationship between spatial or geographic distance.

Table 3.1 Details of the five Web of Science search terms and, the number of hits. A Web of Science search history file is provided in the Supplementary Material.

Search	Search terms	Search results
1	TS = (biogeograph*) AND TS = (bacteria* OR archaea* OR microb* OR microorganism*)	1,872
2	TS = (macroecolog*) AND TS = (bacteria* OR archaea* OR microb* OR microorganism*)	85
3	TS = ("everything is everywhere") AND TS = (bacteria* OR archaea* OR microb* OR microorganism*)	53
4	TS = ("geographic distance") AND TS = (bacteria* OR archaea* OR microb* OR microorganism*)	133
5	TS = ("distance decay") AND TS = (bacteria* OR archaea* OR microb* OR microorganism*)	107

* is a wildcard to allow searches to match multiple terms, e.g. microb* could match “microbiome”, “microbial”, and “microbe”

I focussed on studies that had tested the distance-decay relationship using the Mantel correlation test, as this is the most common method of testing this relationship in microbial ecology (Ramette, 2007; Lisboa *et al*., 2014), and provides an easily interpretable effect size measure (Harrison, 2010). The Mantel test is used to test for correlation between two distance matrices (i.e. community dissimilarity and geographic distance). Mantel correlation coefficients vary between -1 and 1, with values of 1 indicating strong positive correlation, 0 indicates no/weak correlation, and -1 shows strong negative correlation. To standardise correlation coefficients between studies that had used similarity matrices, rather than dissimilarity matrices, I multiplied the former by -1, so that all correlation coefficients reflect the correlation between dissimilarity and geographic distance. For clarity, here a Mantel correlation coefficient of 1 indicates a strong d-d relationship. Partial Mantel statistics (which are able to test for correlation between two matrices whilst controlling for a third) were excluded as they may be heavily influenced by which other variables are included in the test, and are therefore not easily comparable between studies. In order to test our hypotheses, I recorded several variables relating to the ecological context of each study, as well as the methods used (Box 3.1).

Box 3.1 Details of the explanatory variables extracted from each study.

Community characterisation method

This refers to the method used to quantify the species present in their sample and their abundances (if applicable). Each d-d relationship was categorised into either high-throughput sequencing (HTS; Pyrosequencing, Illumina, Ion Torrent, Pac-Bio), community fingerprinting (ARISA, TRFLP, DGGE, PhyloChip), or other (Sanger sequencing, morphological identification).

Sequencing depth

This refers to the sequencing depth in sequencing based studies, or number of individuals counted in morphological based studies. For sequencing studies, we recorded the number of sequences after rarefaction, or if this was not given, the average number of sequences per sample. As it is hard to quantify the resolution of fingerprinting approaches, we recorded these as NA and excluded them from analyses involving sequencing depth.

Sampling effort

This variable represents the number of individual communities/samples used to formulate the d-d relationship.

Dissimilarity index

We recorded the dissimilarity index from which each d-d relationship was calculated. After these had been recorded, we categorised them as abundance based (Bray-Curtis, Horn-Morisita, Euclidean, Hellinger, Theta), binary (Jaccard, Raup-Crick, Sørensen, Simpson, βsim), or phylogenetic (Unifrac, Rao, β-mean nearest taxon distance, β-mean pairwise distance).

Study taxon

We categorised d-d relationships into broad taxonomic categories (Archaea, Bacteria, Eukarya, Fungi). If a d-d relationship was based on multiple taxa, then an appropriate category was added as necessary (I.e. bacteria + archaea).

Scale

We recorded scale as the maximum distance separating communities (in km). If this was not stated in text or provided in supplementary material (e.g. in a geographic distance matrix), it was calculated from given geographic coordinates, or estimated from the d-d graph itself or from scaled maps, if no coordinates were provided.

Biome

We categorised d-d relationships based on their biome (agriculture, air, aquifer, indoor, coral, desert, dune, flower, forest, grassland, ice, lake, marsh, mine, ocean, paddy, river, sediment, sewer, sponge), reflecting the type of environment the communities occupied.

Environmental material

This variable represents the type of material that the sampled communities occupied. We categorised d-d relationships as air, host, sediment, soil, or water.

P-value

As an additional comparison, we also recorded *P*-values for d-d relationships where possible. We recorded unadjusted *P*-values, and here use a global alpha value of 0.05 for simplicity, regardless of multiple tests conducted by each study.

Statistical Analyses

In order to determine whether d-d relationships varied between categoric variables (as in H1, H2, H4, and H6), I used ANOVA tests. To test hypotheses 3 and 5, I used linear regressions. I first log transformed both study scale and sequencing depth as these variables spanned several orders of magnitude.

Results

The Web of Science searches resulted in 2,250 search hits (Table 3.1). After removing duplicate hits (i.e. studies that appeared in multiple searches), this number decreased to 2,031 hits. Manual screening of the abstracts yielded 547 studies that were deemed to be potentially suitable for use in this analysis. A total of 287 Mantel correlation coefficients were successfully obtained from 108 studies represented in 33 journals (Figs. 3.1). Of the 439 studies that were unsuitable for inclusion within this analysis, most had not tested for correlation between geographic distance and community (dis)similarity (although the abstract still contained the search terms), whilst others had used different methods (e.g. multilocus sequence typing on individual species, or spatial eigenvector analysis). Reported Mantel correlation coefficients ranged from -0.24 to 0.95, with a mean of 0.27 (std. error = 0.014).

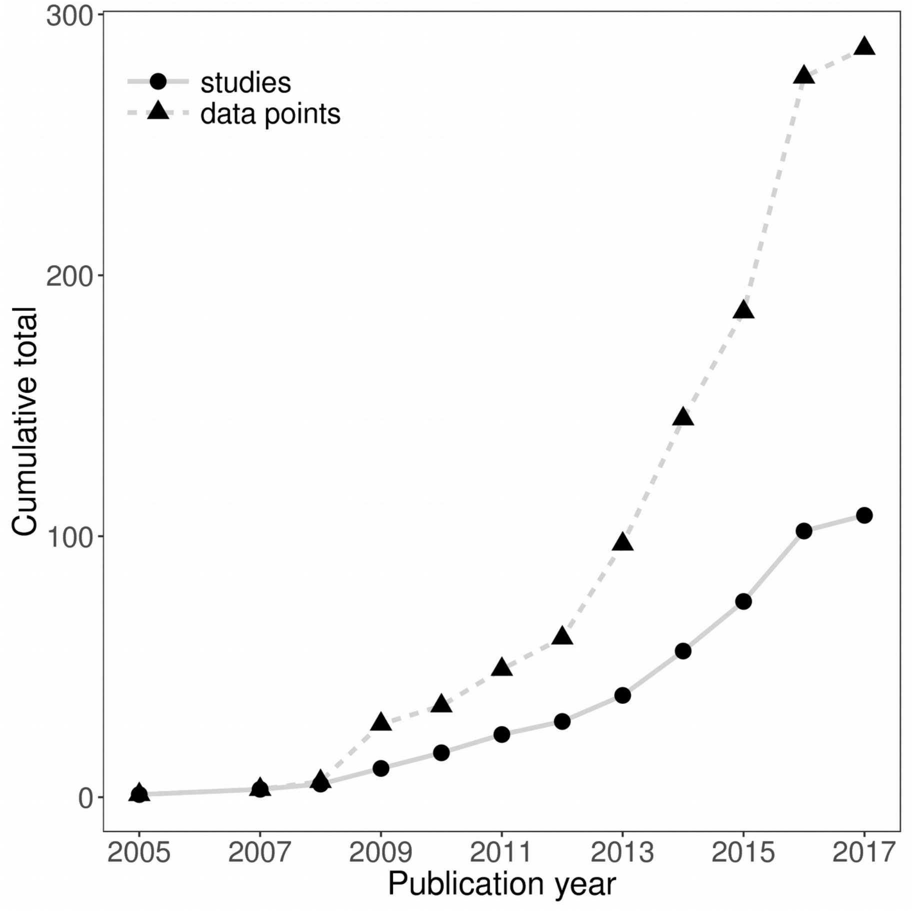


Figure 3.1 The cumulative number of distance-decay (d-d) relationships and publications included in this study, through time. Studies referes to the number of publications in which microbial d-d relationships are tested, whilst data points refers to the number of d-d relationships reported.

Influence of Biological Context on the Distance-Decay Relationship

In order to determine whether different ecological contexts can influence the strength of d-d relationships, the influence of ecological factors including study taxa, study system, and spatial scale were tested. Within the dataset, the most commonly studied taxa were Bacteria, followed by Fungi, micro-Eukaryotes, and Archaea. No significant difference was found in the Mantel coefficients associated with each taxa ($F_{5,\ 281}$ = 1.39, *P* = 0.23), in disagreement with H1. Examining only statistically significant Mantel coefficients revealed marginally significant differences between taxa ($F_{5,\ 172}$ = 2.51, *P* < 0.05) with studies incorporating both bacteria and fungi (*n* = 3) simultaneously, being significantly lower than studies on Archaea (Tukey HSD; *P* < 0.05).

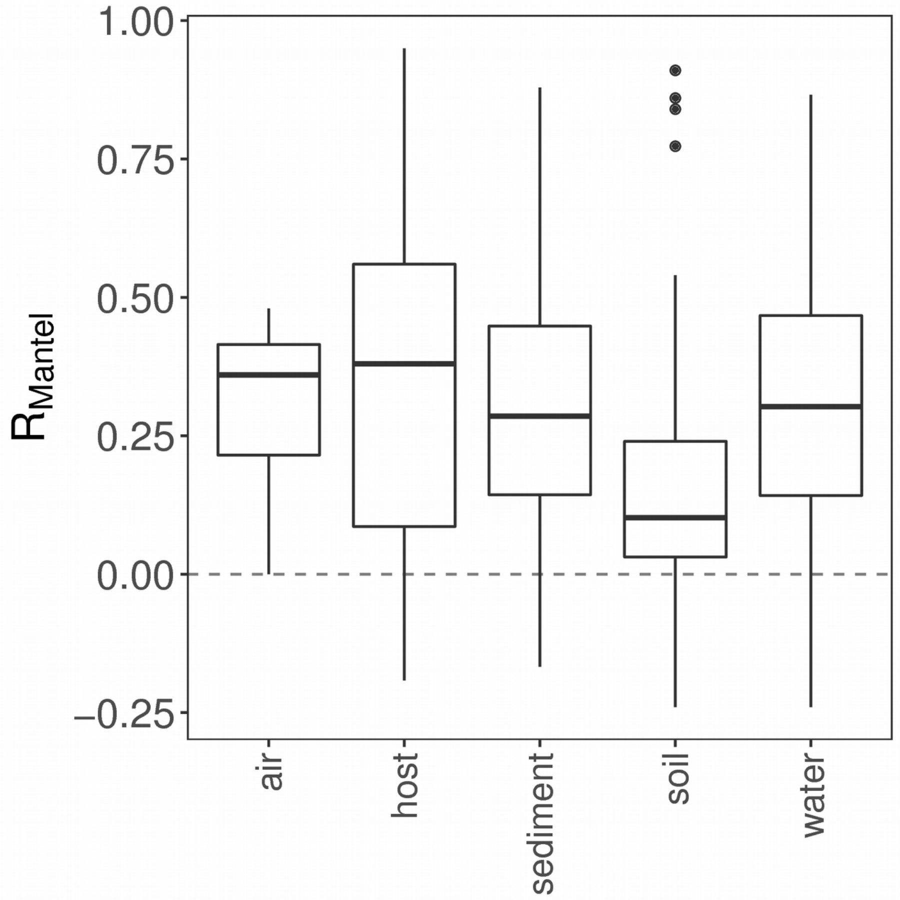


Figure 3.2 Mantel correlation coefficients from microbial communities sampled from different environmental materials. Larger Mantel coefficients indicate stronger distance-decay relationships.

Of the 20 different biomes recorded, 11 had fewer than three d-d relationships, and these biomes were excluded from biome analyses. The most frequently studied biomes were grasslands (*n* = 62), forest (*n* = 57), and lakes (*n* = 44). Mantel coefficients differed significantly between biomes ($F_{8, 262}$ = 8.80, $P < 0.001$), in partial agreement with H2. Specifically, sponge associated communities displayed higher coefficients than all other biomes (Tukey HSD; $P < 0.05$ in all cases), and grassland communities had lower coefficients than most other biomes (Forest, lake, ocean, river, sediment, and sponge. Tukey HSD; $P < 0.05$ in all cases). Furthermore, the different types of environmental materials sampled showed significant differences in Mantel coefficients (Fig. 3.2; $F_{4, 280}$ = 7.35, $P < 0.001$). Surprisingly, soils showed significantly lower coefficients than host-associated, sediment, or water d-d coefficients (Tukey HSD; $P < 0.01$ in all cases), in contrast with H2.

Finally, concordant with H3, there was a significant, positive relationship between the (log) spatial scale and the Mantel coefficient (slope = 0.016, $P <$ 0.001, adj-R^2 = 0.12), showing that studies with larger spatial extents tend to find stronger correlations between community dissimilarity and geographic distance (Fig. 3.3). This relationship held when only significant Mantel coefficients were examined, and after accounting for sampling effort (slope = 0.016, $P < 0.001$, adj-R^2 = 0.13). Sampling effort was not correlated with spatial scale (Pearson's ρ = 0.03, P = 0.64), showing that studies that incorporate larger spatial scales, do not necessarily incorporate more samples.

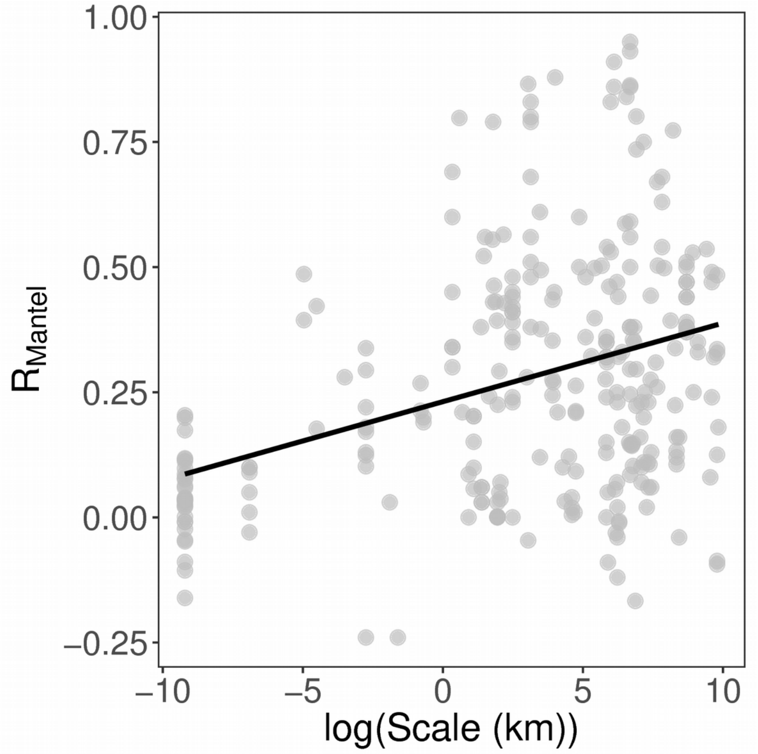


Figure 3.3 The relationship between Mantel correlation coefficients and the geographic extent over which the distance-decay relationship was measured. The solid line shows the fit of a linear model (slope = 0.016, $P < 0.001$, adj-R^2 = 0.12). The positive relationship indicates that larger scale studies tend to record stronger distance-decay relationships.

Influence of Methodological Factors on the Distance-Decay Relationship

To determine whether the microbial distance-decay relationship may be influenced by methodological factors, I tested whether the method of community characterisation, sampling depth, or choice of community similarity index influence the Mantel correlation coefficient. In contrast to H4, high-throughput sequencing methods (HTS) did not result in significantly higher Mantel coefficients compared to fingerprinting methods, or other low resolution methods (Figure 3.4A; $F_{2,\ 284}$ = 0.19, P = 0.83). However, when only statistically significant (alpha = 0.05) Mantel coefficients were examined (Fig. 3.4B), high-throughput sequencing based studies showed higher Mantel coefficients, approaching statistical significance ($F_{2,\ 175}$ = 2.73, P = 0.07).

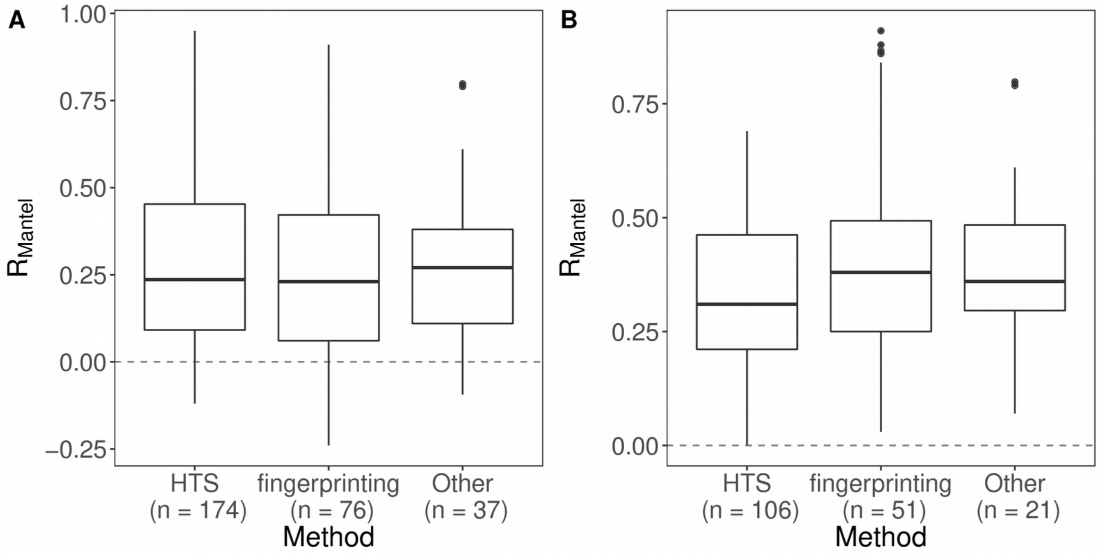


Figure 3.4 (A) All, and (B) only statistically significant, Mantel correlation coefficients (R_{Mantel}) from studies based on high-throughput sequencing (HTS), community fingerprinting approaches (such as DGGE or TRFLP), or other low resolution/throughput methods (morphological identification, Sanger sequencing). Larger Mantel coefficients indicate stronger distance-decay relationships.

Sequencing depth was also significantly and positively related to the Mantel coefficient, albeit with a small effect size (slope = 0.02, P < 0.05, adj-R^2 = 0.02), supporting the hypothesis (H5) that greater sequencing depth would result in stronger d-d relationships. Sequencing depth was not correlated to sampling effort (Pearson's ρ = 0.03, P = 0.64), showing that studies with greater sequencing depth did not necessarily incorporate more samples.

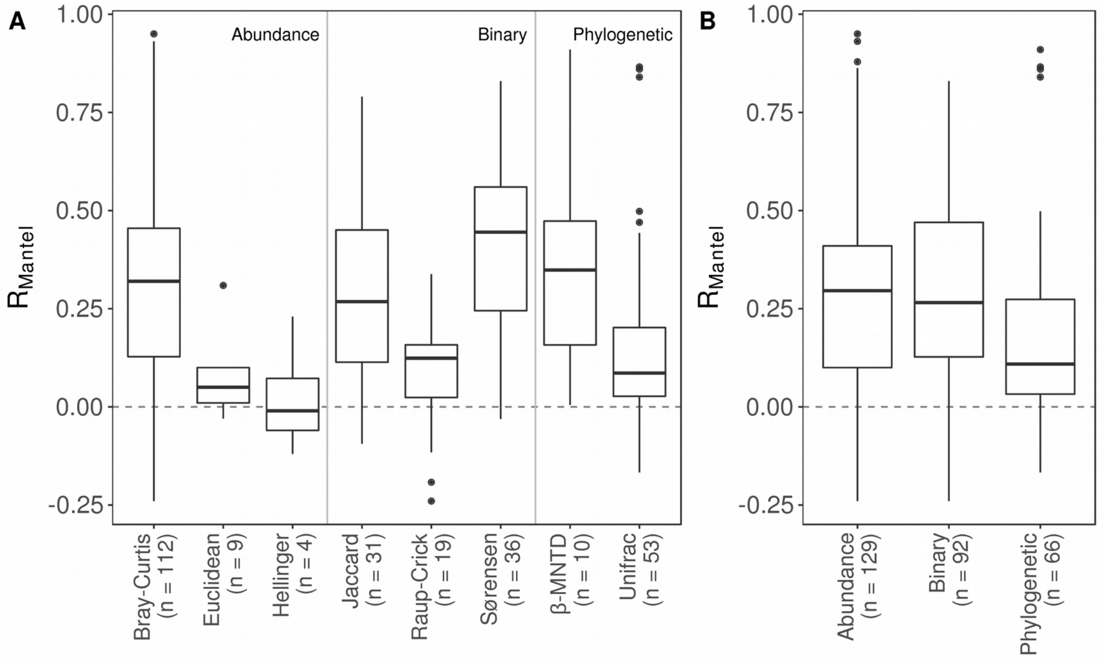


Figure 3.5 Mantel correlation coefficients from distance-decay relationships based on (A) different dissimilarity indices and, (B) different types of dissimilarity index. Index types reflect the different data requirements and type of distance (e.g. community composition or phylogenetic relatedness). Larger Mantel coefficients indicate stronger correlation between community dissimilarity and geographic distance.

In line with H6, significant differences were detected between dissimilarity indices ($F_{14,\ 271}$ = 4.96, P < 0.001). Several indices were excluded from this analysis as they had too few occurrences to calculate a reliable estimate of the central tendency (indices with < 4 occurrences were excluded). Tukey HSD tests showed Mantel coefficients from Raup-Crick and Unifrac indices were significantly lower than Bray-Curtis (*P* < 0.01 in each case, Fig. 3.5A), whilst Sørensen based coefficients were higher than Euclidean, Raup-Crick, and Unifrac indices (*P* < 0.01 in all cases, Fig. 3.5A). Furthermore, Mantel coefficients were significantly different between index types (Fig. 3.4B; $F_{2, 284}$ = 5.41, *P* < 0.01), and Tukey HSD tests showed that Mantel coefficients based on phylogenetic distances were significantly lower than both abundance (*P* < 0.01) and binary based indices (*P* < 0.05), supporting H6.

Discussion

Two decades of research into the spatial ecology of microbial communities has resulted in a highly variable impression of the microbial distance-decay (d-d) relationship. Our meta-analysis of 287 microbial d-d relationships has revealed two main findings. Firstly, d-d relationships may be influenced by methodological choices, including the sequencing depth used and the type of dissimilarity index. Secondly, as expected, the d-d relationship also appears to be dependent on various aspects of biological context, with different d-d relationships observed between different biomes and spatial scales.

The rapid development of methods in microbial ecology has improved our ability to detect and characterise ecological patterns in microbial communities, with high-throughput sequencing (HTS) platforms able to quantify microbial communities in ever increasing detail (Roesch *et al.* 2007; Caporaso *et al.* 2012). The tremendous sequencing depth of HTS platforms allows them to illuminate the “rare biosphere” (Caporaso *et al*., 2012), thus elevating them over other approaches such as “fingerprinting” which tend to capture a smaller proportion of the community. Initially, our results suggested that HTS-based approaches yielded similar strength d-d relationships to lower-resolution methods, such as fingerprinting and lower throughput methods, such as Sanger sequencing, suggesting that the massive sequencing depths offered by HTS platforms are not necessary to capture these ecological patterns (van Dorst *et al.* 2014). However, when I examined only statistically significant d-d relationships, the relationships derived from HTS approaches were stronger than other approaches. The ability of different methods to alter the strength of the d-d relationship is expected for two reasons. Firstly, fingerprinting and HTS approaches capture microbial diversity at different taxonomic resolutions. Comparative approaches have shown that fingerprinting approaches such as ARISA may be comparable to HTS data at the phylum level for instance (Gobet *et al.* 2014). Fingerprinting methods are therefore limited in that they may not detect compositional differences between communities at increasingly fine taxonomic resolutions (Ramette & Tiedje 2007; Bissett *et al.* 2010, Hanson *et al.,* 2012). This may weaken the d-d relationship in instances where communities are similar at the family level, but dissimilar at finer taxonomic levels. Secondly, fingerprinting methods are less able to sample from the “rare biosphere”, unlike HTS approaches. This is significant as, microbial communities often follow an occupancy-abundance relationship in which the most common organisms are also the most widespread, and the rarer organisms are the most restricted (Soininen & Heino 2005; Liu *et al.* 2015). Therefore, sampling only the most common, widespread organisms should flatten the d-d relationship by making communities appear artificially similar in composition (e.g. Zinger *et al.,* 2014). This is in contrast to a recent study, that demonstrated spatial turnover in communities is adequately reflected by “common species” alone in various freshwater communities (Heino & Soininen 2010). However, microbial communities are often enormously diverse and exhibit extremely “long tailed” species abundance distributions, such that the vast majority of microbial species in a community are “rare” (Hong *et al.* 2006; Galand *et al.* 2009; Locey & Lennon 2016). Therefore, it is likely that in microbial communities, common species alone may not adequately reflect patterns in spatial turnover (Galand *et al.* 2009).

Another methodological choice that was found to influence the strength of the microbial d-d relationship is the choice of dissimilarity index. Dissimilarity indices can vary in the type of data they consider (quantitative vs qualitative), the type of distance they quantify (compositional vs. phylogenetic), and the weight they place on common, rare, or absent species (Anderson *et al.* 2011). Within this study, I found significant differences in the d-d relationship between different indices, and between different index types. In particular, d-d relationship using phylogenetic indices were significantly flatter than compositional indices, whereas there was no difference between binary (presence/absence) and abundance based indices. Phylogenetic dissimilarity metrics may result in lower Mantel correlation coefficients for the same reason that fingerprinting methods do; because communities predominantly differ at fine taxonomic resolutions. This means that whilst communities differ in exact species or OTU composition, they can still be phylogenetically closely related, as communities may be highly similar at higher taxonomic ranks. In contrast, community composition metrics give no weight to how related communities are at broader taxonomic levels. The result of this is that communities appear more similar when phylogenetic indices are used (Bryant *et al*., 2008), potentially resulting in flatter d-d relationships (and therefore lower Mantel coefficients). This effect might be exacerbated when all sampled communities are from environmentally similar sites, which select for particular taxonomic groups. For example, extremophilic habitats such as solar salterns, can be highly similar at broad taxonomic levels, yet distinct at the OTU/species level (Zhaxybayeva *et al.* 2013; Clark *et al*., 2017).

Surprisingly, no difference was observed between quantitative and qualitative dissimilarity indices. This suggests that qualitative compositional differences between communities drive d-d relationships rather than quantitative changes in species composition and abundance. In agreement with previous studies that have applied both binary and abundance based indices, these two measures of community similarity are likely to be highly correlated (Martiny *et al.* 2011), and result in similar estimations of d-d relationships (e.g. Green *et al*. 2004, Glassman *et al*. 2015). This analysis also revealed that classic dissimilarity metrics, such as Bray-Curtis or Jaccard's index, are overwhelmingly the most frequently used in studies of microbial d-d relationships. These indices are undoubtedly amongst the most frequently used, not only in microbial ecology, but also more widely in ecology. I draw attention to several contemporary indices that may better suit the types of questions microbial ecologists ask as well as the properties of the data they generate. Classic metrics do not take into consideration co-occurrence information present within the data, which could increase understanding in microbial communities where there are many possible biotic interactions. To this end, a new family of metrics have been defined that account for species co-occurrence as well as shared taxa (Schmidt *et al.* 2017). Additionally, many indices rely on equal sample sizes, and are sensitive to differences in species richness (Green & Bohannan 2006), with potentially confounding effects on d-d relationships (Baselga 2007). Chao *et al*. (2005) therefore extended classic indices such as Jaccard and Sørensen to account for unobserved species, and to make them less sensitive to variable sample sizes, reducing the need for post-sequencing normalisation of sample sizes (McMurdie & Holmes 2014). Finally, many indices (such as Jaccard, Bray-Curtis, and Sørensen) are known to merge true compositional turnover (replacement of species) and nestedness (whereby communities are subsets of one another). To combat this, modified versions of classic indices such as Jaccard, Sorensen, and Bray-Curtis have been developed, allowing the partitioning of community similarity metrics into their turnover and nestedness components. This should enable a more mechanistic understanding of the processes behind d-d relationships (Baselga 2010, 2013; Podani & Schmera 2011). I echo the call of Green and Bohanan (2006) for microbial ecologists to exercise more care in their choice of dissimilarity metrics, especially now that many are implemented in popular and freely accessible analysis software, such as R (e.g. Baselga and Orme 2012).

Whilst significant differences were found between different methodological approaches, I also found differences relating to the biological context of each study. Against expectation, soil based studies had weaker d-d relationships than studies using other environmental materials. Soils are relatively stable habitats, in that they maintain physical structure and are therefore capable of maintaining significant environmental gradients over relatively small spatial scales. Therefore, I expected the combination of high habitat heterogeneity coupled with limited opportunity for dispersal to result in stronger d-d relationships than for example, oceanic waters, where physicochemical gradients are more diffuse. It is possible that the environmental gradients present in soils do not change linearly over geographic distance, for example if the similar environmental conditions are patchily distributed. Alternatively, soil microorganisms may be able to disperse more effectively than previously thought, perhaps via association with other soil organisms (Warmink *et al.,* 2011), migratory species such as birds (Bisson *et al*., 2007), wind blown soil particles (Kellogg & Griffin 2006; Favet *et al.,* 2013), or via bioaerosols (Joung *et al.,* 2017).

Originally, I expected that studies of aquatic microbial communities may show the weakest d-d relationships as riverine or oceanic hydrology may provide an effective dispersal mechanism, thus homogenising microbial communities and presenting more diffuse environmental gradients over larger spatial scales. Contrarily, I found that aquatic communities actually showed stronger d-d relationships indicating increased spatial turnover in aquatic microbial communities. Soininen *et al*. (2007) recorded similar distance-decay rates between terrestrial, marine and aquatic ecosystems, showing that biome-dependent d-d relationships may be a feature of microbial communities. Host-associated communities showed relatively strong, but variable d-d relationships. I suggest that this is caused jointly by the ecology of the host species, in combination with the degree of host specificity with the associated microbial community. For example, if the host is not dispersal limited, and associates with a large variety of microorganisms, then the d-d relationship may be relatively flat. However, if the host is dispersal limited, and associates with a very specific microbiome, the d-d relationship might be steeper. To develop our understanding of the macroecology of host-associated microbial communities, an interesting approach would be to compare microbial d-d relationships of sessile and motile hosts (motile host-associated d-d relationships were excluded in this analysis), as incorporating the ecology of the host (e.g. movement, interactions, range size) would likely provide further explanatory power.

Finally, I also found a relationship between the strength of the d-d relationship and the spatial scale over which the study was conducted. Scale-dependent d-d relationships have previously been reported (Bissett *et al.* 2010; Martiny *et al.* 2011; Soininen *et al.* 2011), albeit with contrasting results. Our results are comparable to those of Martiny *et al*. (2011) and Soininen *et al*. (2011)

who reported that d-d relationships for various microbial communities were generally steeper as greater spatial scales were incorporated. The scale dependence of this relationship may be explained by greater environmental heterogeneity in large scale studies, thus communities are subjected to different environmental filters, resulting in more dissimilar communities. In combination with this, communities separated by very large geographic distances should have minimal dispersal between them, assuming connectivity is linearly related to geographic distance. Alternatively, this observation may be a statistical artefact, caused by studies with very large spatial extents incorporating many zero similarity community comparisons (i.e. communities with no species in common), therefore biasing our quantification of the d-d relationship (Millar *et al.* 2011; Steinbauer *et al.* 2012). This point highlights that careful consideration is required in the statistical analysis of d-d relationships, especially when incorporating large geographic extents or highly dissimilar communities.

Despite its common use in the literature as evidence for neutral processes in microbial ecology, the d-d relationship alone does not provide evidence for neutral processes acting on microbial communities. As discussed previously, d-d relationships can arise from spatially autocorrelated environmental gradients as well as dispersal limitation (Nekola & White 1999). Furthermore, dispersal limitation itself is not solely a property of ecological neutrality. Dispersal limitation may be stochastic as predicted by neutral theory (Chave

2004), but also by asymmetric dispersal abilities between organisms (Salomon *et al.* 2010; Liu & Zhou 2011), thus violating the central tenet of neutral theory; that organisms are ecologically equivalent (Hubbell 2001). Thus, caution is urged in attributing distance-decay relationships to either niche or neutral processes without further evidence, for example from examining species-abundance distributions (e.g. Dumbrell *et al*. 2010). However, this is not to say that examining distance-decay relationships is futile as the relationship jointly reflects species turnover due to historical, environmental, and spatial factors, all of which are important factors to consider in studying biodiversity (Nekola & White 1999).

Moving beyond distance-decay relationships, focussing on other factors that influence the compositional similarity of microbial communities should provide interesting results. For example, quantifying the extent to which microorganisms differ in their dispersal abilities, and what traits are responsible for these differences may help to provide information on the biogeography of microorganisms at the population level, and given appropriate statistical approaches may allow us to predict the range size and habitat occupancy of different microbes. Furthermore, it is commonly assumed that the connectivity between communities is linearly related to the spatial distance between communities. However, given that different dispersal vectors may disperse microorganisms over differing geographic distances, this assumption may not be valid. Therefore, the growing movement towards examining the role of connectivity *per se (*Declerck *et al*. 2013; Vannette *et al*. 2016), rather than using geographical distance as a proxy, will likely provide a fruitful direction for spatial microbial ecology. By modeling the dispersal process itself and quantifying connectivity, a more mechanistic understanding of the spatial ecology of microbial communities could be gained.

References

Anderson MJ, Crist TO, Chase JM, *et al.* (2011) Navigating the multiple meanings of β diversity: A roadmap for the practicing ecologist. *Ecology Letters*, **14**, 19–28.

Baas Becking LGM (1934) *Geobiologie of inleiding tot de milieukunde*. W.P. Van Stockum & Zoon, The Hague, Netherlands.

Barreto DP, Conrad R, Klose M, Claus P, Enrich-Prast A (2014) Distance-decay and taxa-area relationships for Bacteria, Archaea and methanogenic Archaea in a tropical lake sediment. *PLoS ONE*, **9**, e110128.

Baselga A (2007) Disentangling Distance Decay of Similarity from Richness Gradients: Response to Soininen *et al*. 2007. *Ecography*, **30**, 838–841.

Baselga A (2010) Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, **19**, 134–143.

Baselga A (2012) The relationship between species replacement, dissimilarity derived from nestedness, and nestedness. *Global Ecology and Biogeography*, **21**, 1223–1232.

Baselga A (2013) Separating the two components of abundance-based dissimilarity: Balanced changes in abundance vs. abundance. *Methods in Ecology and Evolution*, **4**, 552–557.

Baselga A, Orme CDL (2012) betapart: an R package for the study of beta diversity. *Methods in Ecology and Evolution*, **3**, 808–812.

Beck J, Holloway JD, Schwanghart W (2013) Undersampling and the measurement of beta diversity. *Methods in Ecology and Evolution*, **4**, 370–382.

Bissett A, Richardson AE, Baker G, Wakelin S, Thrall PH (2010) Life history determines biogeographical patterns of soil bacterial communities over multiple spatial scales. *Molecular Ecology*, **19**, 4315–4327.

Bisson IA, Marra PP, Burtt EH, Sikaroodi M, Gillevet PM (2007) A molecular comparison of plumage and soil bacteria across biogeographic, ecological, and taxonomic scales. *Microbial Ecology*, **54**, 65-81.

Bryant JA, Lamanna C, Morlon H, Kerkhoff AJ, Enquist BJ, Green JL (2008) Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. *Proceedings of the National Academy of Sciences*

USA, **105** (Supplement 1), 11505-11511.

Caporaso JG, Lauber CL, Walters WA, *et al.* (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal*, **6**, 1621–1624.

Chao A, Chazdon RL, Colwell RK, Shen TJ (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters*, **8**, 148–159.

Chave J (2004) Neutral theory and community ecology. *Ecology Letters*, **7**, 1–39.

Clark DR, Mathieu M, Mourot L, Dufossé L, Underwood JC, Dumbrell AJ, McGenity TJ (2017) Biogeography at the Limits of Life: Do Extremophilic Microbial Communities Show Biogeographic Regionalisation? *Global Ecology and Biogeography,* **26**, 1435-1446.

Condit R, Pitman N, Leigh EG, *et al.* (2002) Beta-diversity in tropical forest trees. *Science*, **295**, 666–669.

Declerck SAJ, Winter C, Shurin JB, Suttle CA, Matthews B (2013) Effects of patch connectivity and heterogeneity on metacommunity structure of planktonic bacteria and viruses. *The ISME Journal*, **7**, 533–542.

Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH (2010) Relative roles of niche and neutral processes in structuring a soil microbial community. *The ISME Journal*, **4**, 337–345.

Favet J, Lapanje A, Giongo A, *et al.* (2013) Microbial hitchhikers on intercontinental dust: catching a lift in Chad. *The ISME Journal*, **7**, 850–867.

Finlay BJ (2002) Global dispersal of free-living microbial eukaryote species. *Science*, **296**, 1061–1063.

Finlay BJ, Fenchel T (2004) Cosmopolitan Metapopulations of Free-Living Microbial Eukaryotes. *Protist*, **155**, 237–244.

Galand PE, Casamayor EO, Kirchman DL, Lovejoy C (2009) Ecology of the rare microbial biosphere of the Arctic Ocean. *Proceedings of the National Academy of Sciences USA*, **106**, 22427–22432.

Glassman SI, Peay KG, Talbot JM, *et al.* (2015) A continental view of pine-associated ectomycorrhizal fungal spore banks: A quiescent functional guild with a strong biogeographic pattern. *New Phytologist*, **205**, 1619–

1631.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–769.

Gobet A, Boetius A, Ramette A (2014) Ecological coherence of diversity patterns derived from classical fingerprinting and Next Generation Sequencing techniques. *Environmental Microbiology*, **16**, 2672–2681.

Green JL, Bohannan BJM (2006) Spatial scaling of microbial biodiversity. *Trends in Ecology and Evolution*, **21**, 501–507.

Green JL, Holmes AJ, Westoby M, *et al.* (2004) Spatial scaling of microbial eukaryote diversity. *Nature*, **432**, 747–750.

Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JB (2012) Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews Microbiology*, **10**, 497-506.

Harrison F (2011) Getting started with meta-analysis. *Methods in Ecology and Evolution*, **2**, 1-10.

Hazard C, Gosling P, van der Gast CJ, *et al.* (2013) The role of local environment and geographical distance in determining community composition of arbuscular mycorrhizal fungi at the landscape scale. *The ISME Journal*, **7**, 498–508.

Heino J, Soininen J (2010) Are common species sufficient in describing turnover in aquatic metacommunities along environmental and spatial gradients? *Limnology and Oceanography*, **55**, 2397–2402.

Holt RD (2009) Bringing the Hutchinsonian niche into the 21st century: ecological and evolutionary perspectives. *Proceedings of the National Academy of Sciences USA*, **106** (Supplement 2), 19659-19665.

Hong SH, Bunge J, Jeon SO, Epstein SS (2006) Predicting microbial species richness. *Proceedings of the National Academy of Sciences USA*, **103**, 117–122.

Hubbell SP (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press. Princeton, NJ, USA.

Joung YS, Ge Z, Buie CR (2017) Bioaerosol generation by raindrops on soil. *Nature Communications*, **8**, 14668.

Kellogg CA, Griffin DW (2006) Aerobiology and the global transport of desert dust. *Trends in Ecology and Evolution*, **21**, 638–644.

Kivlin SN, Winston GC, Goulden ML, Treseder KK (2014) Environmental filtering affects soil fungal community composition more than dispersal limitation at regional scales. *Fungal Ecology*, **12**, 14–25.

Lajeunesse MJ (2016) Facilitating systematic reviews, data extraction and meta-analysis with the metagear package for R. *Methods in Ecology and Evolution*, **7**, 323–330.

Legendre P, Fortin MJ, Borcard D (2015) Should the Mantel test be used in spatial analysis? *Methods in Ecology and Evolution*, **6**, 1239–1247.

Lisboa FJG, Peres-Neto PR, Chaer GM, da Conceição Jesus E, Mitchell RJ, Chapman SJ, Berbara RLL (2014) Much beyond mantel: bringing procrustes association metric to the plant and soil ecologist's toolbox. *PloS ONE*, ***9***, e101238.

Liu L, Yang J, Yu Z, Wilkinson DM (2015) The biogeography of abundant and rare bacterioplankton in lakes and reservoirs of China. *The ISME Journal*, **9**, 2068–2077.

Liu J, Zhou S (2011) Asymmetry in species regional dispersal ability and the neutral theory. *PLoS ONE*, **6**, e24128.

Locey KJ, Lennon JT (2016) Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences USA*, **113**, 5970–5975.

Martiny JBH, Eisen JA, Penn K, Allison SD, Horner-Devine MC (2011) Drivers of bacterial β-diversity depend on spatial scale. *Proceedings of the National Academy of Sciences USA*, **108**, 7850–7854.

McMurdie PJ, Holmes S (2014) Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, **10**, e1003531.

Millar RB, Anderson MJ, Tolimieri N (2011) Much ado about nothings: Using zero similarity points in distance-decay curves. *Ecology*, **92**, 1717–1722.

Muyzer G (1999) DGGE/TGGE a method for identifying genes from natural ecosystems. *Current Opinion in Microbiology*, **2**, 317–322.

Nekola JC, White PS (1999) The distance decay of similarity in biogeography and ecology. *Journal of Biogeography*, **26**, 867–878.

Podani J, Schmera D (2011) A new conceptual and methodological framework for exploring and explaining pattern in presence - absence data. *Oikos*, **120**, 1625–1638.

R Developement Core Team (2016) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria*.

Ramette A (2007) Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology*, **62**, 142-160.

Ramette A, Tiedje JM (2007) Biogeography: an emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microbial Ecology*, **53**, 197–207.

Roesch LFW, Fulthorpe RR, Riva A, *et al.* (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME Journal*, **1**, 283–90.

Salomon Y, Connolly SR, Bode L (2010) Effects of asymmetric dispersal on the coexistence of competing species. *Ecology Letters*, **13**, 432–441.

Shmida AVI, Wilson MV (1985) Biological determinants of species diversity. *Journal of Biogeography*, **12** 1-20.

Schmidt TSB, Matias Rodrigues JF, von Mering C (2017) A family of interaction-adjusted indices of community similarity. *The ISME Journal*, **11**, 791–807.

Sogin ML, Morrison HG, Huber JA, *et al.* (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere.” *Proceedings of the National Academy of Sciences USA*, **103**, 12115–12120.

Soininen J, Heino J (2005) Relationships between local population persistence, local abundance and regional occupancy of species: Distribution patterns of diatoms in boreal streams. *Journal of Biogeography*, **32**, 1971–1978.

Soininen J, Korhonen JJ, Karhu J, Vetterli A (2011) Disentangling the spatial patterns in community composition of prokaryotic and eukaryotic lake plankton. *Limnology and Oceanography*, **56**, 508–520.

Soininen J, Korhonen JJ, Luoto M (2013) Stochastic species distributions are driven by organism size. *Ecology*, **94**, 660–670.

Soininen J, McDonald R, Hillebrand H (2007) The distance decay of similarity in ecological communities. *Ecography*, **30**, 3–12.

Steinbauer MJ, Dolos K, Reineking B, Beierkuhnlein C (2012) Current measures for distance decay in similarity of species composition are influenced by study extent and grain size. *Global Ecology and Biogeography*, **21**, 1203–1212.

van Dorst J, Bissett A, Palmer AS, *et al.* (2014) Community fingerprinting in a sequencing world. *FEMS Microbiology Ecology*, **89**, 316–330.

Vannette RL, Leopold DR, Fukami T (2016) Forest area and connectivity influence root-associated fungal communities in a fragmented landscape. *Ecology*, **97**, 2374–2383.

Warmink JA, Nazir R, Corten B, van Elsas JD (2011) Hitchhikers on the fungal highway: The helper effect for bacterial migration via fungal hyphae. *Soil Biology and Biochemistry*, **43**, 760–765.

Wilkinson DM, Koumoutsaris S, Mitchell EAD, Bey I (2012) Modelling the effect of size on the aerial dispersal of microorganisms. Journal of B*iogeography*, **39**, 89–97.

Zhaxybayeva O, Stepanauskas R, Mohan NR, Papke RT (2013) Cell sorting analysis of geographically separated hypersaline environments. *Extremophiles*, **17**, 265–275.

Chapter 4

Biogeography at the Limits of Life: Do Extremophilic Microbial Communities Show Biogeographic Regionalisation?

Acknowledgements: Mégane Mathieu, Léonie Mourot, Laurent Dufossé, Graham JC Underwood, Alex J Dumbrell and Terry J McGenity

Abstract

Aim

Biogeographic regions are the fundamental geographic units for grouping Earth's biodiversity. Biogeographic regionalisation has been demonstrated for many higher taxa, such as terrestrial plants and vertebrates, but not in microbial communities. Therefore, we sought to test empirically whether microbial communities, or taxa, show patterns consistent with biogeographic regionalisation.

Location

Within halite (NaCl) crystals from coastal solar salterns of West Europe, the Mediterranean, and East Africa.

Time period

Modern (2006 – 2013).

Major taxa studied

Archaea.

Methods

Using high-throughput Illumina amplicon sequencing, we generated the most high-resolution characterisation of halite-associated archaeal communities to date, using samples from 17 locations. We grouped communities into biogeographical clusters based on community turnover, to test whether these communities show biogeographic regionalisation. To examine if individual taxa, rather than communities, show biogeographic patterns, we also tested whether the relative abundance of individual genera may be indicative of a community's biogeographic origins using machine learning methods, specifically random forest classification.

Results

We found that the rate of community turnover was greatest over sub-regional spatial scales (< 500 km) whilst at regional spatial scales, turnover was independent from geographic distance. Biogeographic clusters of communities were either not statistically robust, or lacked spatial coherence, inconsistent with biogeographic regionalisation. However, we identified several archaeal genera that were good indicators of biogeographic origin, providing classification error rates of < 10%.

Main conclusions

Overall, our results provide little support for the concept of biogeographic regions in these extremophilic microbial communities, despite the fact that some taxa do show biogeographic patterns. We suggest that variable dispersal ability among the halite-associated Archaea may disrupt biogeographic patterns at the community level, preventing the formation of biogeographic regions. This means that the processes that lead to the formation of biogeographic regions operate differentially on individual microbial taxa rather than on entire communities.

Introduction

The classification of Earth’s biota into biogeographic regions separated by dispersal barriers, has captivated ecologists for centuries (Sclater, 1858; Wallace, 1876). The concept of biogeographic regionalisation has yielded insight into the origins of biodiversity and areas of endemism (Lamoreux *et al*., 2006), informed us of species' conservation status (Buckley & Jetz, 2007), and revealed historical connectivity between communities (Cowen *et al.*, 2006). However, early attempts to define these regions have been superseded by more quantitative methods, improving the robustness and reproducibility of region delineations (Kreft & Jetz, 2010; Vilhena & Antonelli, 2015; Dapporto *et al.*, 2015). Coupled with these new methods, the ever increasing availability of species distribution data has renewed interest in the concept of biogeographic regionalisation. As a result, a far greater range of taxa have been studied than ever before in order to define biogeographic regions (Holt *et al*., 2013). Yet, our knowledge about how Earth's biota may be divided into biogeographic regions is still overwhelmingly based on multicellular (and usually large) eukaryotes. Many inconspicuous, but functionally critical organisms, such as microorganisms, remain poorly studied. Consequently, it is unknown whether microbial communities may be grouped into biogeographic regions, similar to those observed for higher taxa.

Microorganisms are arguably the most functionally diverse and important organisms on Earth (Dinsdale *et al*., 2008; Fierer *et al*., 2012), driving every biogeochemical cycle (Zak *et al.*, 2003; Falkowski *et al.*, 2008). Originally, microorganisms were assumed to have cosmopolitan distributions with their small size and high population densities making them effective passive dispersers (Baas Becking, 1934; Finlay, 2002). From this assumption, it follows that biogeographic regionalisation may not be possible because dispersal limitation is required for areas of endemism to occur (Ficetola *et al.*, 2017) and produce regions with distinct communities. In contrast to cosmopolitanism, many recent studies have documented relationships between community turnover (the replacement of species) and geographic distance, indicative of dispersal limitation (e.g. Dumbrell *et al.*, 2010; Lear *et al.*, 2014), hinting that biogeographic regionalisation of microbial communities could be possible. However, whilst the composition of microbial communities has been shown to differ over biogeographic regional scales, a formal test of whether microbial communities exhibit biogeographic regionalisation is lacking.

In order to test for the presence of biogeographic regionalisation in microbial communities, an ideal model community should have relatively low diversity, inhabit isolated environments, and show *a priori* evidence of dispersal limitation. The halite-associated Archaea fulfil these criteria. These Archaea typically belong to the class Halobacteria (more commonly referred to as haloarchaea) and are a major component of halite endolith communities (Henriet *et al.*, 2014). Their entombment into the brine inclusions of halite crystals is believed to be an escape mechanism from desiccation and the increasingly chaotropic conditions present in evaporating brines (Hallsworth *et al*., 2007). Within these pockets they are able to survive over geological time scales (McGenity *et al.*, 2000; Gramain *et al.*, 2011). As with many extremophilic microbial communities, the halite-associated Archaea are typically less diverse than other microbial systems, facilitating more exhaustive sampling of the total diversity and improving detection of the less abundant endemic taxa, which are indicative of biogeographic regions. Furthermore, these Archaea occupy isolated “habitat islands” that are physicochemically distinct from the surrounding environment. Many haloarchaea are obligately halophilic and lyse in less saline conditions (Oren, 1994) such as seawater, rendering the surrounding environment a physiological dispersal barrier. Finally, halite crystals form under highly similar conditions worldwide, i.e. saturated NaCl, thus ensuring that species filtering by the environment is low compared with many other environments. Any physicochemical differences between halite crystals, e.g. caused by underlying geology or climate, should themselves be spatially auto-correlated, meaning that species filtering by the environment should enhance, rather than obscure, biogeographic clustering. Such systems are therefore ideal for studies of community turnover and biogeography (Santos *et al.*, 2016). Previous studies of halophilic microbial communities have found evidence of community turnover at regional scales (Pagaling *et al*., 2009; Zhaxybayeva *et al.*, 2013), suggesting the potential for biogeographic regions to form. Overall, these properties render the halite-associated Archaea an ideal system in which to test for biogeographic regionalisation of microbial communities.

Therefore, we examine the regional turnover (replacement of species over biogeographic regional scales), of halite-associated archaeal communities to test whether communities group together in a manner consistent with biogeographic regionalisation. Using high-throughput Illumina HiSeq amplicon sequencing, we characterise the archaeal communities of halite from 17 locations, spanning three geographic regions. We apply robust biogeographic clustering methods to determine the extent to which archaeal communities, and taxa, show spatial patterns consistent with biogeographic regionalisation. We propose the following three hypotheses:

1: a) There will be a significant relationship between community turnover and geographic distance, and b) the rate of community turnover will be greater at biogeographic regional scales than at within region scales.

2: Communities will form biogeographic clusters that are statistically well supported and spatially coherent.

3: The presence and abundance of some archaeal taxa can predict the (bio)geographic origin of each community.

Methods

We obtained 27 halite samples (in triplicate) from 17 locations in the years between 2006 and 2013 (Fig. 4.1 and Appendix S8). A photographic record of the samples and further details can be found in the supporting information (Appendix S1). We recorded the grain size, which reflects the time taken for the crystals to form, and the impurity colour, which provides a qualitative measure of the types of impurities and physicochemical environment present within the crystal (Sonnenfeld, 1995). Samples were stored in the dark at room temperature.

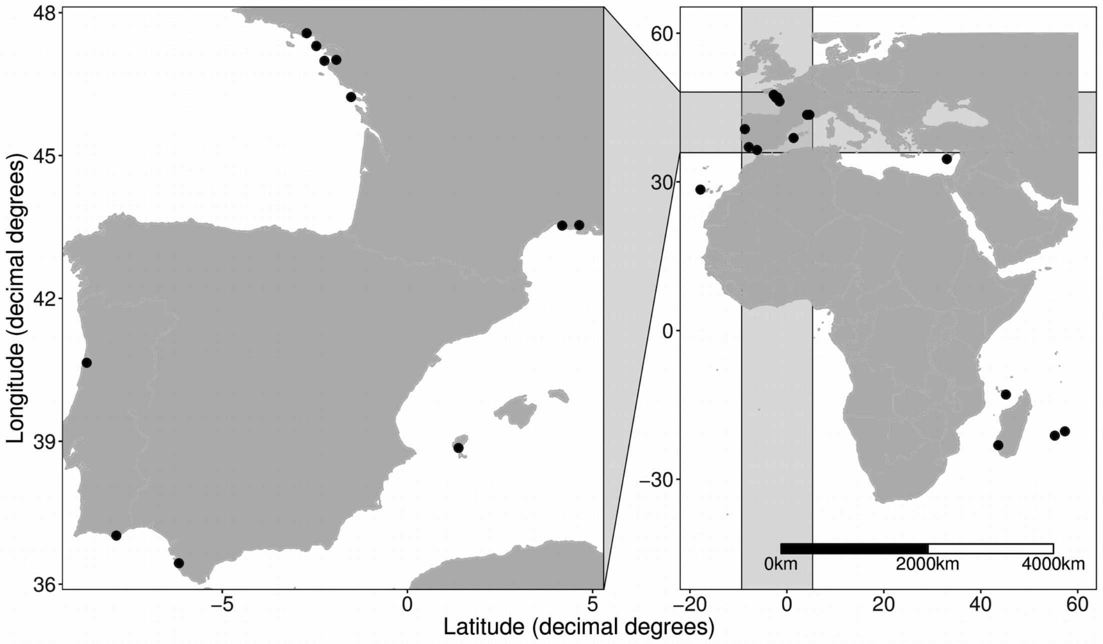


Figure 4.1 Map of sample locations. Further details of samples are available in Appendix S1. The left panel is zoomed in on the grey region in order to distinguish multiple locations along the South West coast of Europe.

Molecular Analyses

DNA was extracted from a 0.25 g aliquot of each sample using MoBio PowerSoil DNA isolation kits following the manufacturer’s instructions (MoBio Laboratories Inc., Carlsbad, CA, USA). To characterise the archaeal communities, we used a Nextera XT dual indexing strategy which involves PCR amplification of a phylogenetic marker gene, followed by a secondary short-cycle PCR amplification in which dual Nextera indices are added to the amplicon for multiplexing of samples. We targetted a ~570 bp region of the 16S rRNA gene with the Archaea specific primers 344F (5'-ACGGGGYGCAGCAGGCGCGA-3', Raskin *et al.*, 1994) and 915R (5'-GTGCTCCCCCGCCAATTCCT-3', Stahl & Amann, 1991), both of which were modified to contain Illumina specific overhang sequences. The 16S rRNA gene was amplified in 25 µl reactions with 12.5 µl of REDTaq® ReadyMixTM (Sigma-Aldrich Co.), 5 µl of each primer (1 µM) and 2.5 µl of template DNA. The PCR protocol included an initial denaturation step at 95°C for 5 min, followed by 32 cycles of 95°C for 45 s, 60°C for 45 s and 72°C for 1 min. After a final extension step of 72°C for 5 min, PCR products were held at 4°C. We purified PCR products using Agencourt AMPure XP PCR Purification beads (Beckman Coulter Ltd, High Wycombe, UK) following Illumina’s “16S Metagenomic Sequencing Library Preparation” document (https://goo.gl/3Y7oY4).

The index PCR was carried out in 50 µl reactions with 25 µl of KAPA HiFi HotStart ReadyMix (KAPA Biosystems, Wilmington, MA, USA), 5 µl each of sample specific i5 and i7 Nextera XT index (Illumina), 10 µl of PCR-water (Bioline Reagents Ltd, UK) and 5 µl of purified PCR product. PCR was conducted with an initial denaturation at 95°C for 3 min, followed by 8 cycles of 95°C for 30 s, 55°C for 30 s and 72°C for 30 s. Again, a final extension step was included at 72°C for 5 min, after which PCR products were held at 4°C. PCR products were purified using Agencourt AMPure XP PCR Purification beads (Beckman Coulter Ltd, High Wycombe, UK) and quantified on a POLARstar Omega (BMG LABTECH GmbH, Germany) plate reader using the PicoGreen® dsDNA assay. PCR products were then pooled in equimolar concentrations. The size and concentration of the resulting pool was checked using an Agilent 2100 Bio-analyser. Sequencing was carried out on an Illumina HiSeq 2500 on rapid-run mode, producing 2 x 300 bp sequences, at The Earlham Institute (formerly The Genome Analysis Centre, Norwich Research Park, Norfolk, UK).

Bioinformatic Analyses

Due to the length of the amplicon, forward and reverse sequences were unable to be pair-end aligned and so all analyses were based on forward sequences only. This approach has been shown to have little effect on ecological conclusions (Werner *et al.*, 2012), and in our case, the forward sequence spans the V3 region of the 16S rRNA gene, which has been shown to perform well for profiling archaeal communities (Yu *et al.*, 2008). Sequences were processed according to guidelines outlined in (Dumbrell *et* *al.*, 2016). Briefly, we quality trimmed sequences using Sickle (Joshi & Fass, 2011) at a threshold of Q20, trimming only the 3′ end of the sequence and discarding sequences with ambiguous nucleotides. Quality-trimmed sequences were error-corrected using the BayesHammer algorithm implemented in SPAdes version 3.10.1, with default parameters (Bankevich *et al*., 2012; Nikolenko *et al.*, 2013). We removed primer sequences, calculated library sizes for each sample, and discarded sequences < 230 nucleotides in length using Linux shell commands. Samples with excessively small library sizes (<20,000 sequences) were excluded from further analyses.

We used VSEARCH (Rognes *et al.*, 2016) to cluster sequences into operational taxonomic units (OTUs). First, sequences were de-replicated and singleton sequences discarded, as they are more likely to be artefacts (Flynn *et al.*, 2015). We then clustered sequences into OTUs at 97% and 99% sequence similarity (referred to as 97% dataset and 99% dataset). The 97% similarity threshold is the most frequently used, corresponding approximately to intra-genus level similarity (Yarza *et al*., 2014). The 99% threshold approximates to species-level similarity. We screened OTUs for chimeras against the RDP database (Cole *et al*., 2009) using VSEARCH, and discarded putative chimeras.

Taxonomy was assigned to OTUs using the RDP classifier (Wang *et al.*, 2007) with a minimum confidence threshold of 0.7. We discarded all non-archaeal OTUs. Specific OTUs of interest were identified using BLAST searches against NCBI's 16S ribosomal RNA sequence database (Altschul *et al.*, 1990).

Statistical Analyses

We rarefied OTU tables to the smallest library size in each dataset (97% dataset; 27,554 sequences, 99% dataset; 26,578). We checked whether sample age was influencing the OTU richness or community composition using a negative binomial GLM and permutation-based multivariate analysis of variance (PERMANOVA), respectively.

In order to address our first hypothesis, we quantified community turnover using the β_{sim} index, which purely quantifies community turnover, the process relevant to biogeographic regionalisation (Baselga, 2010), and not nestedness, whereby communities are subsets of each other. Geographic distances between sampled communities were calculated as geodesic distances (Hijmans, 2016). We then tested for correlation between community turnover and geographic distance using Mantel tests, with Spearman's correlation coefficient and 10,000 permutations. We fitted piece-wise regressions to determine breakpoints in the relationship, showing the geographic distance at which the slope of the relationship changes (Castro-Insua *et al.*, 2016).

To investigate our second hypothesis, we adopted a clustering approach as described by Kreft & Jetz (2010). Briefly, this approach involves clustering communities based on the β_{sim} turnover matrix, creating a dendrogram. This dendrogram can be split into *k* clusters representing bioregions. The quality and biological interpretability of the resulting clusters are then checked via statistical metrics and mapping. Biogeographic regionalisation may be inferred when clustering solutions are both statistically robust and spatially coherent.

To cluster communities, we used three different clustering algorithms to ensure our conclusions were robust. The unweighted pair-group method using arithmetic averages (UPGMA) defines the distance between clusters as the average distance between all the communities within each cluster. Kreft & Jetz (2010) found that UPGMA best preserved information present in the original distance matrix. Dapporto *et al*. (2015) also compared clustering algorithms on datasets of varying completeness. They found that for less well sampled datasets, the Ward method clustered communities most accurately, whereas for intensely sampled datasets, PAM produced the most accurate clusters. To cluster the communities, we used the methodology described by Dapporto *et al*. (2015). This approach overcomes the biases introduced by having zero similarity or tied values in the dissimilarity matrix (Bloomfield *et al.*, 2017), by repeatedly reshuffling the matrix and re-clustering communities. The final clustering solution is then determined by the frequency at which pairs of communities are clustered together in the randomly generated cluster solutions, allowing a more robust final clustering solution. We set the number of matrix randomisations to 50, and the number of clusters (*k*) from 2 to 16. For each value of *k*, we assessed the statistical support of the cluster solution with two metrics, “mean silhouette width” and “explained dissimilarity”. The first metric, “mean silhouette width”, is a commonly used metric to evaluate clustering solutions and ranges from -1, indicating that most communities have been incorrectly clustered, to 1 indicating that most communities are correctly clustered. Values below 0.25 are qualitatively considered to show little evidence of true clustering between the communities (Kaufman & Rousseeuw, 1990). Our second metric, “explained dissimilarity” (Holt *et al*., 2013), is a ratio of sums of mean dissimilarity within regions to total dissimilarity over the entire dissimilarity matrix. “Explained dissimilarity” tends towards 1 as *k* tends towards the number of communities. We follow the approach of Holt *et al*. (2013), who indicated that a threshold of 0.9 provides sufficient support to infer regionalisation. However, we also examined the cluster solution that produced the greatest incremental increase in “explained dissimilarity”, which we refer to as the “knee solution”, as this has been proposed to be a more suitable indicator of optimum cluster number (Kreft & Jetz, 2013). After identifying statistically supported clustering solutions, we determined the spatial coherence of clusters by mapping them. To check whether the measured physicochemical parameters (grain size or impurity) explained any clustering patterns observed, we used PERMANOVA analysis.

We included location as the first variable in the model to account for confounding spatial effects. Statistical significance of physicochemical variables was then assessed based on the “marginal” effects (e.g. after controlling for spatial location), with 999 random permutations. We conducted non-metric multidimensional scaling (NMDS) analysis as a means of visualising these results.

To test our third hypothesis, we investigated whether the relative abundance of halite-associated archaeal genera could predict the biogeographic origin of a given community using the machine learning method, random forest classification. Random forests provide an effective method for classification in ecology (Cutler *et al*., 2007) and are built from an ensemble of classification trees, in which observations of the dependant variable form the leaves and independent variables form the branches. Each tree is trained on a subset of observations and independent variables, and the overall classifier is built by combining predictions from these trees to obtain a more robust classification. We summed the abundances of all OTUs identified to the genus level, and converted these abundances to relative abundances. OTUs not identified to genus were excluded from this analysis. We classified communities (see Appendix S8) based on their biogeographic region (classes: Palearctic, Saharo-Arabian, Madagascan as defined by Holt *et al*. (2013)), geographic region (classes: E Europe, W Europe, Mediterranean or W African), and nearest ocean (classes: Atlantic or Indian). We initialised 10,000 trees and each tree was trained on six archaeal genera. We normalised the sample size from each class to the size of the smallest class to minimise the effects of class size imbalance (e.g. more observations of European communities than African communities). Additionally, for the biogeographic and geographic classifiers, we dropped excessively small classes (Saharo-Arabian; *n* = 4 and West African; *n* = 3), to further reduce the imbalance between classes. We evaluated the overall accuracy of each classifier using the out-of-bag error rate, which quantifies the classifier's ability to correctly classify a given community when it is excluded from the training set. We determined which archaeal genera were the best predictors of biogeographic origin by quantifying variable importance, using the mean decrease in accuracy (MDA), and mean decrease in Gini-index (MDGI). The MDA shows the change in accuracy of the classifier with and without a given variable. Important variables will result in a large decrease in accuracy when they are excluded from the classifier, resulting in large MDA values. MDGI shows the purity of the groups created when the classifier splits the dataset using a given predictor. A good predictor will create homogeneous groups in which all data points belong to the same class, resulting in a large decrease in MDGI. We also examined partial dependence plots (Hastie *et al.*, 2009). In the context of our study, these plots show how the probability of a community being classified into a given biogeographic region changes in relation to the relative abundance of a given archaeal genus.

All analyses were conducted in R (R Developement Core Team, 2016), using the “*vegan*” (Oksanen *et al*., 2015), “*recluster*“ (Dapporto, Ramazzotti, *et al*., 2015), and “*randomForest*” (Liaw & Wiener, 2002) packages.

Results

Diversity of Halite-Associated Archaea

An initial 17.8 million sequences were reduced to 10.33 million after quality trimming. Error correction, length filtering, and removal of small samples further decreased this total to 10.29 million sequences. These sequences clustered into 1,581 and 10,346 OTUs at the 97% and 99% similarity thresholds, respectively. Sixteen non-archaeal OTUs (12 Bacteria, 4 unclassified) were removed from each dataset, comprising a total of 294 sequences (< 0.0001% of total sequences). Of the archaeal OTUs, 45.2% were identified to genus level from the 97% dataset, and 59.5% from the 99% dataset (Appendix S4). At the 99% similarity level, these OTUs represented 40 genera from 5 families (Appendix S2) as identified by the RDP taxonomy; the Halobacteriaceae, Haloferacaceae, Natrialbaceae, Methanosarcinaceae, and Nitrososphaeraceae. Most OTUs (58% from 97% dataset, 79.9% from 99% dataset) were restricted to 20 or fewer samples, but 5 OTUs in the 97% dataset and 3 OTUs in the 99% dataset were detected in every sample (Appendix S5). BLAST analysis of these OTUs revealed their most closely related species as *Halobacterium noricense* (OTU1), *Halorubrum orientale* (OTU2), *Halorubrum sodomense* (OTU21), *Halolamina sediminis* (OTU5), *Halolamina salina* (OTU92510).

Sample age did not significantly affect OTU richness (97% dataset; slope = -0.01, *z*-statistic = -0.30, *P* = 0.77, 99% dataset; slope = -0.02, *z*-statistic =

-0.83, P = 0.41), whilst PERMANOVA analyses showed that age had a small, but significant, effect on turnover (97% dataset; pseudo-$F_{1,\ 74}$ *= 2.50,* R^2 = 0.03, P = 0.03, 99% dataset; pseudo-$F_{1,\ 74}$ = 3.13, R^2 = 0.04, P = 0.003).

How is Community Turnover Related to Geographic Distance?

Mantel tests used to investigate the relationship between community turnover and geographic distance, showed significant and positive relationships for both datasets (97% dataset; r_{Mantel} = 0.26, $P < 0.0001$, 99% dataset; r_{Mantel} = 0.31, $P < 0.0001$), which supports hypothesis 1a. However, piece-wise regressions between geographic distance and community turnover suggested that this correlation was largely driven by high turnover at small spatial scales (Fig. 4.2). For both (97% and 99%) datasets, a steep positive relationship was found at smaller spatial scales with break-points estimated at 420.5 km (standard error = 46.9 km) and 334.6 km (standard error = 23.7 km) respectively. After these breakpoints, community turnover was independent from geographic distance (Fig. 4.2). Davies tests confirmed that the pre-breakpoint slope was significantly greater than the post-breakpoint slope ($P <$ 0.0001 in both cases), showing that the greatest rate of community turnover was at small, sub-regional scales, and so rejecting hypothesis 1b.

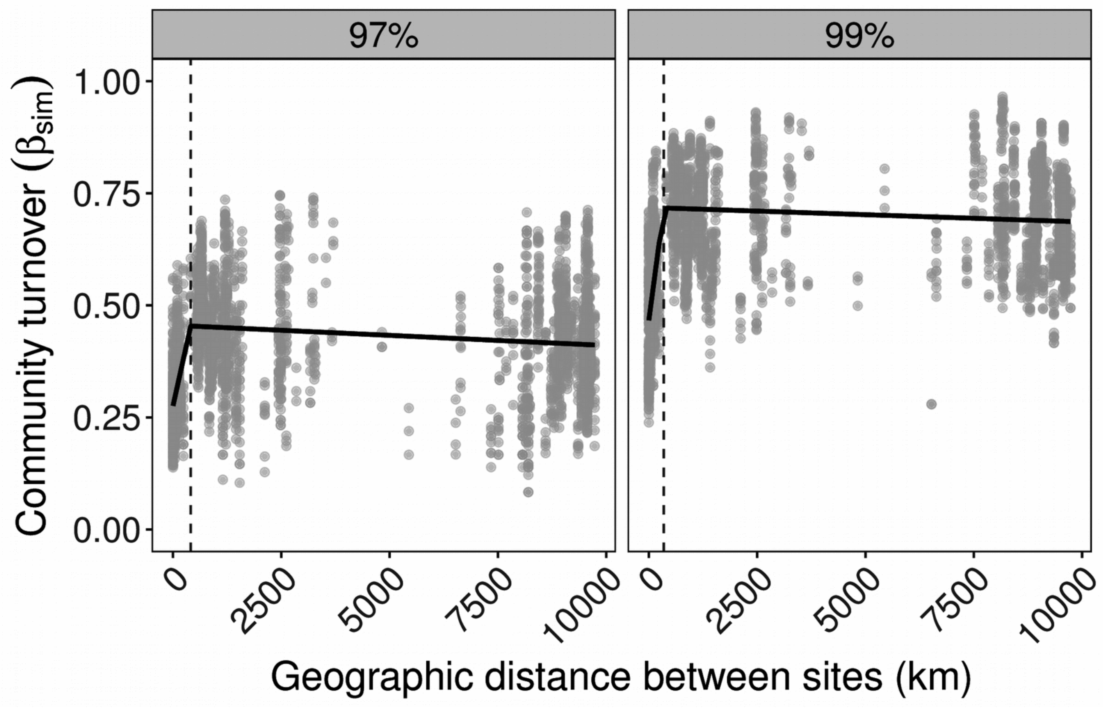


Figure 4.2 The relationship between community turnover and geographic distance, for 97% and 99% similarity operational taxonomic unit (OTU) tables. Values close to 0 indicate pairs of communities highly similar in composition, whereas values close to 1 indicate communities with few OTUs in common. Dashed lines indicate breakpoints (distance in km at which slope changes), which were estimated as 420.5 km (std. error = 46.9 km) and 334.6 km (std. error = 23.7 km). Mantel tests showed statistically significant correlation in both cases ($P < 0.0001$ in both cases).

Do Microbial Communities Cluster into Biogeographic Regions?

We determined whether archaeal communities group into biogeographic regions by applying three different clustering algorithms (UPGMA, Ward, PAM). To assess the degree of biogeographic clustering within these communities, we first determined the appropriate number of clusters (*k*) into which our communities should be grouped by examining the cluster quality (using “mean silhouette width” and “explained dissimilarity”) for values of *k* from 2-16. For the 97% dataset, statistical support for cluster solutions was poor, as the “mean silhouette width” never exceeded 0.25 for any value of *k* (Fig. 4.3A). In contrast, for the 99% dataset, all three clustering algorithms exceeded 0.25 for values of k > 12, showing that reasonable statistical support was gained when communities were grouped into more than 12 regions. All three clustering algorithms yielded similar results when assessed by the “explained dissimilarity” metric (Fig. 4.3B). “Explained dissimilarity” values > 0.9 were considered to provide good support for a given cluster solution. To satisfy this threshold, communities were grouped into 8-10 (97% dataset) or 9-12 (99% dataset) clusters, depending on the cluster algorithm used. For both 97% and 99% datasets, the Ward algorithm required the fewest clusters to reach this threshold, and UPGMA the most. We also identified the number of clusters, (*k*), that resulted in the greatest increase in “explained dissimilarity” (“knee solutions”). For the 97% dataset, this occurred when communities were clustered into 3 (PAM, Ward) or 4 (UPGMA) clusters. Whereas for the 99% dataset, the greatest increase in “explained dissimilarity” was found when communities grouped into 3 (UPGMA, Ward) or 4 (PAM) clusters.

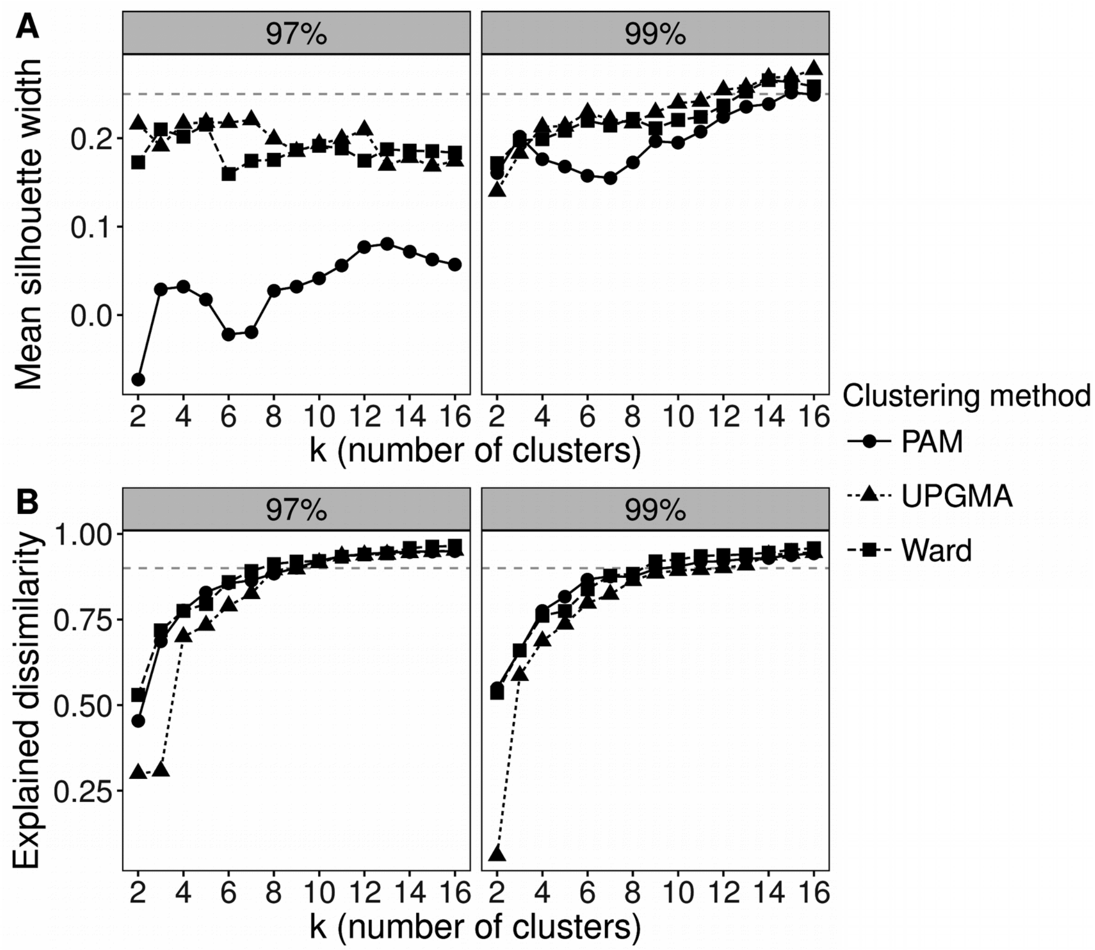


Figure 4.3 The statistical support, quantified as (A) mean silhouette width and (B) explained dissimilarity, of cluster solutions from 2 – 16 clusters, for both 97% and 99% operational taxonomic unit (OTU) datasets. Lines represent three different clustering algorithms used; partitioning around medoids (PAM), unweighted pair group method (UPGMA), and Ward clustering (Ward). In (A), silhouette widths < 0.25 (grey dotted line) are interpreted as showing poor clustering in the data and in (B) explained dissimilarity of > 0.9 indicates a good cluster solution.

We examined the spatial coherence of cluster solutions for the minimum number of clusters (*k*) required to exceed the “explained dissimilarity” threshold of 0.9, as well as solutions that yielded the greatest increase in “explained dissimilarity”. For both 97% and 99% datasets and all three clustering algorithms, mapping revealed poor spatial coherence (Fig. 4), in disagreement with hypothesis 2, suggesting little support for biogeographic regionalisation.

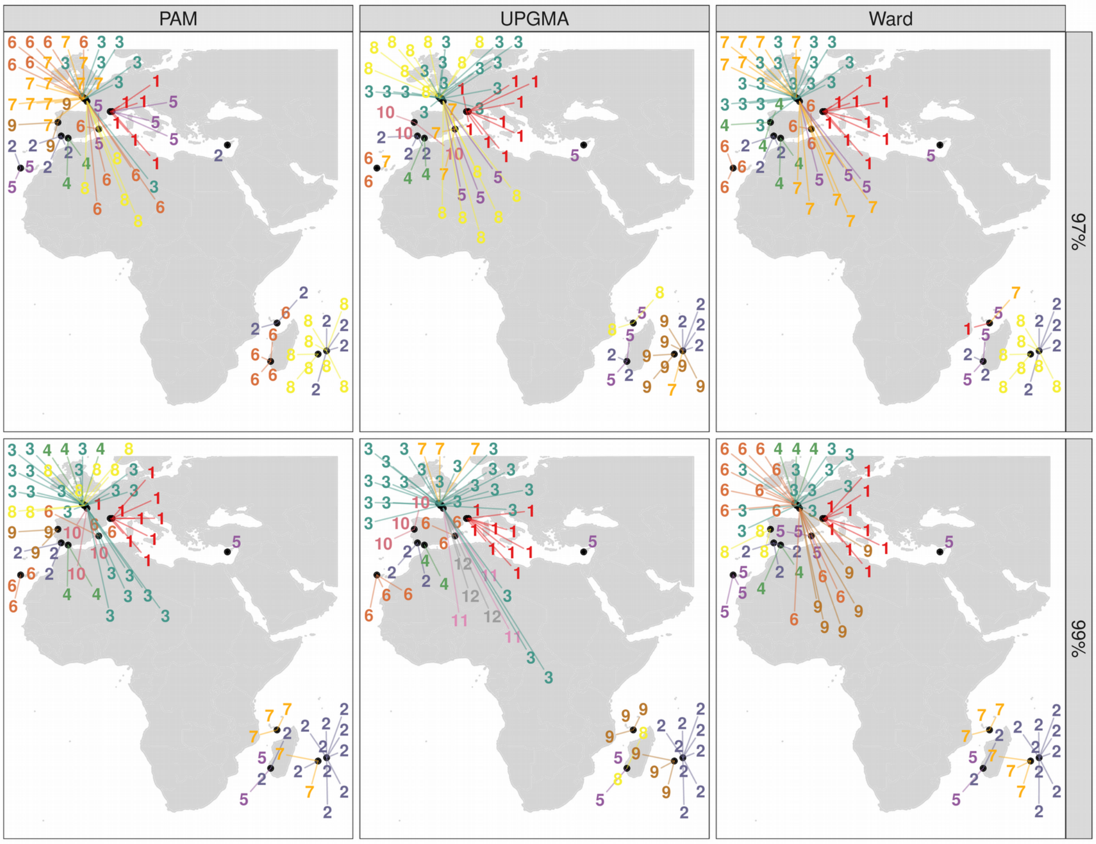


Figure 4.4 The cluster memberships (indicated by colour and number) of communities for each clustering algorithm, for both 97% and 99% operational taxonomic unit (OTU) datasets. For each algorithm, the cluster solution shown is for the minimum value of k (number of clusters) that exceeded the explained dissimilarity threshold of 0.9.

There was a large degree of mixing between communities on the West European coastline, Mediterranean, and Madagascar, counter to our expectation that communities in these regions would cluster separately. Mapping of the “knee solutions” again revealed clusters with poor spatial coherence, with many European communities clustering together with Madagascan communities (Appendix S6). NMDS and PERMANOVA showed that archaeal communities clustered only weakly by impurity, but not by grain size (Fig. 4.5).

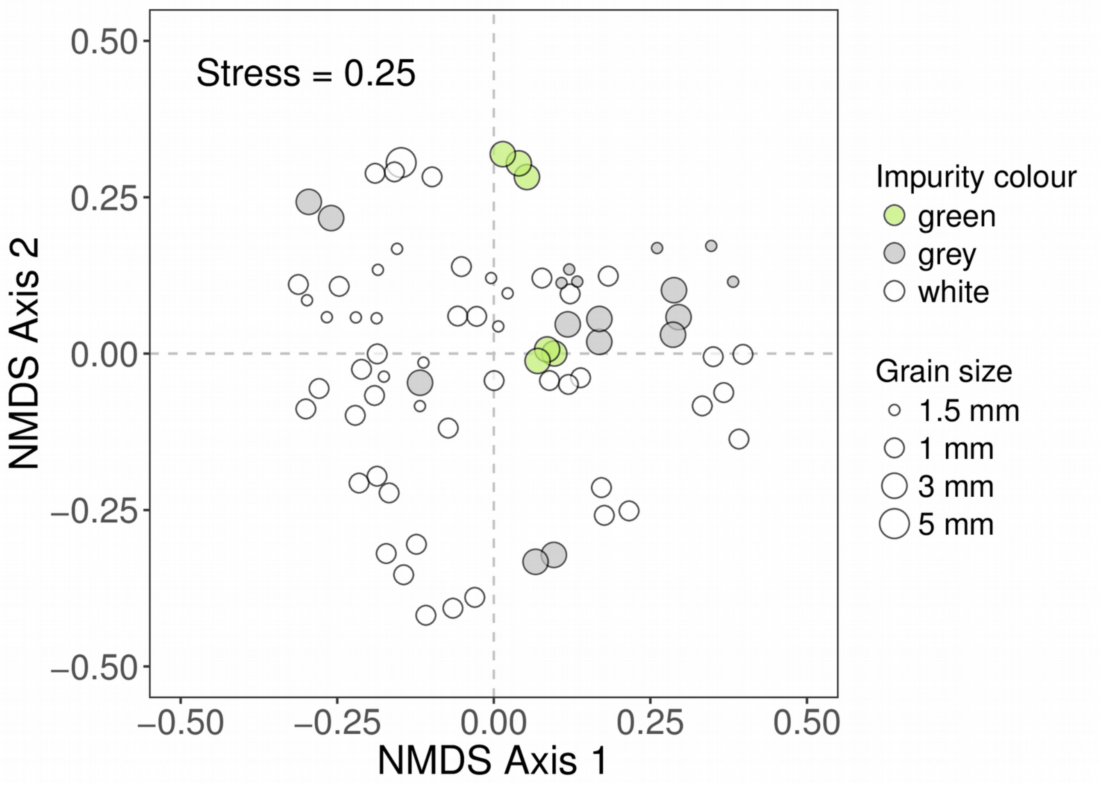


Figure 4.5 Non-metric multidimensional scaling (NMDS) analysis of halite-associated archaeal communities. Each point represents a single community, and points closer together represent compositionally more similar

communities. Communities do not appear to cluster by halite properties. Permutation-based multivariate analysis of variance (PERMANOVA) revealed that, after accounting for spatial location, grain size had no significant effect on community composition (pseudo-$F_{1,\ 55}$ = 2.09, R^2 = 0.01, P = 0.06), whilst impurity had a significant, but negligible effect (pseudo-$F_{1,\ 55}$ = 3.15, R^2 = 0.02, $P < 0.05$).

Can Certain Haloarchaeal Genera Be Used as Indicators of a Community's Biogeographic Origin?

We tested whether the abundance of certain archaeal genera could predict any of three classifiers (biogeographic region, geographic region, and nearest ocean) of a community, using random forest classification (RFC). All three classifiers performed well with comparable accuracies (ocean; error rate = 9.33%, biogeographic region; error rate = 8.45%, geographic region; error rate = 8.33%), showing that the biogeographic origin of a community can be predicted accurately from the relative abundance of individual genera. Each classifier was able to predict communities from different biogeographic origins with similar accuracy, suggesting that archaeal relative abundances were equally useful predictors for all classes. The oceanic RFC classified with similar accuracy those communities nearest to the Atlantic or Indian Ocean, with class errors of 8.9% and 10.5% respectively. The biogeographic region RFC identified communities from the Palearctic region with a 7.7% class error rate, and those from the Madagascan region with a 10.5% class error rate, whilst the geographic region RFC more accurately classified West European (class error = 7.5%) and Mediterranean (class error = 7.7%) communities, than East African communities (class error = 10.5%).

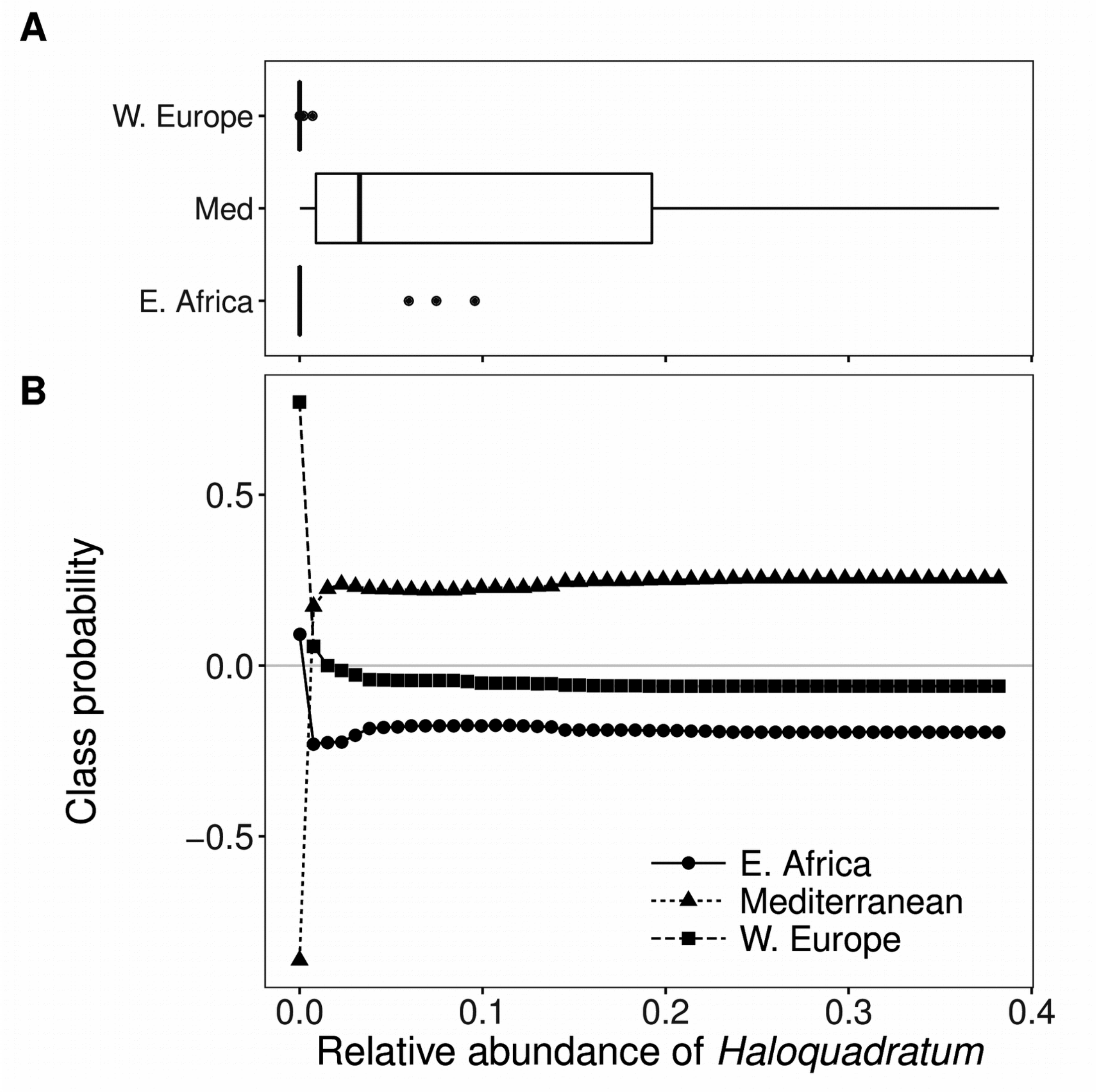


Figure 4.6 (A) The relative abundance of the genus *Haloquadratum*, in samples of different geographic origins (W. Eur = West Europe, Med = Mediterranean, E. Afr = East Africa). (B) A partial dependence plot based on a random forest classification. Class probability shows the probability that the random forest classifies a sample to each class (denoted by different line and

point styles). As the relative abundance of *Haloquadratum* was notably higher in the Mediterranean, the Mediterranean class probability increased rapidly.

To determine which archaeal genera were the best predictors of a community’s oceanic, biogeographic, or geographic origin, we quantified the importance of each variable (genus) to each RFC (Appendix S3).

Haloquadratum was the best genus for classifying geographic region, followed by *Halapricum* and *Halobaculum*. Partial dependence plots revealed that, as the relative abundance of *Haloquadratum* exceeded 0.01, the probability of the community being classified as Mediterranean increased greatly (Fig. 4.6B), reflecting its higher relative abundance in the region (Fig. 4.6A). In contrast, the genera *Halarchaeum* and *Halohasta* were the best for classifying a community’s nearest oceanic or biogeographic region, according to both metrics of variable importance (MDA and MDGI). When the relative abundance of *Halarchaeum* exceeded 0.02, a classification of the community’s nearest ocean and biogeographic region as the Indian Ocean and Madagascan biogeographic region respectively, was most likely (Appendix S7). The finding that certain archaeal genera are good predictors of a community's (bio)geographic origin supports hypothesis 3.

Discussion

We studied halite-associated Archaea to determine whether archaeal communities can be clustered into biogeographic regions comparable to those observed for most higher organisms. Our results show that, despite community turnover correlating with geographic distance over small spatial scales (< 500 km), communities do not cluster into spatially coherent biogeographic regions. We found little statistical support for clustering communities into few (2-3) biogeographic regions, which would be the number of regions expected for higher organisms such as terrestrial vertebrates (Holt *et al*., 2013) or plants (Takhtajan, 1986). Furthermore, when we clustered communities into a greater number of regions, the spatial configuration of these regions was not consistent with biogeographic regionalisation. Lastly, we demonstrated that whilst communities may not show the expected biogeographic patterns, some individual genera do, as their abundances were found to be good predictors of the biogeographic origin of the community.

Numerous studies have demonstrated that microbial communities differ over continental to regional scales (Whitaker *et al.*, 2003; Papke *et al.*, 2003; Lauber *et al.*, 2009), including studies on halophilic microbes (Pagaling *et al*., 2009; Zhaxybayeva *et al*., 2013). However, to our knowledge, no studies have quantitatively tested whether such differences are consistent with the concept of biogeographic regionalisation, thus it remains unknown as to whether the processes that shape microbial communities are capable of forming biogeographic patterns over the spatial scales relevant to other organisms. Glassman *et al*. (2015) examined fungal spore banks of soils across North America showing that community turnover was significantly related to geographic distance and, using ordination techniques, that fungal communities appeared to group in a regional manner. Consistent with our study, they found that the highest rate of community turnover occurred over sub-regional scales, as evidenced by their Mantel correlogram, which shows change from positive to negative correlation over spatial scales of ~500 km. Initially, this might indicate that microbial biogeographic regions are smaller than those defined for higher taxa, and more comparable to sub-regions. However, in our study, this idea is poorly supported by the fact that even for larger values of *k* (indicating more and smaller regions), the spatial coherence of these clusters was poor. A global study of soil fungi (Tedersoo *et al*., 2014) revealed communities that did not cluster in a spatially coherent manner, which is in contrast to the findings of Glassman *et al*. (2015) and in agreement with our results. For instance, fungal communities of Europe clustered with those of North America, and those of Oceania clustered with South America. Furthermore, a study of the bacterial communities on *Tamarix* spp. leaf surfaces showed that communities clustered in a manner at odds with their spatial configuration (Finkel *et al*., 2012). Specifically, communities from around the Dead Sea (Middle East) clustered more closely with those from the Sonoran Desert (N. America) than Mediterranean communities.

Combined with our results, these studies provide further evidence that biogeographic regionalisation may be unlikely in microbial communities.

One possible reason for no evidence of biogeographic regionalisation in these communities is that some halophilic Archaea may be differentially susceptible to long-distance dispersal. Previous studies of halophilic microbial communities have identified several potential mechanisms for long distance dispersal of haloarchaea. Despite the hostility of this environment, animal vectors may passively disperse viable Archaea between sites. Organisms such as birds and invasive invertebrates such as brine shrimp (*Artemia* spp.) have been found to harbour diverse haloarchaea (Brito-Echeverría *et al.*, 2009; Riddle *et al.*, 2013; Yim *et al*., 2015), which may help them spread between habitat islands. Furthermore, wind or human mediated dispersal of halite crystals may disperse entombed haloarchaea. Wind is known to play a role in dispersing free-living microbes over continental distances (Kellogg & Griffin, 2006; Favet *et al*., 2013) and is likely to disperse small halite crystals, along with endolithic microbes, over such distances. Human transport of salt as a commercial product and as a de-icing agent on roads may also aid the dispersal of halite endolithic communities. However, such dispersal would select for those Archaea capable of survival in halite crystals, filtering out some taxa as evidenced by the disparity between brine and halite crystal archaeal communities described previously (Henriet *et al*., 2014). Finally, dispersal via seawater could be possible for some haloarchaeal taxa, as viable cells have been isolated from seawater and coastal sediments (Rodriguez-Valera *et al.*, 1979; Purdy *et al*., 2004). Seawater may also provide a means of dispersal between ancient and modern halite deposits (McGenity *et al.*, 2008). Ancient halite deposits can become exposed in deep water horizons where they may dissolve, creating stratified deep-sea brines, which are a potential source of extremely halophilic Archaea (Antunes *et al.*, 2011). However, while short-term (~24 hour) or partial survival at seawater salinity has been found in a number of haloarchaea (Torreblanca *et al*., 1986), the majority of genera detected in this study, particularly the most abundantly detected genera, are known exclusively from hypersaline habitats, and there cells lyse at seawater salinity. Therefore, seawater is an unlikely medium for their dispersal. Furthermore, the deposition of cells from ancient halite into modern hypersaline environments would most likely occur over regional extents (e.g. due to oceanic currents), thus increasing the compositional similarity of sites within a region. Finally, even with connectivity between ancient and modern halite, there is no guarantee that those cells will become established and multiply (Jones *et al.*, 2017). Therefore, the influence of ancient haloarchaea on the clustering patterns observed here should be minimal. Even so, the degree to which other potential dispersal vectors contribute to connectivity between sites is unknown and warrants further research, as connectivity between contemporary halite deposits may be a better measure of isolation for these communities than geographic distance alone.

An alternative explanation as to why biogeographic clustering was not observed in these archaeal communities is that, environmental filtering due to physicochemical differences between the halite crystals could obscure biogeographic clustering. Within hypersaline systems, salinity (concentration of sodium chloride; NaCl) has been shown to be the predominant physicochemical variable causing environmental filtering of microbial communities (Benlloch *et al*., 2002; Casamayor *et al*., 2002; Baati *et al*., 2008; Herlemann *et al*., 2011). However, the role of physicochemical differences in structuring microbial communities between hypersaline habitats is less well known, as most research has focussed on within-site salinity gradients. Despite this, we suggest that physicochemical differences between halite samples are unlikely to explain the clustering patterns observed. Halite is an evaporite mineral, formed by the precipitation of sodium chloride from concentrated brine. Given that halite precipitates only when the concentration of NaCl (sodium chloride) exceeds approximately 32% w/v (McGenity *et al*., 2000), it is not possible for large differences in NaCl concentrations to occur between sites. Furthermore, since all the halite samples used here were formed in the same way (i.e. by progressive evaporation of seawater), the precipitation point of halite is most likely similar across sites. Other ions are also present in varying concentrations within the source brines that could have an effect on the composition of archaeal communities within the brine. Differences in the concentrations of these ions may be caused by differing underlying geology, or by differing climate. However, both geology and climate are themselves, spatially autocorrelated. Therefore, if physicochemical differences between habitats dictate differences in the microbial communities, we would expect these effects to enhance any biogeographic clustering, because sites within the same region will have physicochemically similar brines. Yet, we observed little evidence of environmental filtering on the microbial communities, suggesting that the physicochemical environment has a minimal influence on our conclusions.

The fact that these dispersal vectors are likely to selectively disperse haloarchaea with differing physiological capabilities may explain why, despite finding no evidence of biogeographic regionalisation at the community level, our population level analyses revealed several haloarchaeal genera with distinct biogeographic patterns. The square haloarchaeon, *Haloquadratum*, was found to be a good indicator of geographic region as it was found in abundance in the Mediterranean, yet was scarce in West Europe and East African. Despite this, *Haloquadratum* has been detected globally in hypersaline brines (Oh *et al.*, 2010; Podell *et al*., 2014; Di Meglio *et al*., 2016). A previous study of halite-associated Archaea found *Haloquadratum* to be a very small component of the halite-associated community, despite being highly abundant in the hypersaline brine that was the source of the halite (Henriet *et al*., 2014). Furthermore, Gramain *et al*. (2011) demonstrated that *Haloquadratum* resumed growth slowly after halite entombment compared with other haloarchaea, inferring that it is a relatively poor survivor in halite. Yet, this fails to parsimoniously explain our finding that *Haloquadratum* was an abundant member of Mediterranean halite samples. Significantly, Gramain *et al*. (2011) also observed that the recovery time of *Haloquadratum* was dramatically enhanced when co-entombed with the geographically widespread halophilic bacterium, *Salinibacter ruber* (Antón *et al*., 2008; Ventosa, *et al.*, 2015; Di Meglio *et al*., 2016). Despite the ubiquity of *S. ruber* in hypersaline environments, metabolomic profiles of geographically distant strains show biogeographic patterns (Rosselló-Mora *et al*., 2008). We speculate that the presence of a particular *S. ruber* variant or other halophilic organism in this region may facilitate the survival of *Haloquadratum* in halite, perhaps via metabolite transfer (Bolhuis *et al.*, 2004; Elevi Bardavid & Oren, 2008). We also identified *Halarchaeum* as the best genus in predicting a sample’s oceanic and biogeographic origins, as it was largely restricted to Madagascan samples. Despite this finding, *Halarchaeum* spp. have been isolated previously from globally distributed commercial salt samples (Minegishi *et al.*, 2010; Youssef *et al.*, 2012; Shimane *et al*., 2015) hinting that, despite its wide distribution, it may only be highly abundant in certain regions.

Conclusions

Overall, we found little evidence to support the existence of biogeographical regions in communities of extremely halophilic Archaea. We demonstrated that, despite finding evidence of a distance decay relationship in these communities, clustering them into regions did not produce spatially coherent regions. We suggest that the cause of this may be long distance dispersal of some haloarchaeal taxa as we identified three particularly abundant and widespread species that were universally detected across all samples. However, certain individual taxa are able to accurately indicate a given community's biogeographic origins, suggesting highly differential dispersal abilities in haloarchaea. Taken together, our results suggest that geographic distance alone may be a poor indicator of isolation in microbial communities, and that more work is needed to examine the role of connectivity in microbial biogeography.

References

Altschul SF, Gish W, Miller WT, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Antón J, Peña A, Santos F, Martínez-García M, Schmitt-Kopplin P, Rosselló-Mora R (2008) Distribution, abundance and diversity of the extremely halophilic bacterium *Salinibacter ruber*. *Saline Systems*, **4**, 15.

Antunes A, Ngugi DK, Stingl U (2011) Microbiology of the Red Sea (and other) deep-sea anoxic brine lakes. *Environmental Microbiology Reports*, **3**, 416–433.

Baas Becking LGM (1934) *Geobiologie of inleiding tot de milieukunde*. W.P. Van Stockum & Zoon, The Hague, Netherlands.

Baati H, Guermazi S, Amdouni R, Gharsallah N, Sghir A, Ammar E (2008) Prokaryotic diversity of a Tunisian multipond solar saltern. *Extremophiles*, **12**, 505–518.

Bankevich A, Nurk S, Antipov D, *et al*. (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, **19**, 455–477.

Baselga A (2010) Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, **19**, 134–143.

Benlloch S, López-López A, Casamayor EO, *et al*. (2002) Prokaryotic genetic diversity throughout the salinity gradient of a coastal solar saltern. *Environmental Microbiology*, **4**, 349–360.

Bloomfield NJ, Knerr N, Encinas-Viso F (2017) A comparison of network and clustering methods to detect biogeographical regions. *Ecography*. 10.1111/ecog.02596

Bolhuis H, Te Poele EM, Rodríguez-Valera F (2004) Isolation and cultivation of Walsby’s square archaeon. *Environmental Microbiology*, **6**, 1287–1291.

Brito-Echeverría J, López-López A, Yarza P, Antón J, Rosselló-Móra R (2009) Occurrence of *Halococcus spp*. in the nostrils salt glands of the seabird *Calonectris diomedea*. *Extremophiles*, **13**, 557–565.

Buckley LB, Jetz W (2007) Environmental and historical constraints on global patterns of amphibian richness. *Proceedings of the Royal Society B:*

Biological Sciences, **274**, 1167–1173.

Casamayor EO, Massana R, Benlloch S, *et al*. (2002) Changes in archaeal, bacterial and eukaryal assemblages along a salinity gradient by comparison of genetic fingerprinting methods in a multipond solar saltern. *Environmental Microbiology*, **4**, 338–348.

Castro-Insua A, Gómez-Rodríguez C, Baselga A (2016) Break the pattern: breakpoints in beta diversity of vertebrates are general across clades and suggest common historical causes. *Global Ecology and Biogeography*, **25**, 1279–1283.

Cole JR, Wang Q, Cardenas E, *et al*. (2009) The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, **37**, D141–D145.

Cowen RK, Paris CB, Srinivasan A (2006) Scaling of connectivity in marine populations. *Science*, **311**, 522–527.

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. *Ecology*, **88**, 2783–2792.

Dapporto L, Ciolli G, Dennis RLH, Fox R, Shreeve TG (2015) A new procedure for extrapolating turnover regionalization at mid-small spatial scales, tested on British butterflies. *Methods in Ecology and Evolution*, **6**, 1287–1297.

Dapporto L, Ramazzotti M, Fattorini S, Vila R, Talavera G, Dennis RLH (2015) recluster: Ordination Methods for the Analysis of Beta-Diversity Indices [R package version 2.8].

Di Meglio L, Santos F, Gomariz M, Almansa C, López C, Antón J, Nercessian D (2016) Seasonal dynamics of extremely halophilic microbial communities in three Argentinian salterns. *FEMS Microbiology Ecology*, **92**, fiw184.

Dinsdale EA, Edwards RA, Hall D, *et al*. (2008) Functional metagenomic profiling of nine biomes. *Nature*, **452**, 629–632.

Dumbrell AJ, Ferguson RMW, Clark DR (2016) Microbial Community Analysis by Single-Amplicon High-Throughput Next Generation Sequencing: Data Analysis -- From Raw Output to Ecology. In: *Hydrocarbon and Lipid Microbiology Protocols: Microbial Quantitation, Community Profiling and Array Approaches* (eds McGenity, T. J., Timmis, K. N., Nogales, B.), pp.

155–206. Springer, Heidelberg, Germany.

Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH (2010) Relative roles of niche and neutral processes in structuring a soil microbial community. *The ISME Journal,* **4**, 337–345.

Elevi Bardavid R, Oren A (2008) Dihydroxyacetone metabolism in *Salinibacter ruber* and in *Haloquadratum walsbyi*. *Extremophiles*, **12**, 125–131.

Falkowski PG, Fenchel T, Delong EF (2008) The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, **320**, 1034–1039.

Favet J, Lapanje A, Giongo A, *et al*. (2013) Microbial hitchhikers on intercontinental dust: catching a lift in Chad. *The ISME Journal*, **7**, 850–867.

Ficetola GF, Mazel F, Thuiller W (2017) Global determinants of zoogeographical boundaries. *Nature Ecology & Evolution*, **1**, 89.

Fierer N, Leff JW, Adams BJ, *et al*. (2012) Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences USA*, **109**, 21390–21395.

Finkel OM, Burch AY, Elad T, Huse SM, Lindow SE, Post AF, Belkin S (2012) Distance-decay relationships partially determine diversity patterns of phyllosphere bacteria on *Tamarix* trees across the Sonoran desert. *Applied and Environmental Microbiology*, **78**, 6187–6193.

Finlay BJ (2002) Global dispersal of free-living microbial eukaryote species. *Science*, **296**, 1061–1063.

Flynn JM, Brown EA, Chain FJJ, MacIsaac HJ, Cristescu ME (2015) Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. *Ecology and Evolution*, **5**, 2252–2266.

Glassman SI, Peay KG, Talbot JM, *et al*. (2015) A continental view of pine-associated ectomycorrhizal fungal spore banks: A quiescent functional guild with a strong biogeographic pattern. *New Phytologist*, **205**, 1619–1631.

Gramain A, Díaz GC, Demergasso C, Lowenstein TK, McGenity TJ (2011) Archaeal diversity along a subterranean salt core from the Salar Grande

(Chile). *Environmental Microbiology*, **13**, 2105–2121.

Hallsworth JE, Yakimov MM, Golyshin PN, *et al*. (2007) Limits of life in MgCl2-containing environments: Chaotropicity defines the window. *Environmental Microbiology*, **9**, 801–813.

Hastie TJ, Tibshirani RJ, Friedman JH (2009) *Boosting and Additive Trees* (eds Hastie TJ, Tibshirani RJ, Friedman JH). Springer, New York, NY, USA.

Henriet O, Fourmentin J, Delincé B, Mahillon J (2014) Exploring the diversity of extremely halophilic archaea in food-grade salts. *International Journal of Food Microbiology*, **191**, 36–44.

Herlemann DPR, Labrenz M, Jürgens K, Bertilsson S, Waniek JJ, Andersson AF (2011) Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *The ISME Journal*, **5**, 1571–1579.

Hijmans RJ (2016) Spherical Trigonometry [R package version 1.5-5].

Holt BG, Lessard JP, Borregaard MK, *et al*. (2013) An Update of Wallace's Zoogeographic Regions of the World. *Science*, **339**, 74–78.

Jones ML, Ramoneda J, Rivett DW, Bell T (2017) Biotic resistance shapes the influence of propagule pressure on invasion success in bacterial communities. *Ecology*, **98**, 1743–1749.

Joshi N, Fass J (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. *Available at https://github.com/najoshi/sickle*, 2011.

Kaufman L, Rousseeuw PJ (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, NY, USA.

Kellogg CA, Griffin DW (2006) Aerobiology and the global transport of desert dust. *Trends in Ecology and Evolution*, **21**, 638–644.

Kreft H, Jetz W (2010) A framework for delineating biogeographical regions based on species distributions. *Journal of Biogeography*, **37**, 2029–2053.

Kreft H, Jetz W (2013) Comment on “An Update of Wallace's Zoogeographic Regions of the World.” *Science*, **341**, 343.

Lamoreux JF, Morrison JC, Ricketts TH, Olson DM, Dinerstein E, McKnight MW, Shugart HH (2006) Global tests of biodiversity concordance and the importance of endemism. *Nature*, **440**, 212–214.

Lauber CL, Hamady M, Knight R, Fierer N (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Applied and Environmental Microbiology*, **75**, 5111–5120.

Lear G, Bellamy J, Case BS, Lee JE, Buckley HL (2014) Fine-scale spatial patterns in bacterial community composition and function within freshwater ponds. *The ISME Journal,* **8**, 1715–1726.

Liaw A, Wiener M (2002) Breiman and Cutler's Random Forests for Classification and Regression [R package version 4.6-12].

McGenity TJ, Gemmell RT, Grant WD, Stan-Lotter H (2000) Origins of halophilic microorganisms in ancient salt deposits. *Environmental Microbiology*, **2**, 243–250.

McGenity TJ, Hallsworth JE, Timmis KN (2008) Connectivity between 'ancient' and 'modern' hypersaline environments, and the salinity limits of life. In *CIESM Workshop Monographs No. 33: The Messinian Salinity Crisis from mega-deposits to microbiology – A consensus report*, 115–120.

Minegishi H, Echigo A, Nagaoka S, Kamekura M, Usami R (2010) *Halarchaeum acidiphilum* gen. nov., sp. nov., a moderately acidophilic haloarchaeon isolated from commercial solar salt. *International Journal of Systematic and Evolutionary Microbiology*, **60**, 2513–2516.

Nikolenko SI, Korobeynikov AI, Alekseyev MA (2013) BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics*, **14**, S7.

Oh D, Porter K, Russ B, Burns D, Dyall-Smith M (2010) Diversity of *Haloquadratum* and other haloarchaea in three, geographically distant, Australian saltern crystallizer ponds. *Extremophiles*, **14**, 161–169.

Oksanen J, Blanchet FG, Kindt R, *et al*. (2015) vegan: Community Ecology Package [R package version 2.3-1].

Oren A (1994) The ecology of the extremely halophilic archaea. *FEMS Microbiology Reviews*, **13**, 415–439.

Pagaling E, Wang H, Venables M, *et al*. (2009) Microbial biogeography of six salt lakes in Inner Mongolia, China, and a salt lake in Argentina. *Applied and Environmental Microbiology*, **75**, 5750–5760.

Papke RT, Ramsing NB, Bateson MM, Ward DM (2003) Geographical isolation in hot spring cyanobacteria. *Environmental Microbiology*, **5**, 650–659.

Podell S, Emerson JB, Jones CM, *et al*. (2014) Seasonal fluctuations in ionic concentrations drive microbial succession in a hypersaline lake community. *The ISME Journal*, **8**, 979–990.

Purdy KJ, Cresswell-Maynard TD, Nedwell DB, McGenity TJ, Grant WD, Timmis KN, Embley TM (2004) Isolation of haloarchaea that grow at low salinities. *Environmental Microbiology*, **6**, 591–595.

R Developement Core Team (2016) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria*.

Raskin L, Stromley JM, Rittmann BE, Stahl DA (1994) Group-specific 16S rRNA hybridization probes to describe natural communities of methanogens. *Applied and Environmental Microbiology,* **60**, 1232–1240.

Riddle MR, Baxter BK, Avery BJ (2013) Molecular identification of microorganisms associated with the brine shrimp *Artemia franciscana*. *Aquatic Biosystems*, **9**, 7.

Rodriguez-Valera F, Ruiz-Berraquero F, Ramos-Cormenzana A (1979) Isolation of extreme halophiles from seawater. *Applied and Environmental Microbiology*, **38**, 164–165.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ Preprints*, **4**, e2409v1.

Rossello-Mora R, Lucio M, Peña A, *et al*. (2008) Metabolic evidence for biogeographic isolation of the extremophilic bacterium *Salinibacter ruber*. *The ISME Journal*, **2**, 242–253.

Santos AMC, Field R, Ricklefs RE (2016) New directions in island biogeography. *Global Ecology and Biogeography*, **25**, 751–768.

Sclater PL (1858) On the General Geographic Distribution of the Members of the Class Aves. *Zoological Journal of the Linnaen Society*, **2**, 130–136.

Shimane Y, Minegishi H, Echigo A, *et al*. (2015) *Halarchaeum grantii* sp. nov., a moderately acidophilic haloarchaeon isolated from a commercial salt sample. *International Journal of Systematic and Evolutionary*

Microbiology, **65**, 3830–3835.

Sonnenfeld P (1995) The color of rock salt-A review. *Sedimentary Geology*, **94**, 267–276.

Stahl AD, Amann R (1991) Development and application of nucleic acid probes. In: *Nucleic Acid Techniques in Bacterial Systematics* (eds Stackebrandt E, Goodfellow M), pp. 205–248. John Wiley and Sons, Chichester, UK.

Takhtajan AL (1986) *The floristic regions of the world*. University of California Press, Berkeley, USA.

Tedersoo L, Bahram M, Põlme S, *et al*. (2014) Global diversity and geography of soil fungi. *Science*, **346**, 1256688.

Torreblanca M, Rodriguez-Valera F, Juez G, Ventosa A, Kamekura M, Kates M (1986) Classification of Non-alkaliphilic Halobacteria Based on Numerical Taxonomy and Polar Lipid Composition, and Description of *Haloarcula* gen. nov. and *Haloferax* gen. nov. *Systematic and Applied Microbiology*, **8**, 89–99.

Ventosa A, de la Haba RR, Sánchez-Porro C, Papke RT (2015) Microbial diversity of hypersaline environments: A metagenomic approach. *Current Opinion in Microbiology*, **25**, 80–87.

Vilhena DA, Antonelli A (2015) A network approach for identifying and delimiting biogeographical regions. *Nature Communications*, **6**, 6848.

Wallace AR (1876) *The Geographical Distribution Of Animals*. Cambridge University Press, Cambridge, UK.

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–5267.

Werner JJ, Zhou D, Caporaso JG, Knight R, Angenent LT (2012) Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. *The ISME Journal,* **6**, 1273–1276.

Whitaker RJ, Grogan D, Taylor JW (2003) Geographic barriers isolate endemic populations of hyperthermophilic Archaea. *Science*, **301**, 976–978.

Yarza P, Yilmaz P, Pruesse E, *et al*. (2014) Uniting the classification of cultured and uncultured Bacteria and Archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, **12**, 635–645.

Yim KJ, Kwon J, Cha IT, *et al*. (2015) Occurrence of viable, red-pigmented haloarchaea in the plumage of captive flamingoes. *Scientific Reports*, **5**, 16425.

Youssef NH, Ashlock-Savage KN, Elshahed MS (2012) Phylogenetic diversities and community structure of members of the extremely halophilic Archaea (order Halobacteriales) in multiple saline sediment habitats. *Applied and Environmental Microbiology*, **78**, 1332–1344.

Yu Z, García-González R, Schanbacher FL, Morrison M (2008) Evaluations of different hypervariable regions of archaeal 16S rRNA genes in profiling of methanogens by Archaea-specific PCR and denaturing gradient gel electrophoresis. *Applied and Environmental Microbiology*, **74**, 889–893.

Zak DR, Holmes WE, White DC, Peacock AD, Tilman D (2003) Plant diversity, soil microbial communities, and ecosystem function: Are there any links? *Ecology*, **84**, 2042–2050.

Zhaxybayeva O, Stepanauskas R, Mohan NR, Papke RT (2013) Cell sorting analysis of geographically separated hypersaline environments. *Extremophiles*, **17**, 265–275.

Chapter 5

A Globally Incoherent Fingerprint of Temperature on Microbial Communities

Acknowledgements: Nikolai Friberg, Terry J McGenity, Graham JC Underwood, Guy Woodward and Alex J Dumbrell

Abstract

The temperature-diversity relationship underpins the near universal increase of biodiversity towards equatorial regions. This relationship is well established across many “higher” taxa, but the effects of temperature on microbial communities are poorly understood, especially as temperature is often confounded by other variables. Therefore, the influence of temperature on the diversity, and composition of microbial communities was investigated using a series of spatially replicated, geothermal stream systems distributed around the Arctic circle (Alaska, Greenland, Iceland, Kamchatka, and Svalbard). Each stream system represents a thermal gradient ranging from ~2℃ to >30℃, allowing the effects of temperature to be quantified in the absence of confounding factors. The diversity and composition of stream sediment microbial communities was quantified using high-throughput DNA metabarcoding. The α- and β-diversity, and composition of communities was then modeled in relation to stream temperature. As expected, large differences in the temperature-diversity relationship were observed between

taxonomic groups, but unexpectedly, also between different stream systems. Temperature-abundance models revealed that Archaea, Bacteria, and Eukarya have different predicted thermal optima, hinting at niche differentiation between broad microbial taxonomic groups. Furthermore, partitioning of β-diversity metrics showed that compositional change in microbial communities along temperature gradients was predominantly due to species turnover, than nestedness (the ordered loss of species). The results show that temperature-diversity relationships do not generalise across microbial taxa, or regional communities. These findings have important implications for our understanding of the potential impacts of global warming on microbial communities, as changes in communities are likely to be highly context-dependent, and warrant further attention.

Introduction

One of the most universal patterns to emerge from macroecology is that biodiversity tends to peak in equatorial regions, and declines towards polar regions. Many hypotheses have been proposed to explain this pattern (Gaston, 2000; Hillebrand, 2004). Among them is the temperature hypothesis, which proposes that the higher metabolic rates in warmer conditions facilitates more rapid rates of evolution, thus generating higher biodiversity within warm, equatorial regions (Clarke & Gaston, 2006). This relationship has been shown in a wide range of conspicuous “higher” taxa in both terrestrial and marine environments, but more rarely in microbial communities (Furhman *et al.,* 2008).

Microorganisms fulfil key roles in most of Earth’s biogeochemical cycles and have profound impacts on ecosystem functioning. Despite this, the effects of temperature on the structure and diversity of microbial communities are not well understood. Many studies have focussed on temperature controlled mesocosm experiments (e.g. Rillig *et al*., 2002; Bálint *et al*., 2015; Yvon-Durocher *et al*., 2015; Treseder *et al*., 2016). However, experimental approaches alone are often unable to replicate the true complexity of natural systems, and are often not run over long enough time periods to yield conclusions of macroecological significance.

Observational studies have the power to incorporate the ecological complexity encountered in natural systems, and are more likely to reflect long-term ecological dynamics. However, observational studies may be confounded by other variables (Woodward *et al*., 2010) such as the physicochemical environment, yielding variable results on the effects of temperature on microbial α- and β-diversity (within and between sample diversity, respectively; Yim *et al.*, 2006; Purcell *et al.*, 2007; Cole *et al.*, 2013; Wang *et al.*, 2013; Plebani *et al.*, 2015; Zhou *et al.*, 2016). Furthermore, many observational studies focus on thermal gradients present within single sites (e.g. Yim *et al*., 2006; Cole *et al*., 2013), or single taxonomic groups, meaning that the potential for site-, or taxon-specific relationships between temperature and community structure remain unaddressed. Therefore, the effects of temperature on microbial biodiversity and community structure, remain largely unclear.

In order to combat issues of confounding factors and ecological realism, whilst also allowing the generality of relationships to be assessed, careful study system selection is required. Ideally, a study system should contain spatially replicated thermal gradients and an absence of confounding factors, whilst still incorporating the complexity and realism of natural systems. Under these criteria, geothermal systems offer great promise as “natural laboratories” in which to determine the effects of temperature on various ecological phenomena (O’Gorman *et al.*, 2014). The geothermal stream systems situated around the Arctic circle (Alaska, Greenland, Iceland, Kamchatka, and Svalbard) present a series of high latitude, replicated thermal gradients, ideal for testing the consistency of temperature effects on microbial communities. Each stream system represents a thermal gradient spanning from ~2°C to > 30°C, and the streams in each system occupy a restricted latitudinal range, minimising the effects of confounding spatial factors (O'Gorman *et al.*, 2014). Additionally, these streams are groundwater fed, meaning that minor differences in water chemistry do not confound changes in temperature (O'Gorman *et al.*, 2014). These properties have yielded great insight into the effects of warming on a number of ecological properties such as body size (Adams *et al*., 2013), productivity (Demars *et al.*, 2016; O'Gorman *et al.*, 2016) and trophic-web dynamics (Woodward *et al.*, 2010).

I therefore exploit the spatially replicated nature of these systems to investigate how temperature effects the diversity and structure of microbial communities *in situ*, across different sites and microbial taxonomic groups. Specifically, I address three main questions: (i) What is the relationship between temperature and diversity in microbial communities, and to what extent does this vary between different geographic regions? (ii) What is the effect of temperature on microbial community structure, and how does this vary within and between taxonomic groups? (iii) Is β-diversity related to changes in temperature in microbial communities, and if so, how do microbial communities change along thermal gradients? To date, this study represents the most well replicated study of microbial community dynamics in response to temperature in natural systems, and is the first to address the generality and causes of temperature effects on microbial community ecology.

Methods

Stream sediment and adjacent soil samples were collected from 5 sites distributed around the Arctic circle (Fig. 5.1), during 2013. The top 2-3 cm of stream sediments were sampled in triplicate, and adjacent soils were sampled once, to provide a comparison between different micro-habitats. Samples were collected from between 4 and 6 points along each stream system, representing a range of stream water temperatures. Samples were preserved in ethanol and stored at 4°C for downstream molecular analyses.

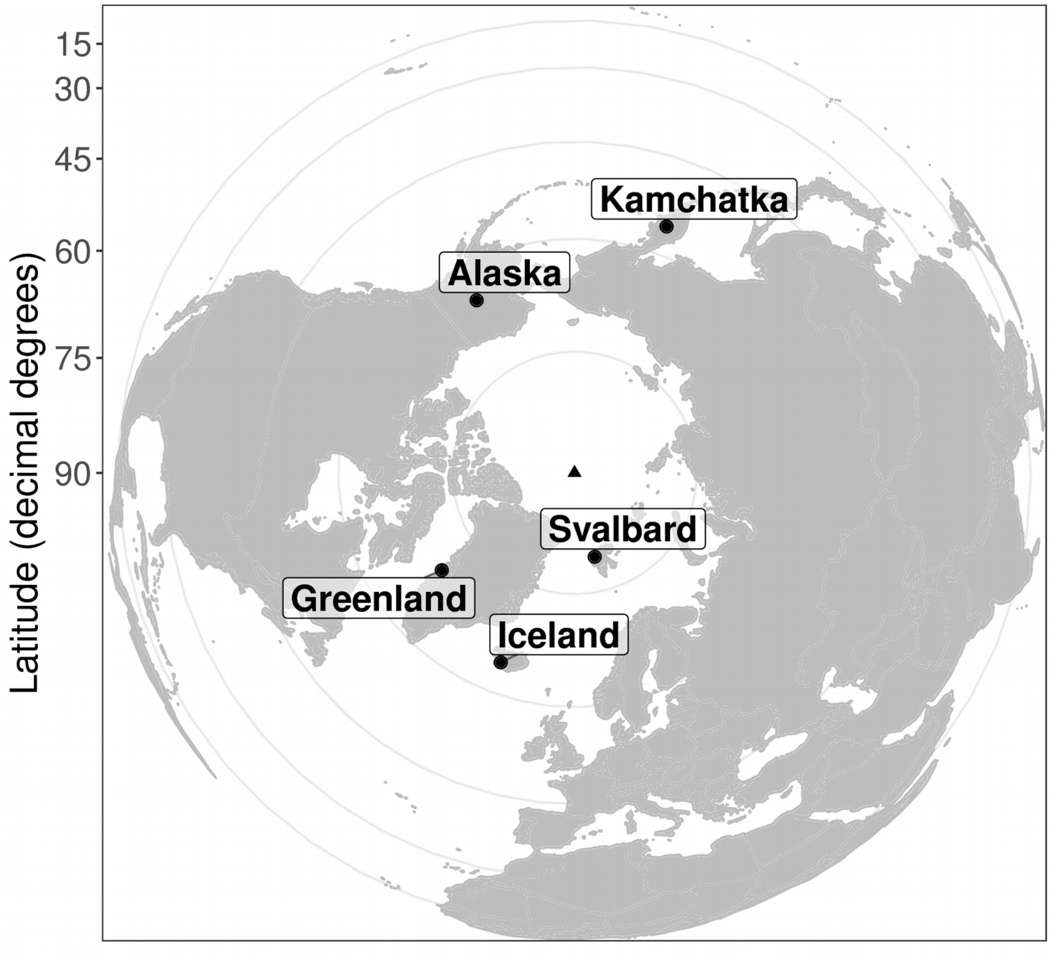


Figure 5.1 Global map of study sites centered on the North Pole (triangle). Rings represent latitudinal increments of 15°.

Molecular Analyses

To profile the microbial communities in our samples, I used next-generation sequencing on an Illumina HiSeq 2500, broadly following Illumina’s “16S Metagenomic Sequencing Library Preparation” protocol. The Nextera XT indexing strategy utilised in this protocol incorporates two rounds of PCR amplification, firstly to generate the amplicon library, and secondly to attach sample specific indices to allow multiplexing of samples. After thoroughly evaporating excess ethanol, DNA was extracted from each sample using MoBio Powersoil DNA extraction kits (MoBio Laboratories Inc., Carlsbad, CA, USA), following the manufacturers instructions. In order to test for taxonomic differences in temperature relationships, I analysed the Archaea, Bacteria, and Eukarya, by targeting them with domain specific primer sets For Bacteria, the primer pair Bact-0341-b-S-17 and S-D-Bact-0785-a-A-21 (Klindworth *et al.*, 2013) were used and for Archaea the primer pair 344F (Raskin *et al.*, 1994) and 915R (Stahl & Amann, 1991) were used, both of which target regions of the 16S rRNA gene. For Eukarya I used 574*f and 1132r (Hugerth *et al.*, 2014) to target a region of the 18S rRNA gene.

All first round PCR amplifications were carried out in 25 µl reactions, using 12.5 µl of REDTaq® ReadyMixTM (Sigma-Aldrich Co.) for Archaea and Bacteria, or KAPA HiFi HotStart ReadyMix (KAPA Biosystems, Wilmington, MA, USA) for Eukaryotes, 5 µl of each primer (1 µM), and 2.5 µl of DNA template. Where necessary, 0.05 µl of T4 Gene32 protein (Roche Diagnostics lid, Sussex, UK) was added to PCR reactions to prevent inhibition from humic acids or other inhibitors (Kreader, 1996). Thermal cycling conditions for the first round of PCR amplifications were as follows; initial denaturation of DNA at 95°C for 5 mins, followed by 32 cycles of 95°C for 45 s, primer annealing for 45 s (Archaea; 60°C, Bacteria; 55°C, Eukarya; 51°C), and 72°C for 1 min. After a final extension step at 72°C for 5 min, PCR products were held at 4°C. PCR products were purified using Agencourt AMPure XP PCR Purification beads (Beckman Coulter (UK) Ltd, High Wycombe, UK), following the protocol in the Illumina “16S Metagenomic Sequencing Library Preparation” document.

For each of the three primer sets used, a second round of PCR was carried out in 50 µl reactions with 25 µl of KAPA HiFi HotStart ReadyMix (KAPA Biosystems, Wilmington, MA, USA), 5 µl each of i5 and i7 Nextera XT index (Illumina), 10 µl of PCR water (Bioline Reagents Ltd, UK) and 5 µl of purified PCR product. Thermocycling conditions followed initial denaturation at 95°C for 3 min, followed by 8 cycles of 95°C for 30 s, 55°C for 30 s and 72 for 30 s. A final extension step was included at 72°C for 5 min, after which PCR products were held at 4°C. PCR products were, again, bead purified using Agencourt AMPure XP PCR Purification beads (Beckman Coulter Ltd, High Wycombe, UK).

Amplicon library DNA concentrations were quantified on a POLAR star Omega (BMG LABTECH GmbH, Germany) plate reader, using the PicoGreen® dsDNA assay. For each primer set, samples were pooled in equimolar concentrations. The size and concentration of the resulting pool was verified on an Agilent 2100 Bio-analyser. Sequencing was carried out on an Illumina HiSeq 2500 set to rapid run mode, which generated 2 x 300 bp reads at The Earlham Institute (formerly The Genome Analysis Centre, Norwich Research Park, Norfolk, UK).

Bioinformatic Analyses

As the length of the amplicons from the archaeal and eukaryotic specific primers exceeded the limit for obtaining reliable paired-end overlaps, I analysed only the forward reads as they were higher quality than reverse reads (data not presented). This approach has been shown to have little effect on ecological conclusions (Werner *et al.*, 2012), and has been used previously for microbial ecological analyses (Clark *et al*., 2017). For the bacterial sequences, paired-end reads could be overlapped. Sickle version 1.33 was used to quality trim the reads (Joshi & Fass, 2011) at the default quality threshold (Q20). Trimming at the 5’ end of the read was suppressed and trimmed reads shorter than 275 nucleotides, or containing ambiguous base calls (“N”) were discarded. Error correction was done using BayesHammer (Nikolenko *et al.*, 2013) as implemented in SPAdes version 3.7.1 (Nurk *et al.*, 2013), with default parameters. The bacterial 16S rRNA sequences were then paired-end aligned using PANDAseq version 1.33 (Masella *et al*., 2012) with the PEAR algorithm (Zhang *et al.*, 2014) and default parameters. VSEARCH version 2.3.2 (Rognes *et al*., 2016) was then used to de-replicate reads and remove singletons, as these are more likely to be non-biological (Behnke *et al.*, 2011; Flynn *et al.*, 2015). Reads were clustered into operational taxonomic units (OTUs) at 97% sequence similarity, using VSEARCH. Taxonomy was assigned to archaeal and bacterial OTUs using the RDP classifier and database (Wang *et al.*, 2007; Cole *et al.*, 2014), with a confidence threshold of 0.7. For the eukaryotic OTUs, I used the RDP classifier on the PR2 database, which contains curated protistan 18S rRNA gene sequences (Guillou *et al.*, 2013). Any non-domain specific OTUs from each dataset were removed prior to statistical analyses. Additionally, any eukaryotic OTUs identified as Metazoan (animal) or Embryophycaea (land plants) were removed to ensure that only microbial Eukarya were included. All analyses were conducted using the Bio-Linux 8 operating system (Field *et al.*, 2006).

Statistical Analyses

In order to determine the relationship between temperature and microbial diversity, I modeled OTU richness using negative binomial generalised linear mixed effects models (GLMMs). Using these generalised models allowed me to directly model the count data (e.g. OTU richness) without transformation, thus preserving interpretability (O'Hara & Kotze, 2010). I specified a random intercept and temperature slope with respect to site to account for, and quantify, differences in the temperature-richness relationship between sites (Bolker *et al.*, 2009). Temperature was also included as a fixed effect with both linear (T_{linear}) and quadratic ($T_{quadratic}$) terms. This allowed the modeled relationship to take on a unimodal shape, rather than a biologically less plausible linear relationship in which, richness monotonically increases or decreases. I accounted for heterogeneity in the number of sequences in each sample (library size) by including log(library size) as the first term in the model. This approach has been applied successfully in other microbial ecology studies to incorporate the effects of differential sampling depth (e.g. Bálint *et al*., 2015) and avoids the statistical pitfalls associated with rarefaction (McMurdie & Holmes, 2014). To test whether accounting for site specificity in the temperature-OTU richness relationship improved model fit, I used likelihood ratio tests to compare models against a model with only a random intercept. If the temperature-richness relationship varied between sites, the model with a site specific random slope should emerge as the better model. To quantify the variance explained these models, I calculated the marginal and conditional R^2 (Johnson, 2014). The marginal R^2 (R^2_{marg}) provides a metric of variance explained by the fixed effects only (e.g. log(no. sequences) + temperature + temperature^2), whereas the conditional R^2 (R^2_{cond}) quantifies the variance explained by fixed and random effects together (site specific intercepts and temperature slopes). As an additional measure of model fit, I calculated the square of Pearson's correlation coefficient between observed and fitted values. Modeling was carried out independently for archaeal, bacterial and eukaryotic datasets.

To examine the effect of temperature on microbes at the population level, I used multivariate generalised linear models (Wang *et al.*, 2012). These models are better able to model the strong mean-variance relationship present in over-dispersed count data such as OTU counts (Warton *et al.*, 2016), providing more robust results than other methods, such as ordination methods. To account for different library sizes, I included log(library size) as an offset term. This approach accounts for differential sample sizes by assuming proportionality between sequence numbers and OTU abundances (e.g. doubling the sequence numbers will double the OTU abundances), which is the expected relationship. Again, temperature was included as both a linear and quadratic term to allow responses to take on a unimodal shape. I modeled OTUs that occurred in three or more samples, as OTUs confined to fewer samples were unlikely to yield much information. In order to test whether different taxa show different temperature preferences, I calculated their predicted thermal optima (T_{opt}). To do this, I used the estimated coefficients for both linear (T_{linear}) and quadratic ($T_{quadratic}$) temperature terms and calculated the thermal optima for each OTU using equation 1. I then tested for differences in T_{opt} between taxa with an ANOVA test, followed by an exploratory Tukey HSD *post-hoc* test.

equation (1): $T_{opt} = - T_{linear} / (2 * T_{quadratic})$

In order to test how β-diversity in microbial communities is related to temperature, I used distance-based correlations. Most methods of quantifying β-diversity merge the two ways in which communities can change; through nestedness (the ordered loss of species, Ulrich & Almeida-Neto, 2012) and pure species turnover (Baselga, 2010). Therefore, to understand how communities change along thermal gradients, I partitioned β-diversity into its turnover and nestedness components, β_{sim} and β_{nes} respectively (Baselga, 2010; Baselga & Orme, 2012). The equations for β_{sim} and β_{nes} are shown below (equations 1 and 2), where *a* is the number of species shared by two communities, *b* is the number of species that only appear in site 1, and *c* is the number of species that appear only in site 2:

equation 1: $$\beta_{sim} = \frac{min(b,c)}{a+min(b,c)}$$

equation 2: $$\beta_{nes} = \frac{max(b,c)-min(b,c)}{2a+min(b,c)+max(b,c)} \times \frac{a}{a+min(b,c)}$$

If β_{sim} is related to temperature, then different species are replaced along the temperature gradient, whereas if β_{nes} is related to temperature, species are not replaced, but decrease or increase their abundance such that communities along the temperature gradient are non-random subsets of each other. I quantified differences in temperature between samples using

Euclidean distance. To test the correlation between the two β-diversity components ($\beta_{sim,}$ β_{nes}) and temperature, I used Mantel tests with Pearson's correlation coefficient and 10,000 Montecarlo permutations.

All statistical analyses were carried out using the R programming language (R Core Team, 2016) with the "lme4" (Bates *et al.*, 2015), "piecewiseSEM" (Lefcheck, 2016), "mvabund" (Wang *et al.*, 2012) and "vegan" (Oksanen *et al.*, 2015) packages. Graphics were created with the "ggplot2" package (Wickham, 2009).

Results

From 1.5 million archaeal, 8.3 million bacterial, and 4 million eukaryotic quality controlled reads, 865 archaeal, 24,658 bacterial, and 21,845 eukaryotic non-singleton OTUs were detected. Archaeal communities were dominated by unclassified OTUs from the phyla Woesarchaeota and Euryarchaeota. Bacterial communities were highly diverse and 8 phyla were represented by > 1,000 OTUs (Acidobacteria, Actinobacteria, Bacteroidetes, Firmicutes, Parcubacteria, Planctomycetes, Proteobacteria and, Verrucomicrobia). Discarding Metazoan and Embryophyte OTUs reduced the number of Eukaryotic OTUs included in our analysis from 21,845 to 18,224. The vast majority (80.2%) of Eukaryotic OTUs were not identified beyond "Eukaryota", suggesting taxonomic novelty and/or poor database coverage during taxonomic assignment. Of those that were identified, many OTUs were identified as fungi, with Cercozoa, Ochrophyta, Ciliophora, Chlorophyta, and Stramenopiles all common.

The Relationship Between Microbial Diversity and Temperature

I modeled the relationship between temperature and microbial diversity using GLMMs. In archaeal and bacterial communities, the relationship between temperature and OTU richness was not statistically significant (for both linear and quadratic terms). In both cases, the random temperature slopes indicated considerable heterogeneity between sites. In particular, archaeal and bacterial communities in Iceland and Svalbard showed dramatically different relationships (Fig. 5.2) compared to Alaska, Greenland, and Kamchatka, which all followed the expected unimodal shape.

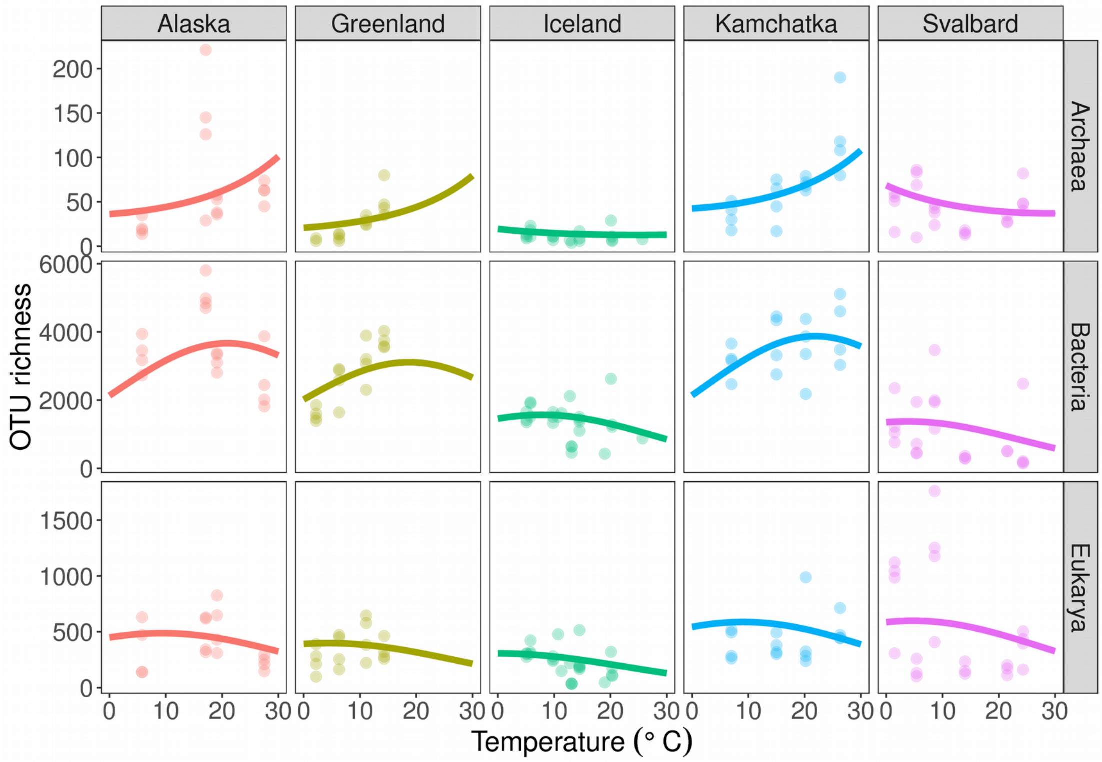


Figure 5.2 The relationship between α-diversity (OTU richness) and temperature in all sites for Archaea, Bacteria, and Eukarya. Lines represent the fit of generalised linear mixed effects models with site dependant random intercepts and temperature slopes. Note that OTU richness values represents raw counts that have not been normalised according to sequencing coverage, whereas model fits have taken differential coverage into account.

Analysis of deviance suggested that for both Archaea and Bacteria, accounting for site specific temperature slopes significantly improved model fit (Archaea; $\chi^2(2) = 9.68$, $P < 0.01$, Bacteria; $\chi^2(2) = 8.54$, $P < 0.05$) . Additionally, conditional R^2 values were far larger than marginal values

(Archaea; R^2_{marg} = 0.44, R^2_{cond} = 0.93, Bacteria; R^2_{marg} = 0.06, R^2_{cond} = 0.99), indicating that accounting for site specific differences in the relationship greatly increased the variance explained by the models. When Iceland and Svalbard communities were removed, T_{linear} was significant for Archaea (z = 4.34, coefficient = 0.53, $P < 0.001$), and both terms became significant for Bacteria (T_{linear}; z = 2.86, coefficient = 0.14, $P < 0.01$, $T_{quadratic}$; z = -2.51, coefficient = -0.08, $P < 0.05$).

In Eukaryotic communities, T_{linear} was significant and negative (z = -2.29, coefficient = -0.16, $P < 0.05$), but $T_{quadratic}$ was not significant, suggesting that the relationship with richness was monotonic. Estimation of the random slopes between sites showed that the relationship was also consistently observed, with a similar response curve fitted to each country. Including random slopes in the model did not significantly improve fit over a model without a random slope term ($\chi^2(2) = 0.74$, $P = 0.69$). Furthermore, the difference between marginal and conditional R^2 values was smaller (R^2_{marg} = 0.64, R^2_{cond} = 0.99) than for the Archaea or Bacteria, showing that random effects did not explain as much extra variation. Predictive performance of models was best for Bacteria, and worst for Eukarya (Fig. S1).

Effects of Temperature on Microbial Community Structure

To determine how temperature affects the structure of microbial communities, I modeled the abundance of OTUs with multivariate GLMs. I found that 252 archaeal, 8,352 bacterial, and 2,437 eukaryotic OTU abundances were significantly (P < 0.05) related to the linear temperature term (T_{linear}), whereas fewer OTU abundances were significantly related to $T_{quadratic}$ (Archaea; 170 OTUs, Bacteria; 5,865 OTUs, Eukarya; 1,550 OTUs). Many OTUs from each taxonomic group responded significantly to both temperature terms (Archaea; 112 OTUs, Bacteria; 3,185 OTUs, Eukarya; 1,289 OTUs). Of all the OTUs that responded significantly to both linear and quadratic temperature terms, 92.3% had negative coefficients for the quadratic term, meaning that the relationship between abundance and temperature formed the expected unimodal hump shape. For these OTUs, thermal optima (T_{opt}) were calculated using estimated coefficients (T_{linear} and $T_{quadratic}$). Archaea were found to have the highest T_{opt} of ~19°C whereas Eukarya had the lowest at ~13.3°C (Fig. 5.3A). Analysis of variance showed that estimates of T_{opt} varied significantly between taxa ($F_{2, 4232}$ = 69.7, *P* < 0.0001). Tukey's HSD test revealed that T_{opt} was significantly different between all three domains (*P* < 0.05 in all cases, Fig. 5.3B).

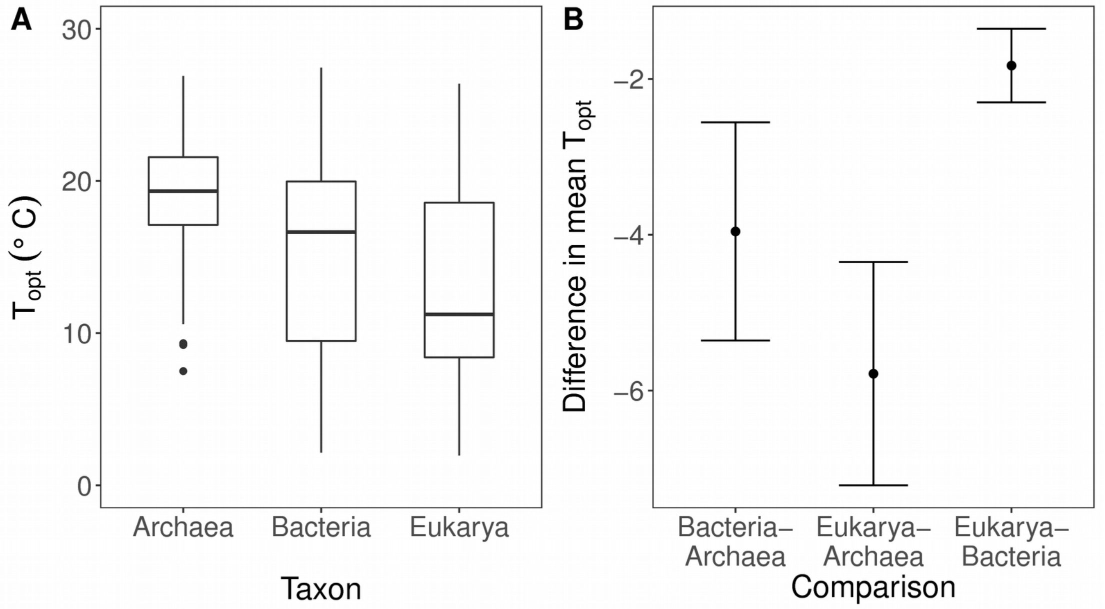


Figure 5.3 (A) The estimated thermal optima (T_{opt}) of OTUs that responded significantly to both temperature terms (T_{linear} and $T_{quadratic}$) and (B) the difference in mean T_{opt} in degrees-Celsius between the microbial taxa.

The Relationship Between β-Diversity and Temperature

To determine how microbial communities change in relation to temperature gradients, I partitioned β-diversity into its turnover and nestedness components, and correlated these with changes in temperature. For all three taxonomic groups, species turnover (β_{sim}) between communities was significantly and positively correlated with changes in temperature (Fig. 5.4A-C). This correlation was stronger for Bacteria and Eukarya (Bacteria; r_{Mantel} = 0.30, $P < 0.001$), Eukarya; r_{Mantel} = 0.25, $P < 0.001$) than for Archaea (r_{Mantel} = 0.13, $P < 0.001$). In contrast, the nestedness component of β-diversity was not related to changes in temperature in any of the three taxonomic groups (Fig. 5.4D-F). For archaeal and bacterial communities, the correlation was both non-significant and weak (Archaea; r_{Mantel} = -0.01, P = 0.61, Bacteria; r_{Mantel} = -0.04, P = 0.83, Eukarya; r_{Mantel} = -0.15, P = 1).

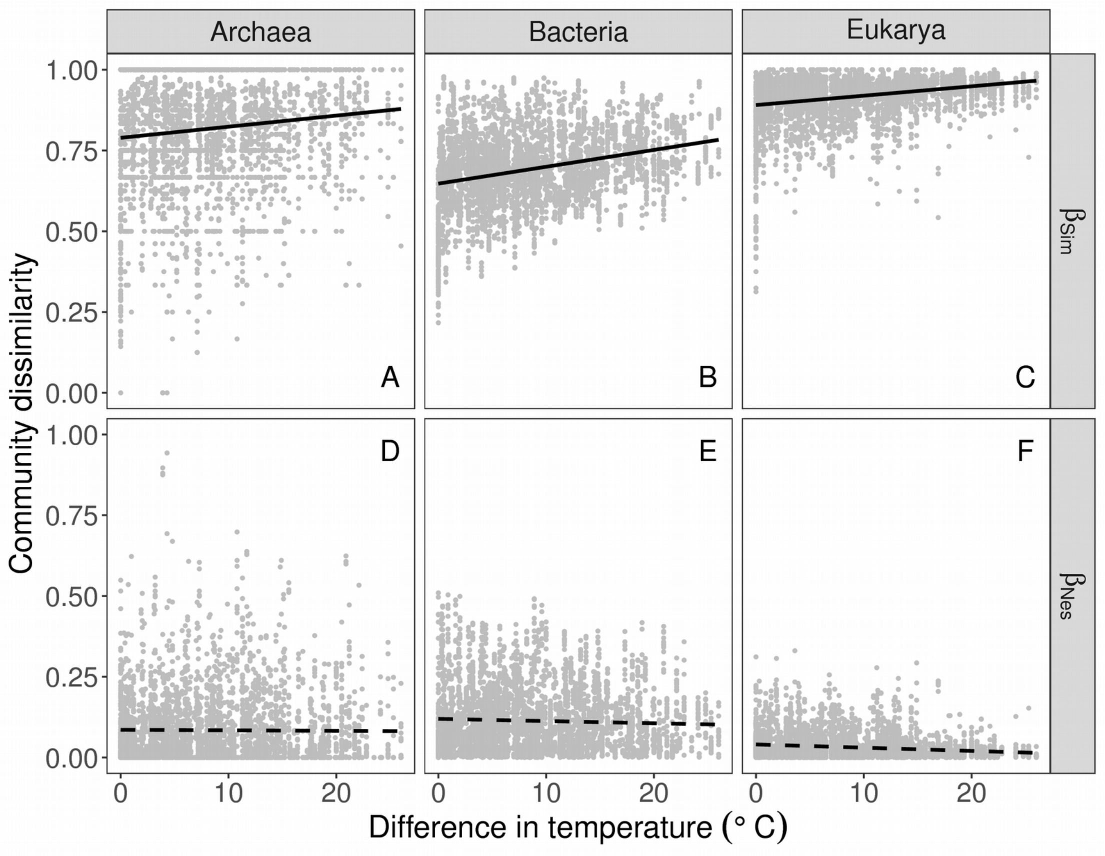


Figure 5.4 The relationship between β-diversity of microbial communities and pairwise differences in temperature between samples. Panels A-C show pure species turnover (β_{sim}) whereas panels D-F show nestedness (β_{nes}). Lines represent the fit of a linear model and are presented to show the direction of correlations, solid lines show statistically significant correlations whereas dashed lines show non-significant correlations.

Discussion

Temperature-Diversity Relationships

I examined microbial communities along five parallel thermal gradients situated around the Arctic Circle. I found that the relationship between temperature and α-diversity was inconsistent across study sites for archaeal and bacterial communities. Iceland and Svalbard communities showed flatter relationships between α-diversity and temperature, and were comparatively depauperate of species in the warmer streams, especially so for archaeal communities. The inconsistency of reported temperature-diversity relationships has previously been discussed by Sharp *et al*. (2014), who proposed that experimental factors such as lack of sequencing depth or sampling effort, too small a temperature gradient, and poor statistical methodology as explanations for this variability. However, these results clearly demonstrate variability in this relationships in an ideal model system, with high sequencing coverage, and robust statistical methodologies. Instead of being an experimental or methodological artefact, this study shows that variability in temperature-diversity relationships is a real ecological phenomenon.

I propose that this finding is due to differential metacommunity dynamics between temperature gradients. The metacommunity describes a set of local communities that are linked by dispersal (Logue *et al*., 2011), and therefore the total diversity of organisms in a site. Within the stream systems studied here, different temperature communities are separated by a maximum of a few kilometers (Woodward *et al*., 2010). Given that microbes show high potential for dispersal (e.g. Favet *et al*., 2013), and that streams offer highly connected dispersal networks (Astorga *et al*., 2012), it seems highly likely that communities within temperature gradient are linked by dispersal. Furthermore, the large geographic distances separating sites suggests that dispersal between them should be minimal. Together, these properties show that the communities studied here offer a relevant system in which to apply metacommunity theory. The metacommunity of each site contains the species available to fill the temperature niche space, thus determining the diversity of communities at different temperatures. Metacommunities are themselves determined by another hierarchy of ecological processes. Differences in regional metacommunities can be caused by differing landscape properties and varying rates and sources of propagule dispersal (Ryberg & Fitzgerald, 2015). The influence of different landscape properties (such as habitat area and isolation) to influence metacommunity composition is analogous to classic Island biogeography theory (MacArthur & Wilson, 1967). In island biogeography, small and more isolated islands tend to be less diverse than larger, more highly connected islands, because fewer dispersing propagules reach these islands, and extinction rates are higher (MacArthur & Wilson, 1967).

In this study, the temperature-diversity relationships of two sites (Svalbard and Iceland) clearly stand out from the rest as they were flatter, and generally contained fewer OTUs. These sites were also the smallest, and most isolated sites, being relatively small islands, separated from large land masses by at least several hundred kilometers. In contrast, the other three sites (Alaska, Greenland, and Kamchatka) are all connected to large, continental land masses. It is perhaps likely that these smaller sites experience reduced rates of dispersal and high local extinction rates, thereby reducing the diversity present in their metacommunities (as evidenced by their lower overall diversity). These relatively species depauperate metacommunities might therefore contain fewer species to fill the available temperature niche space, resulting in markedly different temperature-diversity relationships. This idea is congruent with the metacommunity framework described by Liebold *et al*. (2004), and in particular their description of species-sorting as the process that links local community composition to the metacommunity.

Previous evidence for a role of metacommunity dynamics in determining how microbial communities respond to the environment is scant (though see Lindström & Langenheder (2011) for a comparison of metacommunity concepts with microbial community assembly concepts). However, the presence of different regional metacommunities is a parsimonious explanation for the diversity patterns observed here, and may be of use in explaining other instances where inconsistent community-environment relationships have been found (e.g. Telford *et al*., 2006).

Effects of Temperature on Microbial Community Structure

Multivariate abundance models revealed different responses to temperature by different microbial taxa, highlighting that even at broad taxonomic levels, microbial communities exhibit niche differentiation when faced with strong environmental gradients (Dumbrell *et al*., 2010). Interestingly, temperature optima predicted by the models, generally agreed well with previous knowledge on the physiology and ecology of major microbial lineages. For instance, Eukaryotes are known to have a lower thermal limit to life (Tansey & Brock, 1972; Rothschild & Mancinelli, 2001) than Bacteria, or Archaea (Blöchl *et al.*, 1997; Kashefi & Lovley, 2003) and it has been shown that Archaea generally dominate numerically both in terms of individuals and diversity in warmer habitats.

The finding of thermal niche differentiation between taxonomic groups at the domain-level has important implications for understanding ecosystem functionality. Certain functional roles, of importance at the ecosystem level, are also conserved within these taxonomic groups. Of particular interest in the context of current climate change, is the microbial cycling of methane. Methanogenesis is an exclusively archaeal process (Liu & Whitman, 2008), whereas methanotrophy is predominantly (but not exclusively) carried out by Bacteria, in the Gamma- and Alpha-proteobacteria (McDonald *et al.*, 2008), as well as the Verrucomicrobia (Dunfield *et al.*, 2007). Previous studies have shown that the organisms linked to these processes respond to temperature (Fey & Conrad, 2000, 2003; Börjesson *et al.*, 2004; Mohanty *et al.*, 2007; Fu *et al.*, 2015). Our results suggest that, under warming, microbial communities may switch from predominantly methanotrophic to methanogenic. As methane is a potent greenhouse gas, this represents the potential for a positive feedback cycle as greater atmospheric [methane] will result in increased warming. This possibility is further backed up by stoichiometric analyses which show that the ratio of CH_4:CO_2 increases with increasing temperature (Yvon-Durocher *et al*., 2014). Further analyses of the activity and abundance of methane cycling microorganisms in these habitats warrants further research, and should help to link the niche differentiation of functionally antagonistic microorganisms to ecosystem processes.

Additionally, Eukaryotes were found to have significantly lower predicted thermal optima than Bacteria, or Archaea. Microbial Eukarya, such as protozoa, occupy trophic levels above Archaea and Bacteria, and therefore, complex trophic structure relies on the presence of eukaryotic organisms. Consequently, there is potential for warming to considerably reduce the complexity of trophic webs by decreasing the abundance and diversity of Eukarya in these habitats. In warmer communities, this could lead to predator escape for Bacteria and Archaea, with unknown ecological consequences. Whilst experimental studies show that, in cooler temperatures, predation rates of microbial Eukarya on Bacteria are positively related to temperature

(Sarmento *et al.*, 2010). Thus, the effects of warming on microbial food webs may be dependent on the temperature that communities currently occupy.

β-Diversity Patterns in Relation to Temperature

I found that in all three domains, the species turnover component of β-diversity correlated with changes in temperature, but the nestedness component did not. This shows that different species were present along the thermal gradients, providing further evidence of niche differentiation among the microbial communities. It is well known that distinct species of Archaea and Bacteria occupy warm habitats such as geothermal springs, and therefore not surprising that we detected niche differentiation along the thermal gradients for these communities. Nevertheless, whilst this result is not unexpected, it is novel and important in providing proof that the influence of temperature on microbial β-diversity is due to species replacement, and simply the loss of species with small temperature niches. Interestingly, a study of stream ciliate communities, in the same (Icelandic) stream system as studied here, found that the nestedness component of β-diversity better correlated with changes in temperature (Plebani *et al.*, 2015). In comparison to Bacteria and Archaea, relatively few thermophilic Ciliates are known (Hu, 2014). Therefore, nestedness may be a plausible relationship for Ciliates, as higher temperature communities may comprise of a subset of Ciliates with broad thermal niches, rather than mesophilic species being replaced by thermophilic species (as would be expected if species turnover was dominant).

Conclusions

I investigated how the diversity and composition of microbial communities is related to temperature, using a series of thermal gradients found in geothermal stream systems. In contrast to expectation, I show that discrepancies in the microbial diversity-temperature relationships are not merely methodological or statistical artefacts and do in fact, represent a real ecological phenomenon that may be best explained by differential regional metacommunity dynamics. Furthermore, the results show that even at broad taxonomic levels, microorganisms show thermal niche differentiation, adding to the growing body of evidence for thermal niche differentiation among microbial consortia (Garcia-Pichel *et al.*, 2013). Finally, I show that species turnover, and not nestedness, is the dominant process underlying β-diversity patterns in relation to temperature. The results have serious implications for the composition and functioning of microbial communities under global warming as shifts in diversity may be spatially inconsistent due to metacommunity processes, and communities may show considerable shifts in functionality, especially where functional roles are taxonomically conserved.

References

Adams GL, Pichler DE, Cox EJ, O'Gorman EJ, Seeney A, Woodward G, Reuman DC (2013) Diatoms can be an important exception to temperature-size rules at species and community levels of organization. *Global Change Biology*, **19**, 3540–3552.

Astorga A, Oksanen J, Luoto M, Soininen J, Virtanen R, Muotka T (2012) Distance decay of similarity in freshwater communities: do macro-and microorganisms follow the same rules?. *Global Ecology and Biogeography*, **21**, 365-375.

Bálint M, Bartha L, O'Hara RB, *et al*. (2015) Relocation, high-latitude warming and host genetic identity shape the foliar fungal microbiome of poplars. *Molecular Ecology*, **24**, 235–248.

Baselga A (2010) Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, **19**, 134–143.

Baselga A, Orme CDL (2012) Betapart: An R package for the study of beta diversity. *Methods in Ecology and Evolution*, **3**, 808–812.

Bates D, Mächler M, Bolker BM, Walker SC (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48.

Baas Becking LGM (1934) *Geobiologie of inleiding tot de milieukunde*. W.P. Van Stockum & Zoon, The Hague, Netherlands.

Behnke A, Engel M, Christen R, Nebel M, Klein RR, Stoeck T (2011) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environmental Microbiology*, **13**, 340–349.

Bell T, Ager D, Song JI, Newman JA, Thompson IP, Lilley AK, van der Gast CJ (2005) Larger islands house more bacterial taxa. *Science*, **308**, 1884.

Blöchl E, Rachel R, Burggraf S, Hafenbradl D, Jannasch HW, Stetter KO (1997) *Pyrolobus fumarii*, gen. and sp. nov., represents a novel group of archaea, extending the upper temperature limit for life to 113 degrees C. *Extremophiles*, **1**, 14–21.

Bolgovics Á, Ács É, Várbíró G, Görgényi J, Borics G (2015) Species area relationship (SAR) for benthic diatoms: a study on aquatic islands. *Hydrobiologia*, **764**, 91–102.

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSSS (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, **24**, 127–135.

Börjesson G, Sundh I, Svensson B (2004) Microbial oxidation of CH4 at different temperatures in landfill cover soils. *FEMS Microbiology Ecology*, **48**, 305–312.

Clark DR, Mathieu M, Mourot L, Dufossé L, Underwood JC, Dumbrell AJ, McGenity TJ (2017) Biogeography at the Limits of Life: Do Extremophilic Microbial Communities Show Biogeographic Regionalisation? *Global Ecology and Biogeography,* In press.

Clarke A, Gaston KJ (2006) Climate, energy and diversity. *Proceedings of the Royal Society of London B: Biological Sciences*, **273**, 2257-2266.

Cole JK, Peacock JP, Dodsworth JA, *et al*. (2013) Sediment microbial communities in Great Boiling Spring are controlled by temperature and distinct from water communities. *The ISME Journal*, **7**, 718–729.

Cole JR, Wang Q, Fish JA, *et al*. (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, **42**, D633-D642.

Declerck SAJ, Winter C, Shurin JB, Suttle CA, Matthews B (2013) Effects of patch connectivity and heterogeneity on metacommunity structure of planktonic bacteria and viruses. *The ISME Journal*, **7**, 533–542.

Demars BOL, Gíslason GM, Ólafsson JS, *et al*. (2016) Impact of warming on CO2 emissions from streams countered by aquatic photosynthesis. *Nature Geoscience*, **9**, 758–761.

Domaizon I, Lepere C, Debroas D, *et al*. (2012) Short-term responses of unicellular planktonic eukaryotes to increases in temperature and UVB radiation. *BMC Microbiology*, **12**, 202.

Dunfield PF, Yuryev A, Senin P, *et al*. (2007) Methane oxidation by an extremely acidophilic bacterium of the phylum Verrucomicrobia. *Nature*, **450**, 879–882.

Favet J, Lapanje A, Giongo A, *et al*. (2013) Microbial hitchhikers on intercontinental dust: catching a lift in Chad. *The ISME Journal*, **7**, 850–867.

Fey A, Conrad R (2000) Effect of temperature on carbon and electron flow and on the archaeal community in methanogenic rice field soil. *Applied* *and Environmental Microbiology*, **66**, 4790–4797.

Fey A, Conrad R (2003) Effect of temperature on the rate limiting step in the methanogenic degradation pathway in rice field soil. *Soil Biology and Biochemistry*, **35**, 1–8.

Field D, Tiwari B, Booth T, Houten S, Swan D, Bertrand N, Thurston M (2006) Open software for biologists: from famine to feast. *Nature Biotechnology*, **24**, 801–803.

Flynn JM, Brown EA, Chain FJJ, MacIsaac HJ, Cristescu ME (2015) Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. *Ecology and Evolution*, **5**, 2252–2266.

Fu L, Song T, Lu Y (2015) Snapshot of methanogen sensitivity to temperature in Zoige wetland from Tibetan plateau. *Frontiers in Microbiology*, **6**, 131.

Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, Green JL, Brown JH (2008) A latitudinal diversity gradient in planktonic marine bacteria. *Proceedings of the National Academy of Sciences USA*, **105**, 7774–7778.

Garcia-Pichel F, Loza V, Marusenko Y, Mateo P, Potrafka RM (2013) Temperature drives the continental-scale distribution of key microbes in topsoil communities. *Science*, **340**, 1574–1577.

Gaston KJ (2000) Global patterns in biodiversity. *Nature*, **405**, 220-227.

Guillou L, Bachar D, Audic S, *et al*. (2013) The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, **41**, D597-D604.

Hillebrand H (2004) On the generality of the latitudinal diversity gradient. *The American Naturalist*, **163**, 192-211.

Hugerth LW, Muller EEL, Hu YOO, *et al*. (2014) Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PloS ONE*, **9**, e95567.

Hu X (2014) Ciliates in extreme environments. *Journal of Eukaryotic Microbiology*, **61**, 410-418.

Johnson PCD (2014) Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. *Methods in Ecology and Evolution*, **5**, 944–946.

Joshi N, Fass J (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. *Available at https://github.com/najoshi/sickle*, 2011.

Kashefi K, Lovley DR (2003) Extending the upper temperature limit for life. *Science*, **301**, 934.

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, **41**, e1.

Lefcheck JS (2016) piecewiseSEM: Piecewise structural equation modelling in R for ecology, evolution, and systematics. *Methods in Ecology and Evolution*, **7**, 573–579.

Leibold MA, Holyoak M, Mouquet N, *et al*. (2004) The metacommunity concept: a framework for multi-scale community ecology. *Ecology Letters*, **7**, 601-613.

Lindström ES, Langenheder S (2012) Local and regional factors influencing bacterial community assembly. *Environmental Microbiology Reports*, **4**, 1-9.

Liu Y, Whitman WB (2008) Metabolic, phylogenetic, and ecological diversity of the methanogenic archaea. *Annals of the New York Academy of Sciences*, **1125**, 171-189.

MacArthur RH, Wilson EO (1967) *The theory of island biogeography*. Princeton University Press, Princeton, NJ, USA.

Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD (2012) PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics*, **13**, 1–7.

McDonald IR, Bodrossy L, Chen Y, Murrell JC (2008) Molecular ecology techniques for the study of aerobic methanotrophs. *Applied and Environmental Microbiology*, **74**, 1305–1315.

McMurdie PJ, Holmes S (2014) Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, **10**, e1003531.

Mohanty SR, Bodelier PLE, Conrad R (2007) Effect of temperature on composition of the methanotrophic community in rice field and forest soil.

FEMS Microbiology Ecology, **62**, 24–31.

Nikolenko SI, Korobeynikov AI, Alekseyev MA (2013) BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics*, **14**, S7.

Nurk S, Bankevich A, Antipov D, *et al*. (2013) Assembling genomes and mini-metagenomes from highly chimeric reads. In: *Lecture Notes in Computer Science*, pp. 158–170. Springer, Berlin Germany.

O’Gorman EJ, Benstead JP, Cross WF, *et al*. (2014) Climate change and geothermal ecosystems: Natural laboratories, sentinel systems, and future refugia. *Global Change Biology*, **20**, 3291–3299.

O'Gorman EJ, Ólafsson ÓP, Demars BO, *et al*. (2016) Temperature effects on fish production across a natural thermal gradient. *Global Change Biology*, **22**, 3206-3220.

O’Hara RB, Kotze DJ (2010) Do not log-transform count data. *Methods in Ecology and Evolution*, **1**, 118–122.

Oksanen J, Blanchet FG, Kindt R, *et al*. (2015) vegan: Community Ecology Package. *R package version 2.3-1*.

Peay KG, Bruns TD, Kennedy PG, Bergemann SE, Garbelotto M (2007) A strong species-area relationship for eukaryotic soil microbes: island size matters for ectomycorrhizal fungi. *Ecology Letters*, **10**, 470–480.

Peay KG, Garbelotto M, Bruns TD (2010) Evidence of dispersal limitation in soil microorganisms: isolation reduces species richness on mycorrhizal tree islands. *Ecology*, **91**, 3631–3640.

Plebani M, Fussmann KE, Hansen DM, O’Gorman EJ, Stewart RIA, Woodward G, Petchey OL (2015) Substratum-dependent responses of ciliate assemblages to temperature: a natural experiment in Icelandic streams. *Freshwater Biology*, **60**, 1561–1570.

Purcell D, Sompong U, Yim LC, Barraclough TG, Peerapornpisal Y, Pointing SB (2007) The effects of temperature, pH and sulphide on the community structure of hyperthermophilic streamers in hot springs of northern Thailand. *FEMS Microbiology Ecology*, **60**, 456–466.

R Development Core Team (2016) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, **1**, 409.

Raskin L, Stromley JM, Rittmann BE, Stahl DA (1994) Group-specific 16S rRNA hybridization probes to describe natural communities of methanogens. *Applied and Environmental Microbiology*, **60**, 1232–1240.

Reche I, Pulido-Villena E, Morales-Baquero R, Casamayor EO (2005) Does ecosystem size determine aquatic bacterial richness? *Ecology*, **86**, 1715–1722.

Rillig MC, Wright SF, Shaw MR, Field CB (2002) Artificial climate warming positively affects arbuscular mycorrhizae but decreases soil aggregate water stability in an annual grassland. *Oikos*, **97**, 52–58.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2409v1.

Rothschild LJ, Mancinelli RL (2001) Life in extreme environments. *Nature*, **409**, 1092–1101.

Ryberg WA, Fitzgerald LA (2016) Landscape composition, not connectivity, determines metacommunity structure across multiple scales. *Ecography*, **39**, 932-941.

Sarmento H, Montoya JM, Vázquez-Domínguez E, Vaqué D, Gasol JM (2010) Warming effects on marine microbial food web processes: how far can we go when it comes to predictions? *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **365**, 2137–2149.

Sharp CE, Brady AL, Sharp GH, Grasby SE, Stott MB, Dunfield PF (2014) Humboldt’s spa: microbial diversity is controlled by temperature in geothermal environments. *The ISME Journal*, **8**, 1166–1174.

Stahl DA, Amann R (1991) Development and Application of Nucleic Acid Probes. In: *Nucleic Acid Techniques in Bacterial Systematics*, pp. 205–248, New York, NY, USA.

Tansey MR, Brock TD (1972) The upper temperature limit for eukaryotic organisms. *Proceedings of the National Academy of Sciences USA*, **69**, 2426–2428.

Treseder KK, Marusenko Y, Romero-Olivares AL, Maltz MR (2016) Experimental warming alters potential function of the fungal community in boreal forest. *Global Change Biology*, **22**, 3395–3404.

Ulrich W, Almeida-Neto M (2012) On the meanings of nestedness: back to the basics. *Ecography*, **35**, 865-871.

Valladares F, Matesanz S, Guilhaumon F *et al*. (2014) The effects of phenotypic plasticity and local adaptation on forecasts of species range shifts under climate change. *Ecology Letters*, **17**, 1351–1364.

van der Gast CJ, Lilley AK, Ager D, Thompson IP (2005) Island size and bacterial diversity in an archipelago of engineering machines. *Environmental Microbiology*, **7**, 1220–1226.

Vannette RL, Leopold DR, Fukami T (2016) Forest area and connectivity influence root-associated fungal communities in a fragmented landscape. *Ecology*, **97**, 2374–2383.

Vyverman W, Verleyen E, Sabbe K, *et al*. (2007) Historical processes constrain patterns in global diatom diversity. *Ecology*, **88**, 1924–1931.

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–5267.

Wang Y, Naumann U, Wright ST, Warton DI (2012) mvabund - an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, **3**, 471–474.

Wang S, Hou W, Dong H, *et al*. (2013) Control of temperature on microbial community structure in hot springs of the Tibetan Plateau. *PloS ONE*, **8**, e62901.

Warton DI, Lyons M, Stoklosa J, Ives AR, Schielzeth H (2016) Three points to consider when choosing a LM or GLM test for count data. *Methods in Ecology and Evolution*, **7**, 882–890.

Werner JJ, Zhou D, Caporaso JG, Knight R, Angenent LT (2012) Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. *The ISME Journal*, **6**, 1273–1276.

Wickham H (2009) *ggplot2 Elegant Graphics for Data Analysis*, Springer, New York, NY, USA.

Woodward G, Dybkjær JB, Ólafsson JS, Gíslason GM, Hannesdóttir ER, Friberg N (2010) Sentinel systems on the razor’s edge: Effects of warming on Arctic geothermal stream ecosystems. *Global Change Biology*, **16**, 1979–1991.

Yim LC, Hongmei J, Aitchison JC, Pointing SB (2006) Highly diverse community structure in a remote central Tibetan geothermal spring does not display monotonic variation to thermal stress. *FEMS Microbiology Ecology*, **57**, 80–91.

Yvon-Durocher G, Allen AP, Bastviken D, *et al*. (2014) Methane fluxes show consistent temperature dependence across microbial to ecosystem scales. *Nature*, **507**, 488–491.

Yvon-Durocher G, Allen AP, Cellamare M, *et al*. (2015) Five Years of Experimental Warming Increases the Biodiversity and Productivity of Phytoplankton. *PLoS Biology*, **13**, e1002324.

Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, **30**, 614–620.

Zhou J, Deng Y, Shen L, *et al*. (2016) Temperature mediates continental-scale diversity of microbes in forest soils. *Nature Communications*, **7**, 12083.

Chapter 6

Does Climate Drive The Distributions Of Arbuscular Mycorrhizal Fungi?

Acknowledgements: Joseph D Chipperfield, Terry J McGenity, Graham JC Underwood, and Alex J Dumbrell

Abstract

Climate change is projected to alter the distributions of species as they must move to remain within their climatic niche. This is likely to cause considerable changes in the functionality of ecosystems, and an increased likelihood of extinction if the range of a species shrinks. However, whilst this has been demonstrated in numerous species, the potential for range shifts in microorganisms is unknown, as the link between climate and microbial distributions is rarely examined. Therefore, understanding the extent to which microbial distributions are determined by current climate conditions is important in determining the potential for microbial range shifts to occur under future climatic conditions. To this end, using Bayesian species distribution modeling, I model the distributions of over 300 arbuscular mycorrhizal (AM) fungi using climatic variables. I find that most fungal operational taxonomic units (OTUs) show statistically supported relationships with multiple climatic variables. However, the nature and identity of these climatic relationships vary considerably between OTU definitions, and between and within AM fungal taxa, especially at the genus level. Furthermore, I compared climatic models

to null models using Bayes Factor analysis. For the majority of OTUs, there was no support for climate as a driver of their distribution. Overall, the results highlight whilst most AM fungal OTUs do not have a strong climatic niche, those that do show climatic niche differentiation as their responses to climatic drivers were different. Whilst most OTUs show little support for climatically constrained distributions, the extent to which climate drives the distribution is dependent on the identity of the AM fungus itself, therefore hinting that studies should focus on population level, rather than community, responses to climate. Our findings suggest the potential for novel plant-AM fungal interactions, with unpredictable ecological consequences.

Introduction

The Earth is currently undergoing incredibly rapid climate change. Such large climatic changes are projected to have widespread and detrimental effects on global biodiversity (Bellard *et al.*, 2012) and, by extension, the functioning of ecosystems (Schröter *et al.*, 2005). One mechanism by which this may happen is through climate induced range shifts (Parmesan and Yohe, 2003). As the planet warms, species must move in order to track their thermal and climatic niche, or face extinction. For many species, climate change is projected to shift species' ranges into Polar regions, or higher altitudes (Hickling *et al.*, 2006; Parmesan *et al.*, 1999). This is likely to result in widespread range reductions in most species, and therefore heightened risk of extinction (Thomas *et al.*, 2004a). Understanding the extent to which species' distributions are linked to climate is therefore a prerequisite to predicting their future distributions, potential for extinction, and the functionality of novel ecosystems.

However, whilst much progress has been made in developing tools to quantify the statistical relationships between climate and species' distributions (Guisan and Thuiller, 2005), the range of taxa to which such tools have been applied is relatively narrow. In particular, conspicuous and charismatic taxa such as insects, birds, and trees are well studied (e.g. Thomas *et al*., 2004b). Yet, the potential for range shifts in below-ground biota has lead to the systematic exclusion of inconspicuous but functionally critical taxa, such as microorganisms (Fitter *et al.*, 2000). The vast majority of Earth's biodiversity and biomass belongs to microorganisms (Locey and Lennon, 2016). Their ubiquity, numbers, and functional capabilities mean they are key drivers of Earth's biogeochemical cycles with ecosystem wide impacts (Falkowski *et al.*, 2008). However, whilst the environmental factors determining microbial distributions at small scales are well studied, the extent to which climate determines their distributions is poorly understood. Critically, we therefore do not know the potential for climate change to alter the distribution of key groups of functionally important microorganisms, and by extension the potential for changes in the ecosystem processes that they control.

One such group of functionally diverse microorganisms are the arbuscular mycorrhizal (AM) fungi (phylum Glomeromycota). These fungi are obligate plant-root endosymbionts, that associated with at least 2/3 of terrestrial plant species, making them one of the most common symbioses in nature (Fitter *et al.*, 1997). By increasing the uptake of nutrients such as phosphorous from the soil in exchange for carbon provided by the host plant, AM fungi can be hugely beneficial to the host plant with profound impacts at the ecosystem level (van der Heijden *et al.*, 1998a, 1998b). However, the nature of plant-AM fungal interactions can vary considerably along a spectrum from highly beneficial to the host plant, through neutral, to parasitic (van der Heijden *et al.*, 1998a; Johnson *et al.*, 1997).

Despite their obvious importance to ecosystem functionality, our knowledge about the ecological drivers of AM fungal distributions is predominantly biased towards local-scale, physicochemical factors (e.g. Dumbrell *et al*., 2010). Much attention has also been given to the role of host-plant community composition and diversity in determining the range of AM fungal taxa (Johnson *et al*., 2004; Öpik *et al*., 2009), though recent evidence suggests that over global scales, biotic control (e.g. through host-specificity) over AM fungal distributions is minimal (Lekberg & Waller, 2016). However, comparatively little is known about whether larger scale environmental filters, such as climate, modulate AM fungal distributions and therefore, the potential for range shifts in AM fungi under future climatic conditions remains unknown (Fitter *et al.*, 2000). Range shifts in AM fungi introduce the possibility for novel AM fungal-host plant interactions that, given the highly context-dependent nature of the AM fungal symbiosis, will be difficult to predict the ecological consequences of. Therefore, understanding whether AM fungi are likely to experience range shifts, is key in predicting ecosystem functionality under novel climatic conditions

Experimental approaches have provided evidence that AM fungal communities and taxa may respond to various factors associated with climate change, including elevated CO_2, increased temperature, and decreased precipitation (Compant *et al.*, 2010). Yet, whilst experimental studies are useful in determining the local-scale ecological consequences of climate change, they are not able to determine whether AM fungal distributions themselves are linked to climatic factors and for this reason, are of limited use in informing us about the potential for climate driven range shifts in this functionally important group of microorganisms. Therefore, I sought to test whether climate is linked to the distributions of AM fungal taxa over global scales, using a species distribution modeling (SDM) approach (Pearson & Dawson, 2003). To do this, occurrence data were obtained from the largest available molecular database on AM fungi, and used to test which climatic factors (if any) were related to their distributions. It has been suggested that the taxonomic resolution at which AM fungi are studied can greatly alter our perception of their ecology (Lekberg *et al*., 2014). In particular, overly broad taxonomic resolutions may artificially expand AM fungal ranges by “lumping” distinct AM fungal ecotypes with differing climatic niches (Bruns and Taylor, 2016), obscuring any climatic signal present in the distribution. Thus, by increasing the taxonomic resolution, climatic signal (if present) should be enhanced. I therefore used AM fungal OTUs defined at multiple sequence similarity thresholds to examine whether increasing taxonomic resolution may reveal climatic signal in the distribution of AM fungi. Finally, I determined the extent to which climatic factors are able to predict AM fungal distributions, and therefore how likely AM fungi are to experience climatic range shifts.

Methods

Occurrence Data

AM fungal occurrence data were obtained from the MaarjAM database (Opik *et al.*, 2010). This database contains geo-referenced AM fungal 18S rRNA sequence records, along with various contextual metadata, including sample date and source. I downloaded the entire MaarjAM database (accessed 24/02/2017) and cleaned the data by removing any records with missing geographic coordinates, or genbank accession numbers. Geographical points from which records were obtained were then assigned to 10 arc-minute grid squares covering the terrestrial surface of the Earth. Grid square was then used as a sample identifier with which to label sequence data.

The MaarjAM database automatically assigns sequence records to “virtual taxa” (Öpik *et al.*, 2009), based on phylogenetic and sequence similarity criteria. However, it has been suggested that more specific taxonomic grouping may be required to study AM fungal ranges. Therefore, in order to test whether climatic drivers emerged at higher taxonomic resolution, sequences were clustered into operational taxonomic units (OTUs) at three OTU definitions (97, 98, and 99% sequence similarity). This was achieved using VSEARCH (Rognes *et al.*, 2016), with the virtual taxon “type sequences” as cluster seeds. The number of sequence records within each grid square was recorded, along with the abundance of each OTU, in each grid square.

Climate Data

Climatic predictors were downloaded the “Bioclimatic” variables at a 10 arc-minute resolution from the WorldClim 2 database (Fick and Hijmans, 2017). These data consist of climatic layers, averaged over the past 50 years, and are frequently used in species distribution modeling studies (Hijmans and Graham, 2006). Here, mean annual temperature (bioclim_1), temperature seasonality (variability across seasons, bioclim_4), mean annual precipitation (bioclim_12), and precipitation seasonality (bioclim_15) were used as climatic predictors. These variables are intuitive descriptors of the average climate condition and its seasonal variability, and were statistically representative of the other climatic variables (Fig. 6.1). Details of other, unused variables are provided below, along with their vector names as used in Fig. 6.1.

- bioclim_2 – Mean diurnal range
- bioclim_3 – Isothermality (diurnal range / annual range)
- bioclim_5 – Maximum temperature of the warmest month
- bioclim_6 – Minimum temperature of coldest month
- bioclim_7 – Temperature annual range
- bioclim_8 – Mean temperature of wettest quarter
- bioclim_9 – Mean temperature of driest quarter
- bioclim_10 – Mean temperature of warmest quarter
- bioclim_11 – Mean temperature of coldest quarter
- bioclim_13 – Precipitation of wettest month

- bioclim_14 – Precipitation of driest month
- bioclim_16 – Precipitation of wettest quarter
- bioclim_17 – Precipitation of driest quarter
- bioclim_18 – Precipitation of warmest quarter
- bioclim_19 – Precipitation of coldest quarter

Statistical Analyses

To fit climatic species distribution models to AM fungal occurrence data, binomial generalised linear models (GLMs) were used to model the probability of occurrence of each AM fungal OTU in relation to the climatic predictors. Binomial models are an appropriate GLM for modeling integer based proportions, where the proportion represents the number of “successes” as a result of a number of Bernoulli trials, in our case a success represents the presence of a given fungal OTU in a given grid square, whilst each sequence record represents a Bernoulli trial. Therefore, we are modeling the probability that a given sequence record in a given grid square represents a specific fungal OTU. Models were fit in a Bayesian framework using the “R-INLA” package (Rue *et al.*, 2009). For each model, the independent variable was the abundance of an OTU in each grid square, and the number of Bernoulli trials was set as the number of records for each grid square.

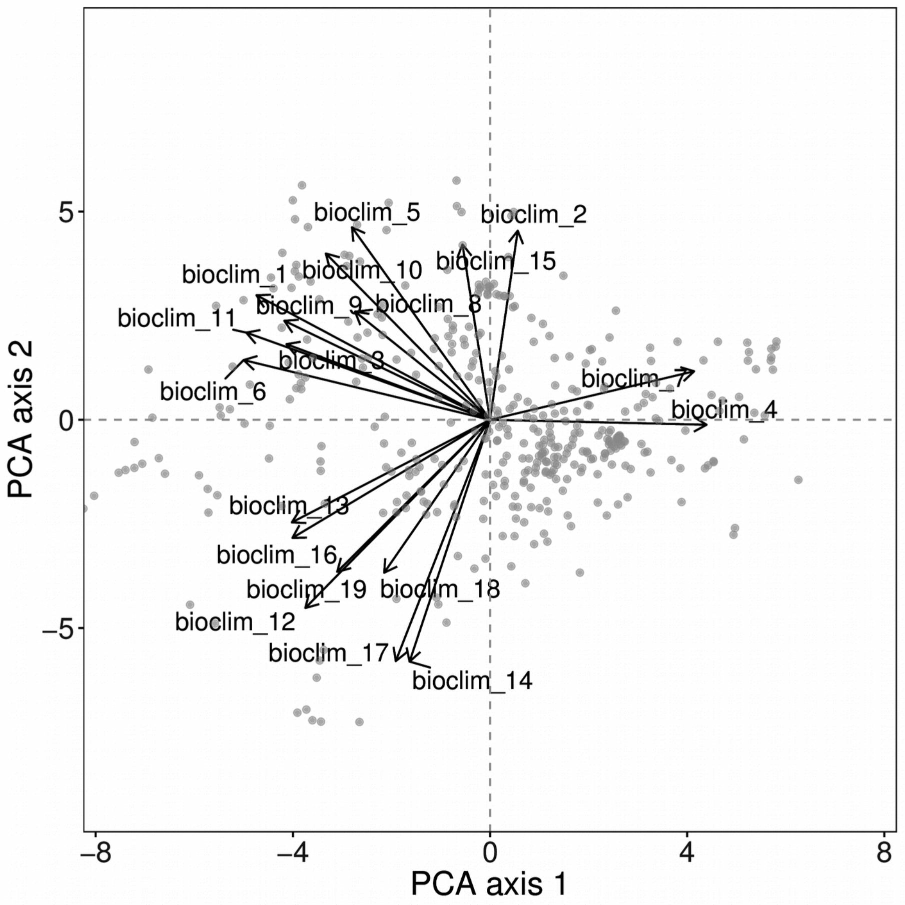


Figure 6.1 PCA analysis of Bioclimatic variables for observation points. I used the variables mean annual temperature (bioclim_1), temperature seasonality (bioclim_4), mean annual precipitation (bioclim_12), and precipitation seasonality (bioclim_15), as these variables are biologically meaningful, and represent the main axes of temperature and precipitation well.

Prior to modeling, climate variables were scaled and centered by subtracting the mean of each variable from each value, then dividing by the standard deviation of each variable. Quadratic terms for both annual precipitation and annual temperature were included in each model, as this allows a hump shaped response curve to be fit. This is a biologically plausible relationship for variables where the probability of occurrence is expected to peak at an “optimal” value. In contrast, seasonality variables were fit as linear terms only, as quadratic terms for these variables would not have an intuitive biological interpretation. Models were created for all AM fungal OTUs, at each OTU definition.

In order to test the importance of individual climatic predictors to each fungus’ distribution, I inspected the posterior coefficient estimates. The median posterior coefficient estimate quantifies the average relationship between the probability of presence, and the climatic covariate of interest, whilst the credible interval of a coefficient is a measure of the certainty in the coefficient. A variable is interpreted as having statistical support if the 95% credible interval does not contain 0 (i.e. is totally negative, or totally positive). Practically speaking, this can be interpreted as showing a 95% probability that there is a (positive or negative) relationship with the covariate, given the data. The relative importance of each covariate to an OTU was estimated by the magnitude of the coefficient, as climatic covariates were scaled prior to modeling. I also examined the shape of the fitted response curves, according to the quadratic coefficients. If these coefficients are negative, the response curve will follow a biologically plausible hump shape, indicating an optimal climatic condition. If the quadratic coefficients are positive, they response curve follows a biologically implausible u-shape, indicating increased probability of presence at extreme climatic values. If the credible interval for quadratic coefficients contained 0, the response curve was considered to be of “indeterminate” shape, as it could be unimodal, u-shaped, or flat.

To determine the relative goodness of fit of the models, comparisons were made between climatic models and null, intercept only models for each OTU. An intercept only model, assumes that the probability of presence is constant across sites. The relative support for climatic, or null models was quantified using Bayes Factors (BFs). Bayes Factors quantify the relative evidence for one model over another, given the data (Good and Hardin, 2009). Pragmatically, a Bayes Factor of 1 indicates no support for either model. A Bayes Factor of < 1, in this study, would indicate support for the null model, and as BF → 0, relative support for the null model increases. Whereas, a BF > 1, indicates support for the climatic model, and as BF → ∞, relative support for the climatic model increases.

Results

From an initial 24,872 records, data cleaning reduced the number of usable, unique sequence records to 22,849. The remaining records represented 402 distinct grid squares (Fig. 6.2A), for which a complete set of climate data was obtained. These records represent a relatively narrow temporal sampling duration, with 95% of the “cleaned” records being published to the MaarjAM database from 2008 onwards (Fig. 6.2B), indicating that differences in sample times between records should have minimal influence on the results.

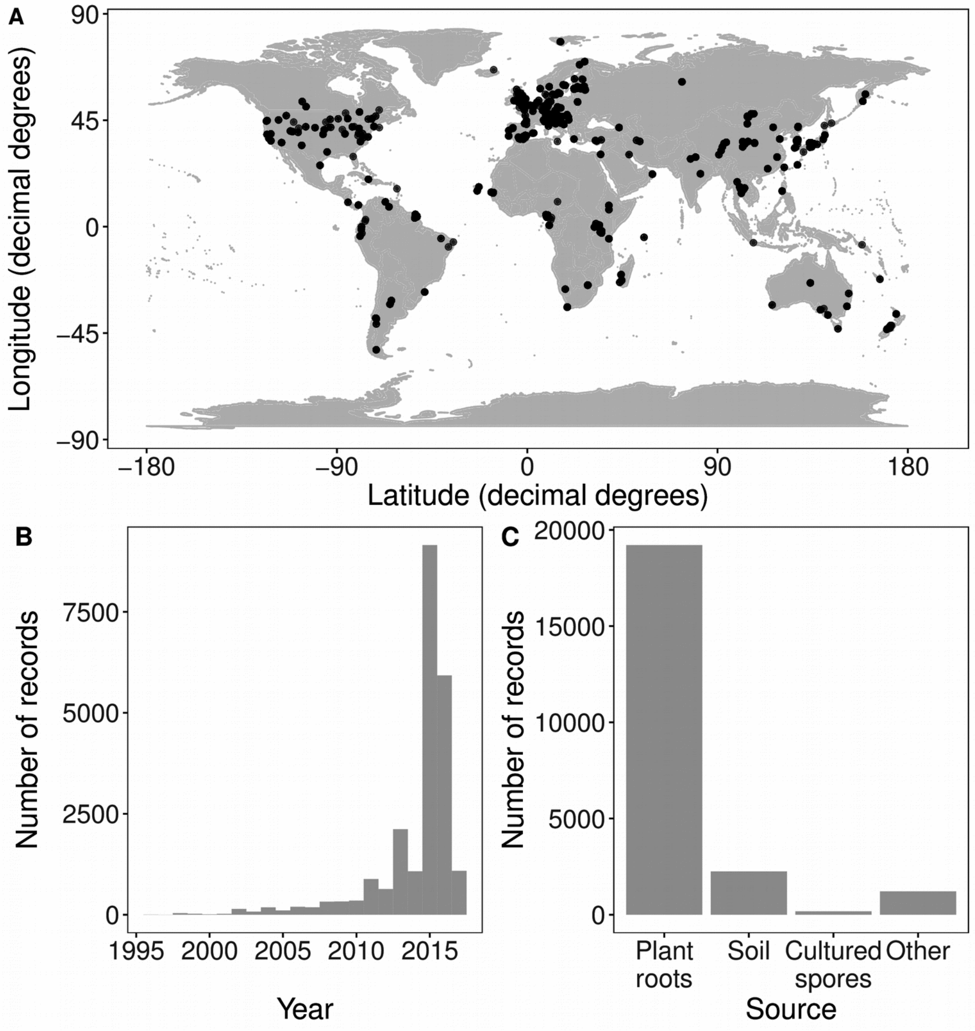


Figure 6.2 (A) A map of all locations from which sequence records were

obtained (n = 402). (B) A histogram of publication years for the sequence records used in this study and, (C) the environmental sources of records used within this study.

Furthermore, the vast majority (84%) of sequence records were obtained from plant root samples, as opposed to (< 1%) spore traps, or (9.8%) soil samples (Fig. 6.2C). Of the 352 virtual taxon type sequences obtained from MaarjAM, 334 were recorded at all three OTU definitions (97, 98, and 99% sequence similarity). However, within the entire set of OTUs (at all similarity thresholds, *n* = 1,018), only 372 had more than 10 presences, therefore only results for these OTUs are presented, as OTUs with fewer presences are unlikely to yield robust models or conclusions.

Determining Climatic Drivers

Of the 372 AM fungal OTUs considered, only 27 did not show statistically supported relationships with any of the climatic covariates considered. For those OTUs that did show statistically supported relationships with at least one climatic covariate, most were related to at least 3 covariates, though this varied between the different OTU definitions (Fig. 6.3). The linear temperature term was the most frequently well supported climatic covariate overall at all three OTU definitions (Table 6.1), whilst many OTUs were also related to precipitation (linear term).

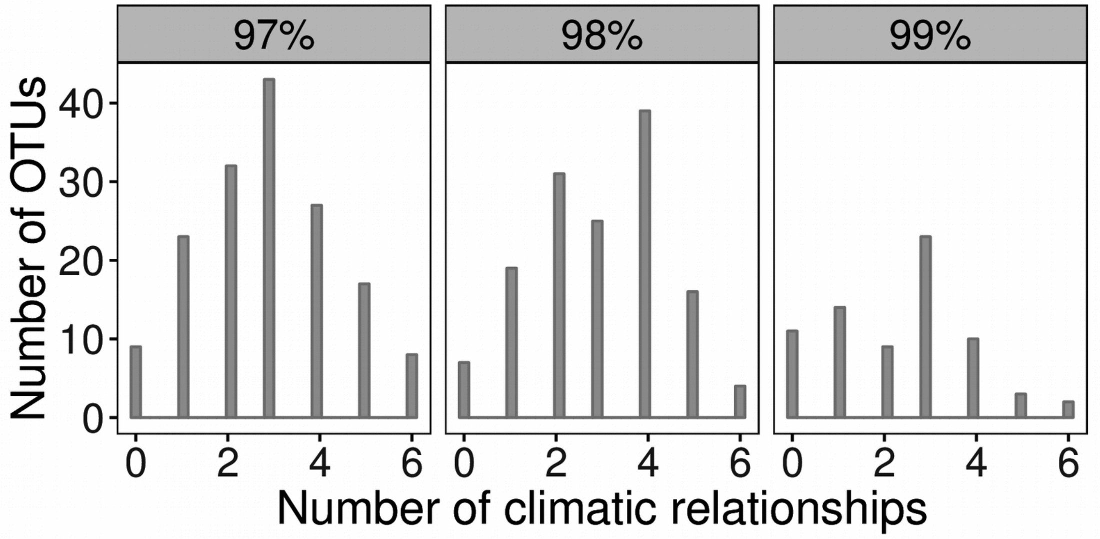


Figure 6.3 The number of climatic variables related to the distributions of AM fungal operational taxonomic units (OTUs), at each OTU definition. Climate variables were considered to show statistically supported relationships if the 95% credible interval did not bound 0. OTUs with a greater number of climatic relationships represent fungi that may have more complex relationships with climatic factors.

Whilst fewer OTUs were related to the quadratic terms for temperature and precipitation variables, the majority of those that were showed negative coefficients (Fig. 6.4). This means that the response curve for these variables forms a hump shaped relationship, as expected. The importance of seasonality (in terms of the number of OTUs related to it) in both temperature and precipitation varied between OTU definitions.

Table 6.1 The number of AM fungal operational taxonomic units (OTUs) with statistically supported relationships to each climatic covariate, at each similarity threshold. A climatic covariate was considered to show a statistically supported relationship if the 95% credible interval excluded 0.

OTU definition[a]	Climatic covariate	Number of OTUs with relationships to covariate
97%	Precipitation	93
	Precipitation (quadratic)	65
	Precipitation seasonality	62
	Temperature	103
	Temperature (quadratic)	62
	Temperature seasonality	72
98%	Precipitation	78
	Precipitation (quadratic)	62
	Precipitation seasonality	63
	Temperature	93
	Temperature (quadratic)	61
	Temperature seasonality	59
99%	Precipitation	37
	Precipitation (quadratic)	17
	Precipitation seasonality	25
	Temperature	37
	Temperature (quadratic)	27
	Temperature seasonality	25

[a] OTU definition refers to the sequence similarity at which OTUs were clustered, using the “type sequences” from the MaarjAM database as cluster seeds.

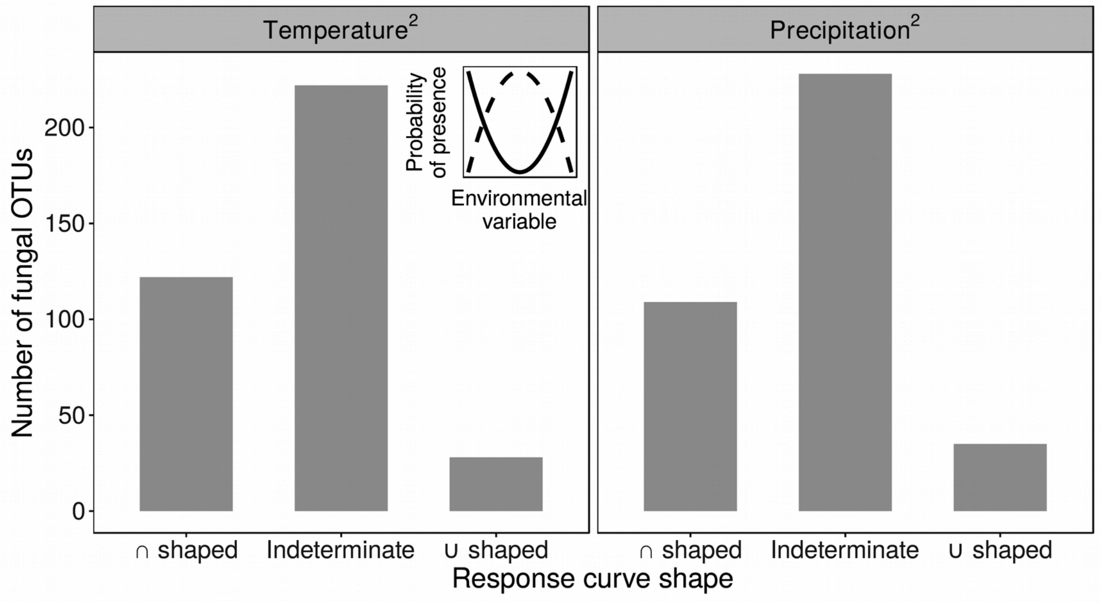


Figure 6.4 The number of OTUs showing different response curves to quadratic temperature and precipitation terms. Inset shows the two response curves for quadratic terms, if coefficients are negative (dashed line) the response curve is a biologically plausible unimodal shape, whereas if they are positive (solid line), a u-shaped curve is formed, possibly indicating a lack of data. The response curve was considered indeterminate if the 95% credible interval contained 0, as this indicates the coefficient could be positive or negative.

Taxonomy Dependent Climatic Relationships

OTUs representing finer taxonomic groups (99%) did not show stronger relationships with environmental covariates than broader OTU definitions (97%), as the magnitude of coefficients did not show any clear increase as finer OTU definitions were used (Fig. 6.5). This pattern was consistent when OTUs were divided into taxonomic groups.

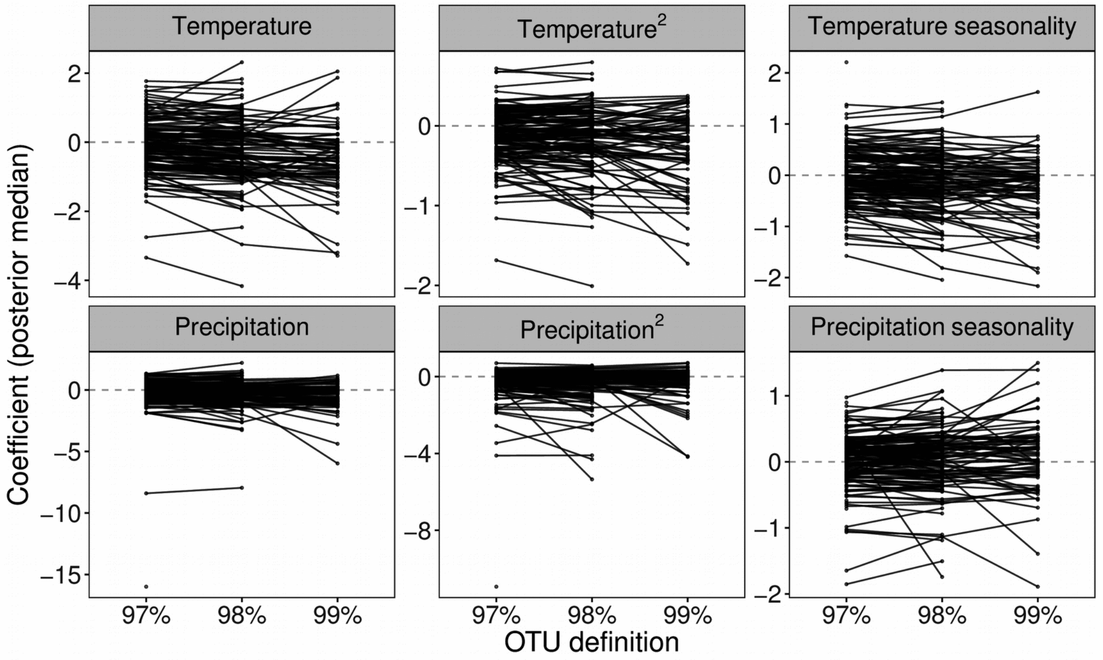


Figure 6.5 Climatic coefficients of operational taxonomic units (OTUs) at each OTU definition. Lines connect OTUs formed from the same sequence. At more specific OTU definitions, the magnitude of responses towards climatic factors did not get stronger, indicating that OTU definition had little effect on the detection of climatic relationships.

At the order level, there were no obvious differences in any of the climatic covariates between taxonomic groups (Fig. 6.6). Whilst at the genus level, different AM fungal genera showed different responses to climatic covariates. In particular, the genera *Paraglomus* and *Scutellospora* appeared distinct from other AM fungal genera in their response to climatic covariates, especially as they were showed predominantly positive responses to the linear temperature covariate, suggesting increased probability of presence in warmer climates (Fig. 6.7).

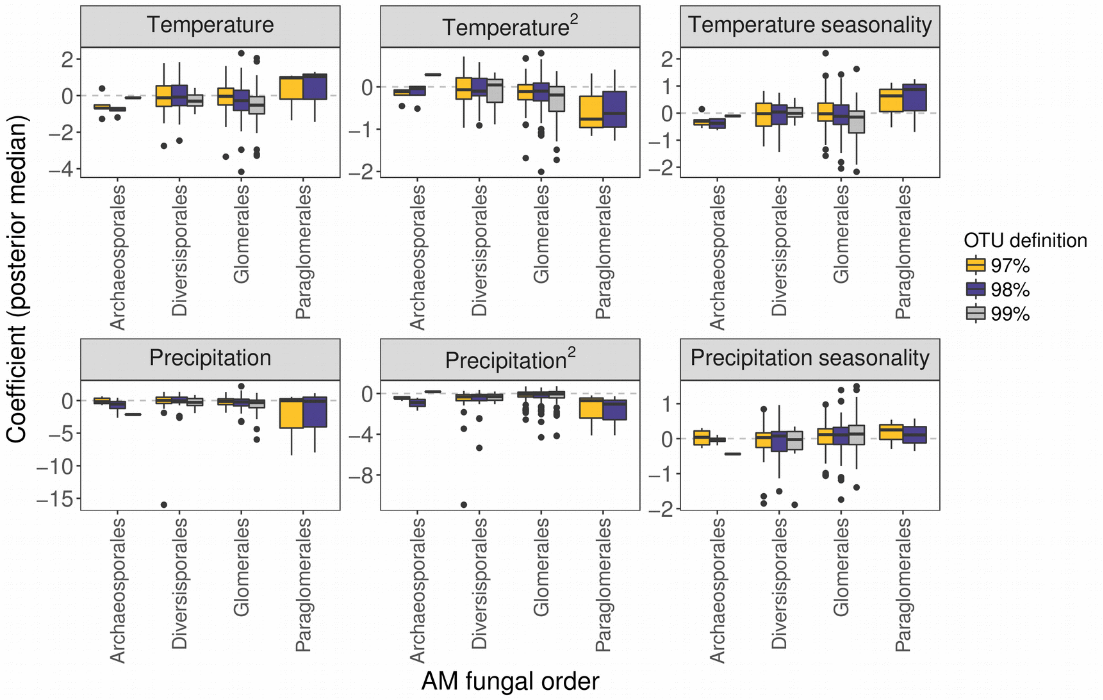


Figure 6.6 Climatic coefficients (median posterior estimates) of operational taxonomic units (OTUs) within different AM fungal orders, at each OTU definition. Only two boxes are present for Paraglomerales as OTUs within this order had too few presences at the 99% OTU definition to build robust models, and were excluded from further analyses.

The genus *Glomus* contained the most OTUs of all the AM fungal genera considered (n = 116 OTUs at the 97% OTU definition), and this genus also showed the greatest variability in responses to climatic covariates, suggesting that climatic niches may vary within, as well as between AM fungal genera.

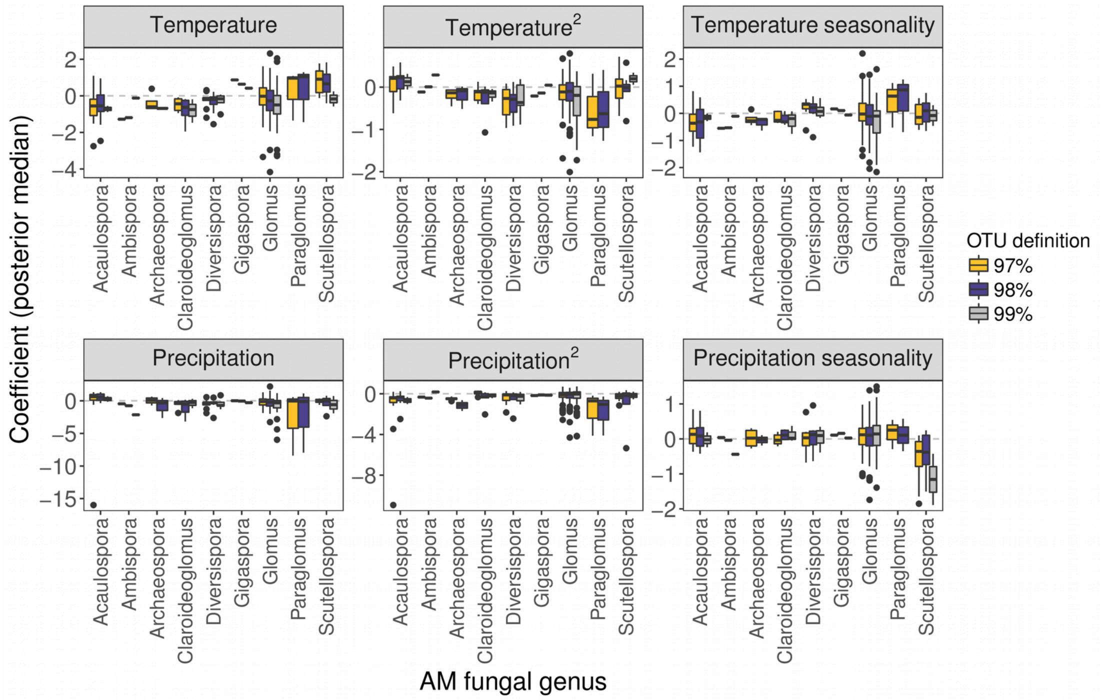


Figure 6.7 Climatic coefficients (median posterior estimates) of operational taxonomic units (OTUs) within different AM fungal genera, at each OTU definition.

Overall Influence of Climate

To test the predictive ability of the climatic SDMs, they were compared to null models, thus testing the relative evidence for climate driven distributions vs. random distributions. Analysis of Bayes factors suggested that, in the majority of OTUs, there was no evidence to suggest that climate based models were better than null models, as the majority of Bayes factors were < 1, and close to zero, supporting the null model over the climate model (Fig. 6.8). In contrast, 105 OTUs in total (45, 42, and 18 at 97, 98, and 99% OTU definitions respectively) had Bayes factors > 1, supporting climate models over null models, to varying extents. Thus, the extent to which climate modulates AM fungal distributions differs between OTUs.

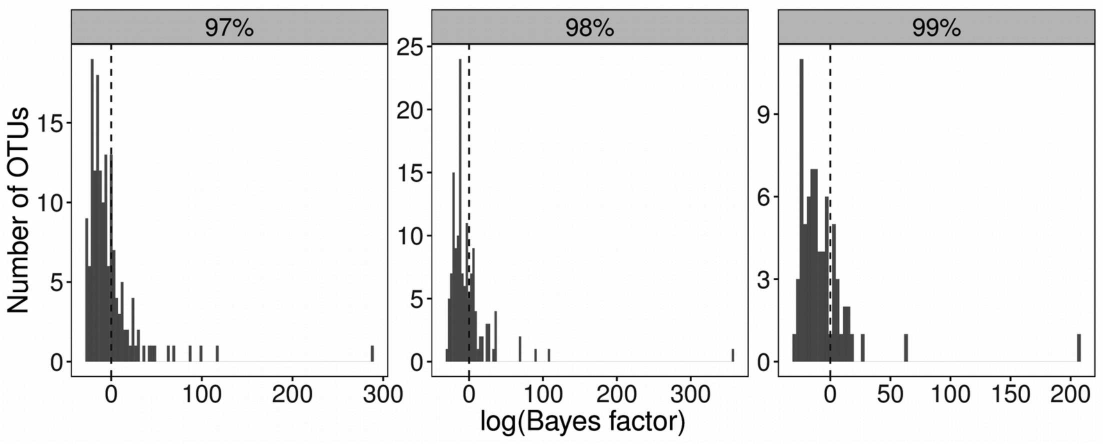


Figure 6.8 Histogram of Bayes Factors (BFs) for each operational taxonomic unit (OTU), at each OTU definition. Note that the natural log of BFs are presented for ease of visualisation. Bayes Factors to the left of the dashed line (0 = log(1)), indicate support of a null intercept only model, whereas BFs > 0 indicate support for climatic models. The further away from 0 a BF is, the stronger the support for either model.

Discussion

This study used a species distribution modeling approach to determine whether climate controls the distribution of arbuscular mycorrhizal fungi at global scales. The results suggest that many AM fungi show at least some relationship to climatic variables, although the nature and extent of these relationships vary between taxa, particularly at fine taxonomic resolutions (sub-genus level). However, for the vast majority of AM fungi, climate-driven distributions are not well supported and therefore, the importance of climate to AM fungi appears to be restricted to certain AM fungi.

Most AM Fungi Are Not Climatically Controlled

These analyses offer the first insight into whether the distributions of multiple AM fungi are linked to large-scale current climatic conditions. Strikingly, we found that the majority of AM fungal OTUs showed little, or no support for climate effects on their distributions. In fact, for many of the OTUs, null models (assuming random distributions) gained greater statistical support, further highlighting the weak predictive power of climatic variables on the distributions of many AM fungi.

Previous research into the role of climate on the ecology of AM fungi has tended to focus on community level effects (Torrecillas *et al.*, 2013), making it challenging to determine individual taxon responses to climate (Kivlin *et al.*, 2017), making it impossible to generalise to other AM fungal taxa.

Observational studies of AM fungi have reported changes in community composition (Kivlin *et al.*, 2011), diversity (Torrecillas *et al.*, 2013), and colonisation (Zhang *et al.*, 2016; Hu *et al.*, 2013), in relation to climatic factors such as temperature or precipitation. By manipulating specific climatic variables, experimental approaches have also demonstrated climatic influences on aspects of AM fungal ecology. Hawkes *et al*. (2011) found that under experimentally induced drought conditions, AM fungal communities were more diverse and abundant than under non-drought conditions, whilst root colonisation increased under elevated precipitation conditions. In addition, temperature manipulations have demonstrated positive effects on root colonisation (Rillig *et al.*, 2002), and spore size and density under elevated temperatures (Zhang *et al.*, 2016). These studies show that climatic factors can influence AM fungi from the population level (changes in root colonisation and hyphal length), through to community level changes (changes in total fungal biomass, and diversity). Coupled with the results gathered here, this suggests that many AM fungi will be affected by climate change, but these changes may manifest themselves at more local scales than the global range of the species.

The lack of a relationship between climate and the distribution of many AM fungi found in this study hints that climatically driven range shifts or changes in range size may be unlikely for most AM fungi. If AM fungi are highly host-specific, given the extensive evidence of climatic range shifts in host plant species (Kelly and Goulden, 2008; Lenoir *et al.*, 2008), then climate could modulate the distributions of AM fungi indirectly by manipulating the distributions of their host-plant species. Under this scenario, there would be no expected change in the functionality of the mycorrhizal symbiosis, as the identity of the plant and AM fungus will remain the same. However, global studies of AM fungal ranges show that there is little coupling between the ranges of specific AM fungi and host-plant species (Lekberg & Waller, 2016). In this case, there may be a disconnect between the future ranges of host-plants and AM fungi, as the plants may shift their ranges, whilst the AM fungi will not. Here, the host-plant species may acquire novel AM fungal symbionts, the composition of which may be determined by the local environment. Furthermore, the local AM fungal community might play an important role in the establishment of new host-plant species that may be experience range shifts. AM fungi have previously been shown to influence the success of alien plant species (Moora *et al.*, 2011; Menzel *et al.*, 2017), and it is therefore likely that AM fungi could have significant effects on future plant communities through below-ground facilitation of host plants in novel climates. In this case, the disconnect in the future ranges of host-plant species and AM fungi will likely result in novel plant-microbe interactions, that in turn will lead to unpredictable ecological functionality.

Climatic Drivers and Influence Vary Between Taxa

Whilst many of the fungal OTUs showed little evidence of climatic distributions, some did, suggesting that the effects of climate on AM fungi are likely to vary between and within AM fungal taxonomic groups. I found considerable variation in the nature of climatic relationships between different AM fungal genera, and even within certain genera, such as *Glomus* sp., in which OTUs showed highly variable responses to climatic variables. This finding is consistent with previous research that showed that colonisation responses induced by drought conditions were differentiated between two species of *Glomus* (Davies *et al.*, 2002). Furthermore, Klironomos *et al*. (1998) showed that under elevated CO_2 conditions, the colonisation and sporulation of four AM fungal species was variably affected. More broadly speaking, niche differentiation with respect to physicochemical parameters has been shown to structure AM fungal communities, particularly along strong physicochemical gradients (Dumbrell *et al.*, 2010), and in wider fungal communities (Geml *et al.*, 2012).

These results suggest that some AM fungi are likely to show climatic niche differentiation. This finding means that the ecological consequences of climate change on mycorrhizal mediated ecosystem processes are likely to be unpredictable, as the extent to which mycorrhizal communities will change in a given environment is dependent on the individual AM fungal taxa present. Thus, further research should therefore consider population level effects in parallel with community level effects in order to gain a more holistic view of AM fungal ecology under future climatic conditions.

Extensions and Applications of Microbial Species Distribution Modeling

Whilst the statistical approaches used here are robust, it would be remiss not to discuss the limitations and potential future directions of climatic niche modeling of microorganisms, and especially, AM fungi. A common caveat of SDM approaches is the assumption of spatial independence, that is to say, that each observation has no influence on other observations (Dormann, 2007). Biologically speaking, this assumption is rarely validated as many species are dispersal limited, meaning that observations close to a known presence are more likely to also be presences than observations from geographically distant sites (Lennon, 2000). For AM fungi, the validity of this assumption is unclear as evidence for whether AM fungi are dispersal limited or not, and at which spatial scales, is mixed (Lekberg *et al.*, 2007; Dumbrell *et al.*, 2010; Davison *et al.*, 2015; Kivlin and Hawkes, 2016). Additionally, from community based studies, it is not possible to tell the dispersal status of individual taxa. Therefore, an obvious extension to this study would be to incorporate the effects of spatial autocorrelation within the data. Bayesian approaches such as those implemented within the INLA software would probably offer the most computationally tractable method to do this, particularly as INLA is becoming more widely using in the SDM community (e.g. Blangiardo *et al.*, 2013). In addition to improving the fit of SDMs, determining the importance of dispersal related processes to AM fungi would offer further insight into the nature of range shifts under climate change. If AM fungi are unable to disperse efficiently over large geographic distances, they may not be able to keep up with the rapid rates of climatic change projected to occur over the coming decades (Engler *et al.*, 2009). In this scenario, some AM fungi may experience range contractions, and potentially increased risk of local extinction. Better characterisation of the dispersal abilities of individual microbial taxa would therefore improve our understanding about the potential ecological consequences of climate change on microorganisms. Such knowledge would require the integration of macroecological approaches with trait based modeling, informed by physiological and experimental studies of AM fungal taxa.

Another potential direction for future work on AM fungal distribution modeling would be to incorporate biotic interactions. Evidence for host specificity in AM fungi is mixed, and may be context-dependent (Douhan *et al.*, 2005; Helgason *et al.*, 2007; Torrecillas *et al.*, 2012). However, jointly considering the climatic niches of AM fungi and host-plant species may help to pinpoint areas where the climatic niches of AM fungi and their host plants do, or do not overlap. In turn, this would help identify areas of potential conservation interest (where plant-fungus interactions are conserved), or research interest, in areas where novel plant-fungus interactions may be likely to occur. Recent developments in the field of statistical modeling now mean that joint species distribution models (jSDMS) are computationally tractable. This class of models allow simultaneous modeling of the distributions of multiple species and importantly, it is possible to partition the effects of climatic covariates from biotic interactions (Pollock *et al.*, 2014; Ovaskainen *et al.*, 2015). However, such models have remained largely under-utilised within the microbial ecology literature to date, despite their obvious usefulness (Björk *et al.*, 2017).

Finally, consideration of the climatic niches of the AM fungi present in a habitat as “dormant” spores may help to build a predictive framework of the types of plant-fungal interactions under future climate conditions. AM fungal community shifts resulting from activation of sporulated fungi are likely to occur more rapidly than as a result of geographical range shifts. Therefore, characterising the climatic niches of dormant fungi would allow prediction of which members of the AM fungal spore bank may be more likely to become active in the future. By incorporating knowledge on the types of interactions these fungi have with their plant host (e.g. beneficial, neutral, parasitic; Chaudhary *et al*., 2016), it may even be possible to predict the potential for changes in ecosystem functionality from above- and below-ground interactions such as the AM fungal-plant symbiosis.

Conclusions

In summary, climatic factors do not drive the distributions of most AM fungi, suggesting that local-scale factors may be of greater importance. However, some AM fungi do appear to show climatic niches, indicating that the potential for range shifts to occur under future climatic conditions is taxon-specific, highlighting the need for microbial ecologists to consider individual taxa, as well as communities in future research. Furthermore, the identity of climatic variables driving some AM fungal distributions vary between taxa, adding to a growing awareness that microorganisms may show climatic niche differentiation. The results highlight the applicability of species distribution modeling in understanding the potential impacts of climate change on functionally important microorganisms.

References

Bellard C, Bertelsmeier C, Leadley P, Thuiller W, Courchamp F (2012) Impacts of climate change on the future of biodiversity. *Ecology Letters,* **15**, 365–377.

Björk JR, Hui FKC, O'Hara RB, Montoya JM (2017) Uncovering the drivers of animal-host microbiotas with joint distribution modeling. *BioRxiv,* 137943.

Blangiardo M, Cameletti M, Baio G, Rue H (2013) Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology,* **7**, 39–55.

Bruns TD, Taylor JW (2016) Comment on 'Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism'. *Science,* **351**, 826.

Chaudhary VB, Rúa MA, Antoninka A, *et al.* (2016) MycoDB, a global database of plant response to mycorrhizal fungi. *Scientific Data,* **3**, 160028.

Compant S, van der Heijden MGA, Sessitsch A, *et al.* (2010) Climate change effects on beneficial plant-microorganism interactions. *FEMS Microbiology Ecology,* **73**, 197–214.

Davies FT, Olalde-Portugal V, Aguilera-Gomez L, Alvarado MJ, Ferrera-Cerrato RC, Boutton TW (2002) Alleviation of drought stress of Chile ancho pepper (Capsicum annuum L. cv. San Luis) with arbuscular mycorrhiza indigenous to Mexico. *Scientia Horticulturae,* **92**, 347–359.

Davison J, Ainsaar L, Burla S, *et al.* (2015) Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. *Science,* **127**, 970–973.

Dormann CF (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography,* **16**, 129–138.

Douhan GW, Petersen C, Bledsoe CS, Rizzo DM (2005) Contrasting root associated fungi of three common oak-woodland plant species based on molecular identification: Host specificity or non-specific amplification? *Mycorrhiza,* **15**, 365–372.

Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH (2010) Relative roles of niche and neutral processes in structuring a soil microbial community. *The ISME Journal,* **4**, 337–345.

Engler R, Randin CF, Vittoz P, *et al.* (2009) Predicting future distributions of mountain plants under climate change: Does dispersal capacity matter? *Ecography,* **32**, 34–45.

Eom A-H, Hartnett DC, Wilson GWT (2000) Host plant species effects on arbuscular mycorrhizal fungal communities in tallgrass prairie. *Oecologia,* **122**, 435–444.

Falkowski PG, Fenchel T, Delong EF (2008) The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science,* **320**, 1034–1039.

Fick SE, Hijmans RJ (2017) WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology, e-pub ahead of print, doi: 10.1002/joc.5086.

Fitter AH, Heinemeyer A, Staddon PL (2000) The impact of elevated CO2 and global climate change on arbuscular mycorrhizas: A mycocentric approach. *New Phytologist,* **147**, 179–187.

Fitter AH, Moyersoen B, Silvertown J, Franco M, Harper JL (1997) Evolutionary trends in root-microbe symbioses. P*hilosophical Transactions: Biological Sciences*, **351**, 1367-1375.

Geml J, Timling I, Robinson CH, *et al.* (2012) An arctic community of symbiotic fungi assembled by long-distance dispersers: phylogenetic diversity of ectomycorrhizal basidiomycetes in Svalbard based on soil and sporocarp DNA. *Journal of Biogeography,* **39**, 74–88.

Good PI, Hardin JW (2009) Common Errors in Statistics (and How to Avoid Them). John Wiley & Sons, Hoboken, NJ, USA.

Guisan A, Thuiller W (2005) Predicting species distribution: Offering more than simple habitat models. *Ecology Letters,* **8**, 993–1009.

Hawkes CV, Kivlin SN, Rocca JD, Huguet V, Thomsen MA, Suttle KB (2011) Fungal community responses to precipitation. *Global Change Biology,* **17**, 1637–1645.

Helgason T, Merryweather JW, Young JPW, Fitter AH (2007) Specificity and resilience in the arbuscular mycorrhizal fungi of a natural woodland community. *Journal of Ecology,* **95**, 623–630.

Hickling R, Roy DB, Hill JK, Fox R, Thomas CD (2006) The distributions of a wide range of taxonomic groups are expanding polewards. *Global*

Change Biology, **12**, 450–455.

Hijmans RJ, Graham CH (2006) The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology,* **12**, 2272–2281.

Hu Y, Rillig MC, Xiang D, Hao Z, Chen B (2013) Changes of AM Fungal Abundance along Environmental Gradients in the Arid and Semi-Arid Grasslands of Northern China. *PLoS ONE,* **8**, e57593.

Johnson D, Vandenkoornhuyse PJ, Leake JR, *et al.* (2004) Plant communities affect arbuscular mycorrhizal fungal diversity and community composition in grassland microcosms. *New Phytologist,* **161**, 503–515.

Johnson NC, Graham JH, Smith FA (1997) Functioning of mycorrhizal associations along the mutualism-parasitism continuum. *New Phytologist,* **135**, 575–586.

Kelly AE, Goulden ML (2008) Rapid shifts in plant distribution with recent climate change. *Proceedings of the National Academy of Sciences USA,* **105**, 11823–11826.

Kivlin SN, Hawkes CV (2016) Tree species, spatial heterogeneity, and seasonality drive soil fungal abundance, richness, and composition in Neotropical rainforests. *Environmental Microbiology,* **18**, 4662–4673.

Kivlin SN, Hawkes CV, Treseder KK (2011) Global diversity and distribution of arbuscular mycorrhizal fungi. *Soil Biology and Biochemistry,* **43**, 2294–2303.

Kivlin SN, Muscarella R, Hawkes CV, Treseder KK (2017) The Predictive Power of Ecological Niche Modeling for Global Arbuscular Mycorrhizal Fungal Biogeography. In: *Biogeography of Mycorrhizal Symbiosis* (ed Tedersoo L), pp 143–158. Springer, Cham, Switzerland.

Klironomos JN, Ursic M, Rillig M, Allen MF (1998) Interspecific differences in the response of arbuscular mycorrhizal fungi to *Artemisia tridentata* grown under elevated atmospheric CO2. *New Phytologist,* **138**, 599–605.

Lekberg Y, Gibbons SM, Rosendahl S (2014) Will different OTU delineation methods change interpretation of arbuscular mycorrhizal fungal community patterns? *New Phytologist*, **202**, 1101-1104.

Lekberg Y, Koide RT, Rohr JR, Aldrich-Wolfe L, Morton J. (2007) Role of niche restrictions and dispersal in the composition of arbuscular mycorrhizal fungal communities. *Journal of Ecology,* **95**, 95–105.

Lekberg Y, Waller LP (2016) What drives differences in arbuscular mycorrhizal fungal communities among plant species? *Fungal Ecology*, **24**, 135-138.

Lennon JJ (2000) Red-shifts and red herrings in geographical ecology. *Ecography,* **23**, 101–113.

Lenoir J, Gegout JC, Marquet PA, de Ruffray P, Brisse H (2008) A Significant Upward Shift in Plant Species Optimum Elevation During the 20th Century. *Science,* **320**, 1768–1771.

Locey KJ, Lennon JT (2016) Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences USA,* **113***,* 5970-5975.

Menzel A, Hempel S, Klotz S, *et al.* (2017) Mycorrhizal status helps explain invasion success of alien plant species. *Ecology,* **98**, 92–102.

Moora M, Berger S, Davison J, *et al.* (2011) Alien plants associate with widespread generalist arbuscular mycorrhizal fungal taxa: evidence from a continental-scale study using massively parallel 454 sequencing. *Journal of Biogeography,* **38**, 1305–1317.

Öpik M, Metsis M, Daniell TJ, Zobel M, Moora M (2009) Large-scale parallel 454 sequencing reveals host ecological group specificity of arbuscular mycorrhizal fungi in a boreonemoral forest. *New Phytologist,* **184**, 424–437.

Opik M, Vanatoa A, Vanatoa E, *et al.* (2010) The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytologist,* **188**, 223–241.

Ovaskainen O, Roy DB, Fox R, Anderson BJ (2016) Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution*. **7**, 428-436.

Parmesan C, Ryrholm N, Stefanescu C, *et al.* (1999) Poleward shifts in geographical ranges of butterfly species associated with regional warming. *Nature,* **399**, 579–583.

Parmesan C, Yohe G (2003) A globally coherent fingerprint of climate change impacts across natural systems. *Nature,* **421**, 37–42.

Pearson RG, Dawson TP (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography*, **12**, 361-371.

Pollock LJ, Tingley R, Morris WK, *et al.* (2014) Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution,* **5**, 397–406.

Rillig MC, Wright SF, Shaw MR, Field CB (2002) Artificial climate warming positively affects arbuscular mycorrhizae but decreases soil aggregate water stability in an annual grassland. *Oikos,* **97**, 52–58.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ,* **4**, e2409v1.

Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (statistical methodology),* **71**, 319–392.

Schröter D, Cramer W, Leemans R, *et al.* (2005) Ecosystem service supply and vulnerability to global change in Europe. *Science,* **310**, 1333–1337.

Thomas CD, Cameron A, Green RE, *et al.* (2004a) Extinction risk from climate change. *Nature,* **427**, 145–148.

Thomas JA, Telfer MG, Roy DB, *et al.* (2004b) Comparative Losses of British Butterflies, Birds, and Plants and the Global Extinction Crisis. *Science,* **303**, 1879–1881.

Torrecillas E, Alguacil MM, Roldán A (2012) Host preferences of arbuscular mycorrhizal fungi colonizing annual herbaceous plant species in semiarid mediterranean prairies. *Applied and Environmental Microbiology,* **78**, 6180–6186.

Torrecillas E, Torres P, Alguacil MM, *et al.* (2013) Influence of habitat and climate variables on arbuscular mycorrhizal fungus community distribution, as revealed by a case study of facultative plant epiphytism under semiarid conditions. *Applied and Environmental Microbiology,* **79**, 7203–7209.

van der Heijden MGA, Boller T, Wiemken A, Sanders IR (1998a) Different arbuscular mycorrhizal fungal species are potential determinants of plant community. *Ecology,* **79**, 2082–2091.

van der Heijden MGA, Klironomos JN, Ursic M, *et al.* (1998b) Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature,* **396**, 69–72.

Zhang X, Johnston ER, Li L, Konstantinidis KT, Han X (2016) Experimental warming reveals positive feedbacks to climate change in the Eurasian Steppe. *The ISME Journal,* **11**, 885–895.

Chapter 7

General Discussion

Summary of Thesis Findings

Chapter 2 Summary

In Chapter 2, I compared whether metagenomic, or amplicon sequencing recovers the most diversity from microbial communities, and whether the additional cost of metagenomic sequencing is a worthwhile compromise. I assembled a dataset of published sequence data in which samples had been sequenced by both metagenomic, and amplicon sequencing. Results showed that, once differences in sequence numbers had been controlled for, metagenomic sequencing recovered a greater number of operational taxonomic units (OTUs). Furthermore, metagenomic sequencing recovered more taxonomic diversity than amplicon sequencing, even at basal taxonomic levels (from genus, to phylum). However, in all studies (except one), the cost of metagenomic sequencing to generate the same level of coverage provided by amplicon sequencing was at least an order of magnitude higher. I conclude that whilst metagenomic sequencing offers increased ability to recover microbial diversity, currently, the cost of this approach is still prohibitive in many cases (Neufeld, 2017).

Chapter 3 Summary

In Chapter 3, I tested the generality of macroecological relationships in microbial communities, and whether such relationships vary due to ecological, or methodological reasons. To do this, I conducted a meta-analysis of microbial distance-decay (d-d) relationships. For each distance-decay relationship, I recorded factors relating to methodological approach, including the sequencing depth, sampling effort, and molecular approach, and factors describing the biological context of each relationship, such as the scale, focal taxa, and study system. Both methodological and biological contextual aspects significantly influence the strength of microbial distance-decay relationships. Factors relating to community coverage tended to weakly, but significantly affect the strength of d-d relationships, whereas choice of dissimilarity index had a stronger effect. In particular, phylogenetic distance metrics resulted in weaker d-d relationships. Additionally, these results also showed that biological factors such as scale and study system resulted in different d-d relationships. These results highlight that methodological choices are capable of biasing our perception of macroecological patterns in microbial communities. Furthermore, the results from Chapter 3 demonstrate that macroecological relationships vary between biological contexts, and are therefore not universal within microbial communities.

Chapter 4 Summary

In Chapter 4, I examine the role of spatial processes in structuring microbial communities over macroecological scales, and whether such processes result in macroecological patterns comparable to higher organisms. I characterised the extremophilic archaeal communities from halite crystals, gathered over continental scales. I found that over small spatial scales (< 300 km), species turnover in these communities was strongly related to geographic distance, indicating limited dispersal between sites. However, at larger scales, turnover was not related to distance, and communities did not cluster together in a spatially coherent manner. Finally, I tested for archaeal genera indicative of specific biogeographic regions, and was able to identify several genera that were good indicators of geographic origin. Overall, results from Chapter 4 indicate that extremophilic microbial communities may be organised by spatial factors operating at small, rather than regional, spatial scales. Furthermore, the influence of spatial processes is likely to be different between microbial taxa. Collectively, these results indicate the need to consider ecological patterns at the population level, as well as at the community level.

Chapter 5 Summary

Within Chapter 5, I examine the relationship between environmental gradients and microbial communities, and whether the relationship generalises across microbial taxa and geographic distance. I comprehensively characterised the sediment microbial communities (Archaea, Bacteria, and Eukarya) from five parallel geothermally warmed stream systems situated around the Arctic circle. The results showed that in three sites, microbial communities followed the expected unimodal, hump shaped relationship between temperature and diversity, whereas in two sites, the relationship is flat. Multivariate models revealed that the predicted thermal optima of microbial taxa was different between Eukarya (lowest thermal optima), Bacteria, and Archaea (highest thermal optima). Finally, by partitioning β-diversity into nestedness and species turnover components, the results showed that microbial communities change along environmental gradients predominantly due to species replacement (turnover), rather than species loss (nestedness). Results from Chapter 5 highlight that regional-scale metacommunity dynamics may influence the extent to which the environment structures microbial communities (Telford *et al.*, 2006), as the two smallest and most isolated sites showed very different relationships to the other three. Furthermore, by showing that microbial taxa have different thermal optima, and that species turnover occurs over temperature gradients, the results comprehensively demonstrate that niche differentiation plays a major role in structuring microbial communities over strong environmental gradients. The results of Chapter 5 have important implications in understanding how microbial communities may change under global warming over macroecological scales.

Chapter 6 Summary

In Chapter 6, I investigated whether the distributions of microorganisms, specifically the arbuscular mycorrhizal (AM) fungi, are driven by modern climatic conditions. If they are, microbes may experience range shifts under future climate change scenarios, which in the case of AM fungi, may result in novel plant-fungus interactions, with unknown consequences for plant productivity. I obtained distribution data from the AM fungal database, Maarjam (Opik *et al.*, 2010), and used species distribution models to investigate whether their distributions are linked to climate. Results showed that different AM fungi respond to different climatic drivers, and that this climatic niche differentiation emerged at the genus level, rather than more broad taxonomic levels. However, the results also showed that for most AM fungi, at most taxonomic resolutions, climate was a relatively poor predictor of their distribution, and in most cases climatic species distribution models were not statistically supported over random null models. These results suggest that climatically driven geographic range shifts are unlikely in many AM fungi. They also suggest that climatic niche differentiation between different AM fungi means that microbial communities should be considered at the population level when modeling the potential effects of climate change. Additionally, the results hint at the possibility of novel plant-fungus-environment interactions, which are likely to have highly unpredictable ecological consequences at the ecosystem level.

Sampling Microbial Communities for Macroecological Studies

Sequencing Microbial Communities

Throughout this thesis, I have utilised high-throughput sequencing data to characterise the composition and diversity of microbial communities. The ability of modern molecular methods to delve into previously unobserved parts of the microbial biosphere facilitates tests of macroecological hypotheses in microbial communities (Barberán *et al.*, 2014). In Chapter 2, I show that metagenomic sequencing allows superior sampling of microbial diversity when compared to amplicon sequencing. However, the requirement for much greater sequencing depth in metagenomic sequence datasets means that the financial cost of such an approach could be an order of magnitude higher. Coupled with the fact that macroecological datasets often require high sampling effort in order to gain the necessary statistical power, the trade-off between sequencing depth and sampling effort represents an unacceptable compromise. In comparison, amplicon sequencing offers a more favourable cost:sequencing depth ratio, and has the benefit of being able to target specific groups of organisms. Therefore the most optimal method of sampling microbial biodiversity for testing macroecological hypotheses is currently still amplicon sequencing, although decreasing costs will eventually make metagenomics a more favourable option (Neufeld, 2017).

Within Chapters 4 and 5, I used extremely high-throughput amplicon sequencing using the Illumina HiSeq platform. In Chapter 4, the extremely high coverage offered by this approach near comprehensive coverage of the haloarchaea present within halite crystals. In the context of the chapter, this allowed quantification of even the very rare organisms, that are more likely to be endemic to specific regions (Liu *et al.*, 2015). A lower coverage approach may have missed these taxa, and therefore communities might appear artificially similar to each other. In turn, this may have affected the clustering of communities and the ecological conclusion that extremophilic microbial communities do not form biogeographic regions.

In Chapter 5, amplicon sequencing was used to specifically target different microbial taxonomic groups present in Arctic stream sediments. Here, the ability to target specific taxonomic groups of organisms was invaluable. Within this chapter, it was found that microbial taxonomic groups differ in their temperature-richness relationships. By targeting specific groups of organisms using amplicon sequencing, I was able to ensure that sequencing was of sufficient depth to provide adequate coverage for each taxonomic group (Archaea, Bacteria, and Eukarya) in order to robustly model the temperature-diversity relationship of each taxonomic group. In this instance, if lower sequencing depth approaches were used, the most diverse samples would have been inadequately sampled, inevitably leading to flatter temperature-diversity relationships. Chapter 5 therefore highlights the benefit of being able to allocate sequencing coverage to specific taxonomic groups in order to ensure that they are adequately sampled before making ecological inferences.

Databases and Open Data in Microbial Macroecology

Whilst the generation of new empirical datasets is often necessary to test (macro)ecological hypotheses, the culture of data sharing in microbial ecology allows ecological or methodological hypotheses to be tested on previously published datasets. In Chapter 2, I utilised previously published sequence datasets that contained both metagenomic, and amplicon sequence data, in order to test which sequencing method recovered the most diversity. Sequence data were obtained from a number of sources including the European nucleotide archive (Leinonen *et al.*, 2011a), the sequence read archive (Leinonen *et al.*, 2011b), and the MG-Rast server (Meyer *et al.*, 2008). The resulting dataset contained more than 1.1 billion sequences in total, spanning multiple sequencing platforms, sequencing depths, and biomes. The results showed that metagenomic sequencing recovered more diversity than amplicon sequencing, and that this conclusion held across sequencing platforms, biomes, and different sequencing depths. However, the difference in cost between the two sequencing approaches was dependent on the platform that had been used for each sequencing approach. In this context, the ability to combine datasets from a wide variety of biomes and sequencing technologies allowed far greater insight and more generalisable conclusions than if all sequence data had been generated in one sequencing run. The findings of this chapter therefore built on previous comparisons of metagenomic and amplicon sequencing that had generated data on a restricted range of sequencing platforms or environmental contexts (Poretsky *et al.*, 2014; Tessler *et al.*, 2017). The ability to access large volumes of raw sequence data quickly in order to test hypotheses highlights the value of sequence data repositories to macroecological, and bioinformatics research.

In addition to data repositories, more specific, purpose-built and curated databases also enable the assemblage of datasets of use for testing macroecological hypotheses. In Chapter 6, I utilised the database MaarjAM in order to obtain global occurrence data for a variety of arbuscular mycorrhizal (AM) fungal taxa (Opik *et al.*, 2010). The main purpose of this database is to contain DNA sequences associated with AM fungi, in order to build a large taxon-specific database of all known AM fungal sequences. However, the additional strength of this database is in the metadata associated with the sequence records, which provides information including geographic coordinates, sampling year, and sample source. This metadata allows the sequence records to be more useful to other researchers by providing more contextual information in an accessible format. In turn, this facilitates the integration of these data with other datasets, such as climate data (Chapter 6) in order to test hypotheses. Furthermore, datasets with global extent (as is the case with MaarjAM) offer the opportunity to test macroecological theory without the enormous logistical and financial difficulty associated with generating new global datasets.

Data sharing in ecology is still a hotly debated (Reichman *et al.*, 2011; Hampton *et al.*, 2013), but widely encouraged ethos. In microbial ecology, many of the field specific journals request that any sequence data are uploaded to publicly accessible data repositories as a condition of publication. Whilst enforced, this has lead to the development of a minimum set of standards with which sequence data should be shared (Yilmaz *et al.*, 2011), thus ensuring that such data remain relevant and useful in the future.

Statistical Methods to Analyse Macroecological Relationships

The widespread uptake of high-throughput sequencing in microbial ecology has resulted in datasets with unusual or idiosyncratic properties that require careful consideration during statistical analyses. Challenges associated with the analysis of microbial datasets are the often extremely large numbers of species, sparseness (high proportion of zeros), a strong mean-variance relationship, and non-normality, all of which may violate the assumptions of common statistical tests (Warton *et al.*, 2012; Bálint *et al.*, 2016). Throughout this thesis, I have used advanced statistical methods, rarely applied in microbial ecology studies, in order to overcome these problems. In Chapter 5, the strong mean-variance relationship of OTU richness data, combined with the fact that count data are non-normally distributed (because counts are integers and bound by 0) meant that the assumption of standard linear regression analyses are violated (Ver Hoef & Boveng, 2007; O’Hara & Kotze, 2010). In particular, failing to account for the mean-variance relationship can influence estimates of model parameters (Ver Hoef & Boveng, 2007) and common model selection criteria such as AIC (Richards, 2008). Therefore, the use of generalised linear models circumvented the need for data transformation, resulting in more intuitive model outputs (O’Hara & Kotze, 2010). Furthermore, explicitly accounting for the strong mean-variance relationship present in the data allowed models to account for over-dispersion, and therefore make accurate parameter estimates (Warton *et al.*, 2016).

In addition to the challenging statistical properties associated with microbial datasets, the choice of appropriate indices to describe community properties, such as β-diversity, in microbial ecology has contributed to difficulty in testing macroecological theory. The past two decades has seen a considerable body of research devoted to the development of improved indices to quantify β-diversity in ecology, which has lead to more accurate and interpretable indices. Yet, the uptake of these new indices to quantify β-diversity has been slow within microbial ecology, perhaps due to uncertainty about how applicable such indices are to microbial communities, or more likely due to a lack of integration between the fields of microbial and statistical ecology. In Chapter 3, I recorded the range of indices used to test for distance-decay relationships within microbial ecology, and found that classic indices such as Bray-Curtis and Jaccard's index are overwhelmingly the most frequent choice. However, these indices have been shown to blur the two ways in which communities can change, through species turnover or nestedness (Baselga, 2010). This makes it difficult to determine the nature of changes occurring in microbial communities; are species being replaced along environmental gradients, or are species being lost from the community, resulting in nested subsets? A new generation of turnover indices have been developed that allow the separation of these two components, allowing for better insight into how microbial communities are changing in space, or along environmental gradients (Baselga, 2010, 2012). In Chapter 4, I used the β_{sim} index to cluster communities in order to test for biogeographic regions. This index represents a better choice than classic indices as it purely quantifies species turnover, which is the process relevant to biogeographic regionalisation, and not nestedness (e.g. biogeographic regions are not nested subsets of each other). Furthermore, in Chapter 5, I partitioned the β-diversity of stream sediment microbial communities into it's nestedness and turnover components, in order to determine the nature of community changes along thermal gradients. This analysis revealed that communities change along temperature gradients through species turnover, as distinct species occupy different parts of the temperature gradient. In contrast, there was no relationship between changes in temperature and the nestedness component of community turnover, showing that cold or hot sediment communities are not merely subsets of diverse “warm” communities, in contrast to previous results (Sharp *et al.*, 2014).

Recently developed community similarity indices are now able to account for unobserved species (Chao *et al.*, 2006), uncertainty in species' occurrences (Barbosa, 2015), and interactions between species (Schmidt *et al.*, 2017). These indices, along with those described previously (e.g. Baselga, 2010), represent exciting opportunities for microbial ecologists to move “beyond

Bray", and to more adequately test their hypotheses about the nature of changes in microbial community composition. The results of Chapter 3, along with Chapters 4 and 5, illustrate that the choice of similarity index can influence our perception of microbial macroecology, and careful consideration is needed in order to choose the most appropriate index to address a given hypothesis.

Microbial Study Systems for Macroecological Studies

The choice of study system can dramatically enhance or obscure patterns and processes in macroecological studies of microbial communities (Chapter 3). Microbial communities have been shown to respond to environmental variables that are structured at a variety of spatial scales (e.g. climate vs pH; Dumbrell *et al*., 2010; Pajunen *et al*., 2016). Therefore, choosing an appropriate study system to test for the processes or patterns of interest is key. Previous research has argued that experimental systems are ideal for testing hypotheses about microbial ecology (Jessup *et al.*, 2004). However, it is often not possible to test macroecological theories in a laboratory setting, whilst field based experiments offer relatively limited opportunities (Bell, 2010). Furthermore, in congruence with "macroecological thinking", field based studies often fail to replicate the complexity found within nature, and often provide over-simplified results that are of limited generality to the natural world (O'Gorman *et al.*, 2014). This point is particularly relevant to Chapter 6, which examined the extent to which current climate determines the distribution of arbuscular mycorrhizal fungi over global scales. *In situ* manipulations of single climatic factors such as temperature or precipitation often show strong community responses to these factors (e.g. Hawkes *et al*., 2011) yet, at the global scale, the distributions of many AM fungi are only weakly related to climate (Chapter 6). Therefore, whilst natural systems may be challenging due to the presence of confounding factors, by accounting for the complexity of the natural world, rather than trying to remove it, better generalisations can be made, and unifying ecological principles can be determined.

Microbial Macroecology

Generality of Macroecological Relationships

One of the major themes of macroecological research is the search for general patterns and unifying theories that unite the ecologies of different organisms (Keith *et al.*, 2012). In microbial macroecology, it has been suggested that the application of existing ecological theory to microbial communities will likely be the most fruitful approach, circumventing the requirement for new theory (Prosser *et al.*, 2007). However, whilst studies comparing the macroecological patterns of micro- and “macroorganisms” are useful and interesting (e.g. Horner-Devine *et al*., 2007; Astorga *et al*., 2012), this has meant that the generality of macroecological patterns within microorganisms has remained largely under-explored.

In Chapter 3, I found that the strength of the distance-decay relationship in microbial communities can vary according to several biological contextual aspects including study system, and scale, showing that the rate at which community similarity decays with distance is not universal across microbial communities in different habitats. Furthermore, in Chapter 5 I find that the relationship between microbial community diversity and temperature can vary considerably between study systems separated by thousands of kilometers. This is in contrast to previous studies showing that this relationship is conserved across microbial communities (Sharp *et al.*, 2014), and suggests that metacommunity dynamics and the availability of species to fill specific niches might have a role in determining the generality of macroecological relationships in microbial communities (Telford *et al.*, 2006). Finally, in Chapter 6, I find that the drivers and explanatory power of climatic variables on arbuscular mycorrhizal fungal distributions varies considerably between and within different AM fungal taxa. This suggests that AM fungi show climatic niche differentiation, and therefore will show different patterns of occurrence in relation to climatic variables.

Combined, these results highlight that macroecological relationships in microbial communities are unpredictable, whilst addressing some of the key unanswered questions in microbial macroecology (Lennon & Locey, 2017). The results gathered here suggest that metacommunity dynamics (Chapter 5), taxon-dependent niche differentiation (Chapters 5 and 6), biological context (Chapter 3), and methodology (Chapter 3) may all influence the presence and detectability of macroecological relationships in microbial communities. Therefore, understanding the generality of macroecological dynamics, and the factors that lead to idiosyncratic relationships in microbial macroecology could result in new theory to describe the macroecological processes that determine microbial community structure.

Spatial Processes in Microbial Macroecology

Due to the provocative nature of EiE, a common theme present in the microbial ecology literature is the balance between the environment and spatial processes in determining the composition and structure of microbial communities (van der Gast, 2015). Previous research has frequently framed such questions by investigating whether microbial communities assemble by niche (environmental) or neutral (random dispersal and speciation) processes (e.g. Dumbrell *et al*., 2010; Ofiteru *et al*., 2010; Lekberg *et al*., 2007). This body of research has yielded much insight into the often strong effect of the environment on microbial communities, but also into the less well understood role of spatial processes in determining microbial community composition over a range of spatial scales. Often, the decay of similarity in community composition with increasing geographic distance between communities is interpreted as evidence of neutral processes in the assembly of microbial communities, as neutral theory predicts that species will disperse more to neighbouring habitat patches than to distant ones (Rosindell *et al.*, 2011).

However, species within microbial communities can have very different distributions, and many species may be found in distant sites, but not in neighbouring sites (Chapter 4), making them poor indicators of a community's geographic origin. Given that the environment in which we sampled these communities was highly similar meaning that environmental effects are less likely, two explanations remain. The first is that connectivity, rather than geographic distance *per se*, determines the probability of dispersal between sites. In previous distance-decay studies within microbial ecology, distance and connectivity are assumed to be closely related, and therefore close sites are presumed to be more well connected, facilitating dispersal between them. However, empirical tests of this assumption are lacking (Müller *et al.*, 2014; Vannette *et al.*, 2016). Certain dispersal mechanisms or vectors could provide dispersal “motorways” connecting communities separated by large geographic distances. Intuitive examples in which connectivity may “override” distance are in stream networks (Niño-García *et al.*, 2016), ocean currents (Müller *et al.*, 2014), or animal migration routes (e.g. birds). In all of these examples, there is a clear mechanism by which long distance dispersal might connect geographically distant communities, or there is the potential for asymmetric dispersal between sites. Connectivity may therefore be a viable explanation for the lack of biogeographic regionalisation seen in communities of extremophilic microorganisms (Chapter 4). However, connectivity is potentially difficult to quantify, and may be easier to infer *ad hoc* (e.g. after examining compositional similarity), especially if the dispersal vector is not known. Therefore, careful consideration of study system is required in order to address hypotheses relating to connectivity in microbial communities.

The second explanation is that dispersal itself is a non-neutral process (Lowe & McPeek, 2014). If dispersal is a neutral process, all species should disperse equally successfully, and therefore the commonest species in a community will have the best chance to disperse to distant sites. However, in extremophilic communities, this is not necessarily true, as less abundant species were often observed in geographically distinct communities, making them poor indicators of a community's geographic origin (Chapter 4). Asymmetric dispersal ability between species could explain these species' wide distributions, as they may possess traits or behaviour that facilitates their long-distance dispersal. In microorganisms, dispersal-related traits may include the ability to enter a vegetative state in order to survive suboptimal environmental conditions (Norros *et al.*, 2014, 2015), or having a small cell size to facilitate aeolian (wind) dispersal (Wilkinson, 2001; Wilkinson *et al.*, 2012). Alternatively, the ability to survive in dispersible environmental material, such as halite crystals, could allow long distance dispersal, via wind blown particles for example.

In addition to determining the composition of microbial communities, spatial processes may also influence how microbial communities are shaped by environmental gradients (Chapter 5, Telford *et al*., 2006). The relationship between community diversity and an environmental gradient is determined by the niche use of the species pool available to colonise the habitat. Therefore, spatial processes such as dispersal may therefore determine the size of the species pool available to colonise a habitat, which in turn, will determine how diversity at the community level is related to an environmental gradient (Telford *et al.*, 2006).

Climate change and Microorganisms

Microorganisms may be the most vulnerable organisms to climate change, and may also have the largest impact on ecosystem functioning under climate change (Singh *et al.*, 2010). Microorganisms do show relationships with factors associated with climate (change) including temperature (Chapter 5, Zhou *et al*., 2016; Sharp *et al*., 2014), and precipitation (Chapters 6, Angel *et al*., 2010). However, the results gathered here indicate that the impacts of climate change on microbial communities are unlikely to be uniform. Changes in the diversity and composition of microbial communities are likely to vary spatially according to the regional metacommunity (Chapter 5), whereas changes in the functionality of microbial communities may depend on the identity of organisms present within the community and their relative contributions to ecosystem functioning (Chapter 6).

Evidence for whether functionality is linked to diversity within microbial communities is mixed (Nannipieri & Ascher, 2003; Peter *et al.*, 2011; Peter &

Sommaruga, 2016), and may depend on the function and identity of taxa within a community. This suggests that whilst climate change may affect the diversity of microbial communities, it is currently unclear as to whether this will lead to a change in functionality. Therefore, a better understanding of the identity of microorganisms that drive functionality, and their susceptibility to climate change is needed in order to build a more predictive model of how climate change will affect microbial ecosystem processes (Compant *et al.*, 2010).

In addition to effecting microbial communities, climatic changes may also act differentially on microorganisms at other levels of biological organisation, such as the population level. Niche differentiation with respect to climate means that the relative effects of climate on different microbial taxa will likely be variable, as the distributions of different taxa are linked to different climatic drivers, or the strength of relationship between climate and distribution is different (Chapter 6). In this case, the potential for climate to affect microbially mediated ecosystem processes is highly dependent on identity of the organisms present in a habitat, meaning that specific knowledge on the functionality of individual taxa is required in order to build a predictive understanding of climate change.

A “one size fits all” approach to predicting the impacts of climate change on microbial communities, populations, and functioning is therefore not appropriate and will result in poor predictions. In order to build more accurate predictive models for microbial functioning under climate change, all levels of biological organisation should be considered including metapopulation dynamics, and niche differentiation between populations.

Future Work

Are Biotic Interactions Important?

Over macroecological scales, the distributions of microbial taxa are determined by interactions with the environment, as well as their dispersal and connectivity between habitat patches. However, biotic interactions may also play a large role in controlling the distributions of microorganisms (Larsen *et al.*, 2012). Microbial taxa do not exist in isolation, they are part of diverse communities and may interact with other microorganisms, or "macroorganisms". These interactions may be obvious in some cases, for example between endosymbiotic fungi and their host plants, but often may be more subtle, for example between microorganisms occupying different steps of biogeochemical cycles. Such interactions may determine the spatial configuration of microorganisms, particularly at small scales e.g. in stratified biofilms (Elias & Banin, 2012). However, it is largely unknown as to whether biotic interactions between microorganisms, or between microorganisms and "macroorganisms" could be capable of structuring microbial communities over macroecological scales. In particular it is unknown whether strong biotic interactions could maintain the coexistence of species outside of their physicochemical niches. An example of one such possible interaction would be the co-occurrence of the archaeon, *Haloquadratum walsbyi*, with the bacterium, *Salinibacter ruber*. (Gramain *et al.*, 2011) showed that the survival of *H. walsbyi* in laboratory formed halite crystals was dramatically enhanced in the presence of *S. ruber,* perhaps due to co-metabolism of certain compounds (Elevi Bardavid & Oren, 2008). Other evidence suggests that such interactions may be relatively common in halophilic microbial communities (Elevi Bardavid *et al.*, 2008). In this instance, the traditional niche concepts of the fundamental niche (the entire range of survivable conditions), and the realised niche (the conditions in which the species actually occurs) are not intuitive. Biotic interactions are usually thought to restrict the fundamental niche, with the resulting realised niche representing a subset of the fundamental niche. However, in this case biotic interactions extend the fundamental niche, resulting in a larger realised niche. It is arguable that the fundamental and realised niche concepts do not adequately describe this concept, and the development of new niche concepts to incorporate the role of facilitation may be required, although arguably the concept of biotic interactions extending the fundamental niche has not been well characterised in the wider field of ecology (Bruno *et al.*, 2003).

Uniting Experimental Manipulations with Metacommunity Theory

It has been argued that *in situ* experimental manipulations of microbial communities may offer better insight into the effects of environmental change on microbial communities, than through observation based studies (Jessup *et al.*, 2004). Experimental manipulations have been widely employed to study the potential effects of climate change on microbial communities (e.g. Hawkes *et al*., 2011; Steven *et al*., 2012; Heinemeyer and Fitter, 2004). However, these studies have yielded varying results, perhaps due to regional metacommunity processes (Chapter 5), making it difficult to discern general impacts of climatic change in microbial communities. Therefore, explicitly accounting for macroecological processes, such as metacommunity dynamics, by conducting spatially replicated experimental manipulations over large geographic extents represents an elegant way of integrating the reductionist approach of experimental ecology with the complexity and unifying principles of macroecology (Lessard *et al.*, 2012). For example, a series of artificial warming experiments set out over a latitudinal gradient would allow a test of the whether regional metacommunities determine the response to temperature at local scales. If they do, one might expect that equatorial regions should show a higher thermal richness optima, as there are likely to be more species adapted to warm conditions present in the metacommunity. In contrast, polar regions might have lower thermal richness optima as there are fewer warm-adapted microbes, and therefore the majority of microbial diversity would be cold-adapted. Such a study could be conducted empirically or, given the number of similar previous experiments, may be possible via meta-analysis of previously published experiments. By uniting small scale experimental approaches with macroecological theory, a more realistic understanding of how climate change affects microbial communities could be gained.

References

Angel R, Soares MIM, Ungar ED, Gillor O (2010) Biogeography of soil archaea and bacteria along a steep precipitation gradient. *The ISME Journal*, **4**, 553–563.

Astorga A, Oksanen J, Luoto M, Soininen J, Virtanen R, Muotka T (2012) Distance decay of similarity in freshwater communities: Do macro- and microorganisms follow the same rules? *Global Ecology and Biogeography*, **21**, 365–375.

Bálint M, Bahram M, Eren AM, *et al*. (2016) Millions of reads, thousands of taxa: Microbial community structure and associations analyzed via marker genesa. *FEMS Microbiology Reviews*, **40**, 686–700.

Barberán A, Casamayor EO, Fierer N (2014) The microbial contribution to macroecology. *Frontiers in Microbiology*, **5**, 203.

Barbosa AM (2015) fuzzySim: Applying fuzzy logic to binary similarity indices in ecology. *Methods in Ecology and Evolution*, **6**, 853–858.

Baselga A (2010) Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, **19**, 134–143.

Baselga A (2012) The relationship between species replacement, dissimilarity derived from nestedness, and nestedness. *Global Ecology and Biogeography*, **21**, 1223–1232.

Bell T (2010) Experimental tests of the bacterial distance-decay relationship. *The ISME Journal*, **4**, 1357–1365.

Bruno JF, Stachowicz JJ, Bertness MD (2003) Inclusion of facilitation into ecological theory. *Trends in Ecology and Evolution*, **18**, 119–125.

Chao A, Chazdon RL, Colwell RK, Shen TJ (2006) Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics*, **62**, 361–371.

Compant S, van der Heijden MGA, Sessitsch A, *et al*. (2010) Climate change effects on beneficial plant-microorganism interactions. *FEMS Microbiology Ecology*, **73**, 197–214.

Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH (2010) Relative roles of niche and neutral processes in structuring a soil microbial community. *The ISME Journal*, **4**, 337–345.

Elevi Bardavid R, Oren A (2008) Dihydroxyacetone metabolism in *Salinibacter ruber* and in *Haloquadratum walsbyi*. *Extremophiles*, **12**, 125–131.

Elevi Bardavid R, Khristo P, Oren A (2008) Interrelationships between *Dunaliella* and halophilic prokaryotes in saltern crystallizer ponds. *Extremophiles*, **12**, 5–14.

Elias S, Banin E (2012) Multi-species biofilms: Living with friendly neighbors. *FEMS Microbiology Reviews*, **36**, 990–1004.

Gramain A, Díaz GC, Demergasso C, Lowenstein TK, Mcgenity TJ (2011) Archaeal diversity along a subterranean salt core from the Salar Grande (Chile). *Environmental Microbiology*, **13**, 2105–2121.

Hampton SE, Strasser CA, Tewksbury JJ, *et al*. (2013) Big data and the future of ecology. *Frontiers in Ecology and the Environment*, **11**, 156–162.

Hawkes CV, Kivlin SN, Rocca JD, Huguet V, Thomsen MA, Suttle KB (2011) Fungal community responses to precipitation. *Global Change Biology*, **17**, 1637–1645.

Heinemeyer A, Fitter AH (2004) Impact of temperature on the arbuscular mycorrhizal (AM) symbiosis: Growth responses of the host plant and its AM fungal partner. *Journal of Experimental Botany*, **55**, 525–534.

Horner-Devine MC, Silver JM, Leibold MA, *et al*. (2007) A comparison of taxon co-occurrence patterns for macro- and microorganisms. *Ecology*, **88**, 1345–1353.

Jessup CM, Kassen R, Forde SE, Kerr B, Buckling A, Rainey PB, Bohannan BJM (2004) Big questions, small worlds: Microbial model systems in ecology. *Trends in Ecology and Evolution*, **19**, 189–197.

Keith S a, Webb TJ, Böhning-Gaese K, *et al*. (2012) What is macroecology? *Biology letters*, **8**, 904–906.

Larsen PE, Field D, Gilbert JA (2012) Predicting bacterial community assemblages using an artificial neural network approach. *Nature Methods*, **9**, 621–625.

Leinonen R, Akhtar R, Birney E, *et al*. (2011a) The European nucleotide archive. *Nucleic Acids Research*, **39**, D28–D31.

Leinonen R, Sugawara H, Shumway M (2011b) The sequence read archive. *Nucleic Acids Research*, **39**, D19-21.

Lekberg Y, Koide RT, Rohr JR, Aldrich-Wolfe L, Morton JB (2007) Role of niche restrictions and dispersal in the composition of arbuscular mycorrhizal fungal communities. *Journal of Ecology*, **95**, 95–105.

Lennon JT, Locey KJ (2017) Macroecology for microbiology. *Environmental Microbiology Reports*, **9**, 38–40.

Lessard JP, Belmaker J, Myers JA, Chase JM, Rahbek C (2012) Inferring local ecological processes amid species pool influences. *Trends in Ecology and Evolution*, **27**, 600–607.

Liu L, Yang J, Yu Z, Wilkinson DM (2015) The biogeography of abundant and rare bacterioplankton in the lakes and reservoirs of China. *The ISME Journal*, **9**, 2068–2077.

Lowe WH, McPeek MA (2014) Is dispersal neutral? *Trends in Ecology and Evolution*, **29**, 444–450.

Meyer F, Paarmann D, D’Souza M, *et al.* (2008) The metagenomics RAST server—a public resource for the automatic phylo- genetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Müller AL, de Rezende JR, Hubert CRJ, *et al*. (2014) Endospores of thermophilic bacteria as tracers of microbial dispersal by ocean currents. *The ISME Journal*, **8**, 1153–1165.

Nannipieri P, Ascher J, Ceccherini M, Landi L, Pietramellara G, Renella G (2003) Microbial diversity and soil functions. *European Journal of Soil Science*, **54**, 655-670.

Neufeld JD (2017) Migrating SSU rRNA gene surveys to the metagenomics era. *Environmental Microbiology Reports*, **9**, 23–24.

Niño-García JP, Ruiz-González C, del Giorgio PA (2016) Interactions between hydrology and water chemistry shape bacterioplankton biogeography across boreal freshwater networks. *The ISME Journal*, **10**, 1755–1766.

Norros V, Rannik Ü, Hussein T, Petäjä T, Vesala T, Ovaskainen O (2014) Do small spores disperse further than large spores? *Ecology*, **95**, 1612–1621.

Norros V, Karhu E, Nordén J, Vähätalo A V., Ovaskainen O (2015) Spore sensitivity to sunlight and freezing can restrict dispersal in wood-decay fungi. *Ecology and Evolution*, **5**, 3312–3326.

O’Gorman EJ, Benstead JP, Cross WF, *et al*. (2014) Climate change and geothermal ecosystems: Natural laboratories, sentinel systems, and future refugia. *Global Change Biology*, **20**, 3291–3299.

O'Hara RB, Kotze DJ (2010) Do not log-transform count data. *Methods in Ecology and Evolution*, **1**, 118–122.

Ofiteru ID, Lunn M, Curtis TP, Wells GF, Criddle CS, Francis CA, Sloan WT (2010) Combined niche and neutral effects in a microbial wastewater treatment community. *Proceedings of the National Academy of Sciences USA*, **107**, 15345–15350.

Opik M, Vanatoa A, Vanatoa E, *et al*. (2010) The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytologist*, **188**, 223–241.

Pajunen V, Luoto M, Soininen J (2016) Climate is an important driver for stream diatom distributions. *Global Ecology and Biogeography*, **25**, 198–206.

Peter H, Sommaruga R (2016) Shifts in diversity and function of lake bacterial communities upon glacier retreat. *The ISME Journal*, **10**, 1545–1554.

Peter H, Beier S, Bertilsson S, Lindström ES, Langenheder S, Tranvik LJ (2011) Function-specific response to depletion of microbial diversity. *The ISME Journal*, **5**, 351–361.

Poretsky R, Rodriguez-R LM, Luo C, Tsementzi D, Konstantinidis KT (2014) Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS ONE*, **9**, e93827.

Prosser JI, Bohannan BJM, Curtis TP, *et al*. (2007) The role of ecological theory in microbial ecology. *Nature Reviews Microbiology*, **5**, 384–392.

Reichman OJ, Jones MB, Schildhauer MP (2011) Challenges and Opportunities of Open Data in Ecology. *Science*, **331**, 703–705.

Richards SA (2008) Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, **45**, 218-227.

Rosindell J, Hubbell SP, Etienne RS (2011) The unified neutral theory of biodiversity and biogeography at age ten. *Trends in Ecology and Evolution*, **26**, 340–348.

Schmidt TSB, Matias Rodrigues JF, von Mering C (2017) A family of interaction-adjusted indices of community similarity. *The ISME Journal*,

11, 791–807.

Sharp CE, Brady AL, Sharp GH, Grasby SE, Stott MB, Dunfield PF (2014) Humboldt’s spa: microbial diversity is controlled by temperature in geothermal environments. *The ISME Journal*, **8**, 1166–1174.

Singh BK, Bardgett RD, Smith P, Reay DS (2010) Microorganisms and climate change: terrestrial feedbacks and mitigation options. *Nature Reviews Microbiology*, **8**, 779–790.

Steven B, Gallegos-Graves LV, Yeager CM, Belnap J, Evans RD, Kuske CR (2012) Dryland biological soil crust cyanobacteria show unexpected decreases in abundance under long-term elevated CO2. *Environmental Microbiology*, **14**, 3247–3258.

Telford RJ, Vandvik V, Birks HJB (2006) Dispersal Limitations Matter for Microbial Morphospecies. *Science*, **312**, 1015–1015.

Tessler M, Neumann JS, Afshinnekoo E, *et al*. (2017) Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Scientific Reports*, **7**, 6589.

van der Gast CJ (2015) Microbial biogeography: the end of the ubiquitous dispersal hypothesis? *Environmental Microbiology*, **17**, 544–546.

ver Hoef JM, Boveng PL (2007) Quasi-poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, **88**, 2766–2772.

Vannette RL, Leopold DR, Fukami T (2016) Forest area and connectivity influence root-associated fungal communities in a fragmented landscape. *Ecology*, **97**, 2374–2383.

Warton DI, Wright ST, Wang Y (2012) Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, **3**, 89–101.

Warton DI, Lyons M, Stoklosa J, Ives AR, Schielzeth H (2016) Three points to consider when choosing a LM or GLM test for count data. *Methods in Ecology and Evolution*, **7**, 882–890.

Wilkinson DM (2001) What is the upper size limit for cosmopolitan distribution in free-living microorganisms? *Journal of Biogeography*, **28**, 285–291.

Wilkinson DM, Koumoutsaris S, Mitchell EAD, Bey I (2012) Modelling the effect of size on the aerial dispersal of microorganisms. *Journal of*

Biogeography, **39**, 89–97.

Yilmaz P, Kottmann R, Field D, *et al*. (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, **29**, 415–420.

Zhou J, Deng Y, Shen L, *et al*. (2016) Temperature mediates continental-scale diversity of microbes in forest soils. *Nature Communications*, **7**, 12083.

Appendices

Appendix 1

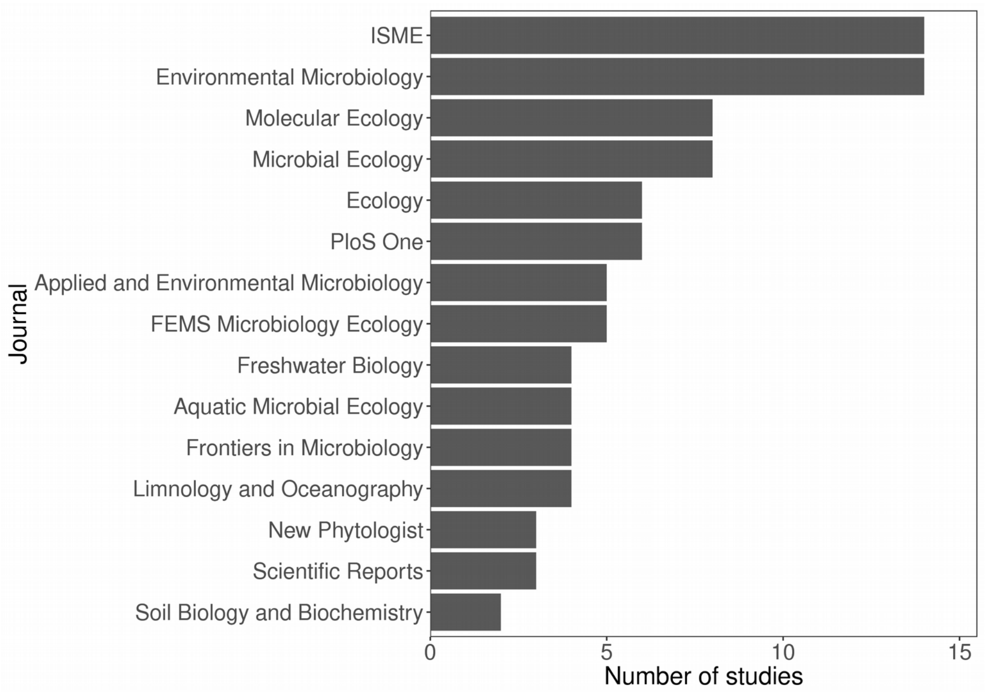


Figure 3.S1 The number of distance-decay relationships used in this analysis from different journals. Only the most frequent 15 journals are shown.

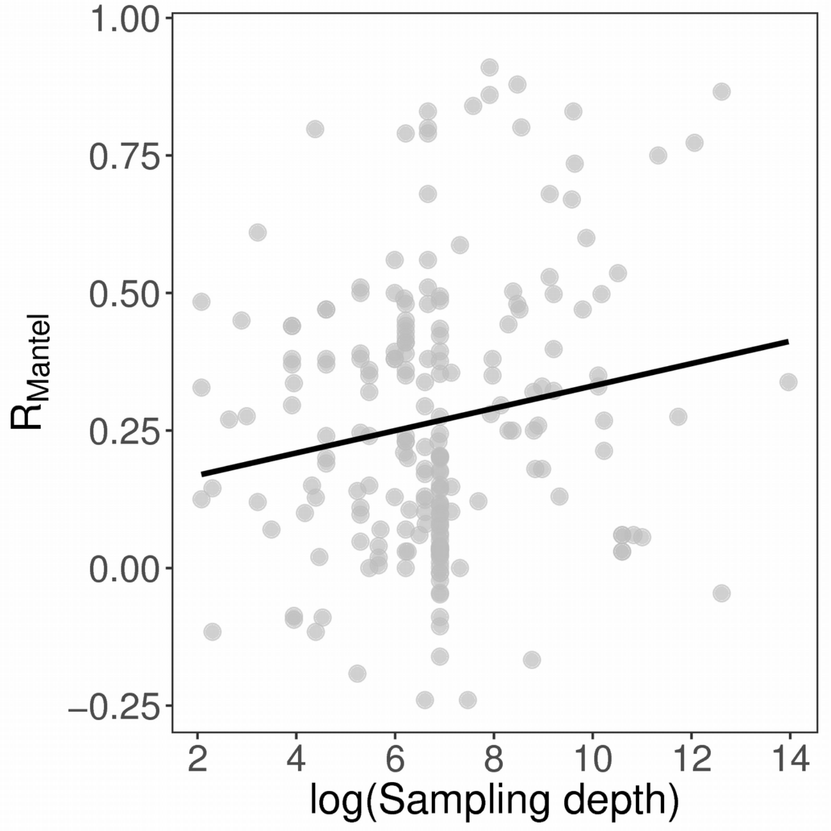
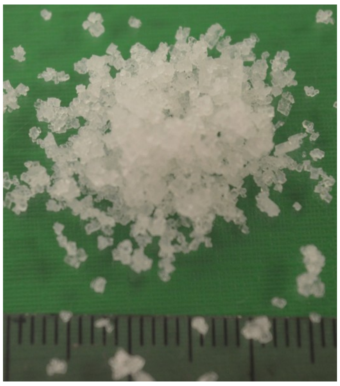
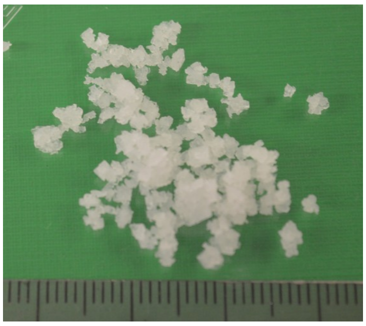
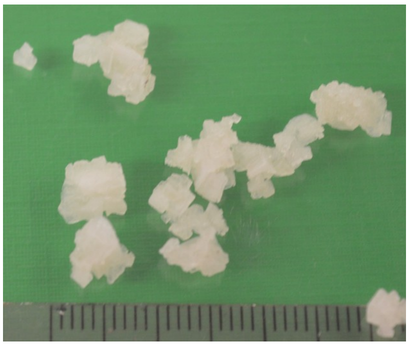
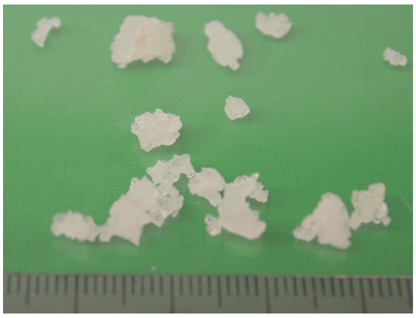
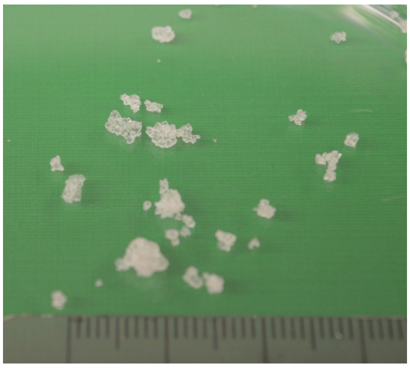
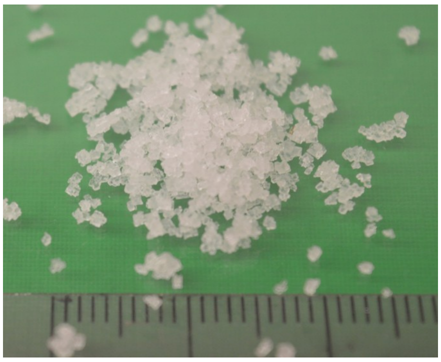
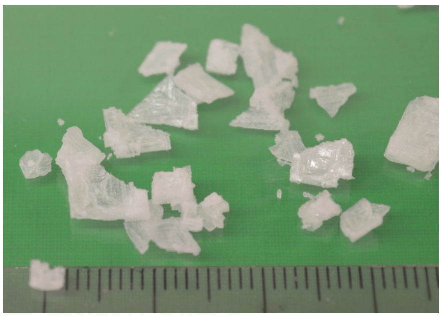


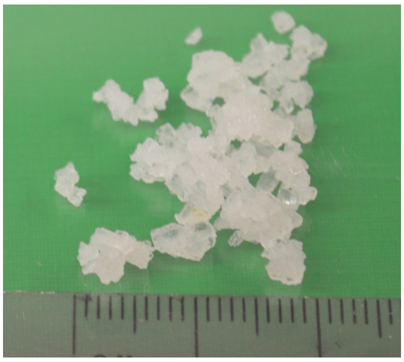
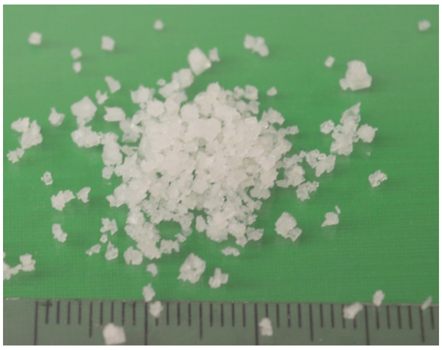
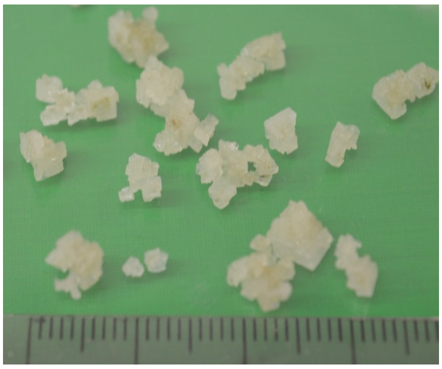
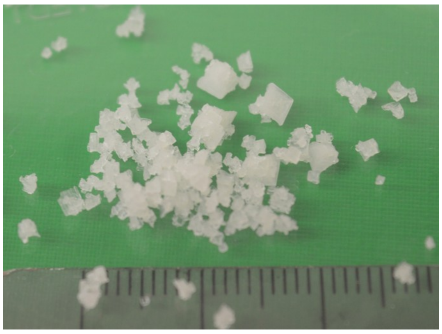
Figure 3.S2 The relationship between Mantel correlation coefficients and sampling depth. The solid line is the fit from a linear model (slope = 0.02, *P* < 0.05, adj-R^2 = 0.02). Sampling depth refers to the sequencing depth of sequence-based approaches, or the number of individuals counted for morphological studies. Fingerprinting studies are excluded from this analysis.

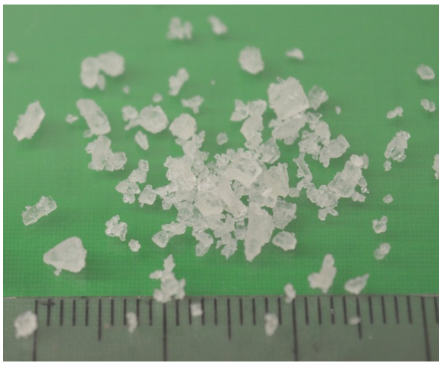
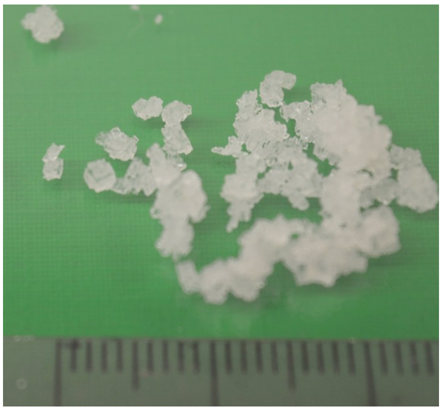
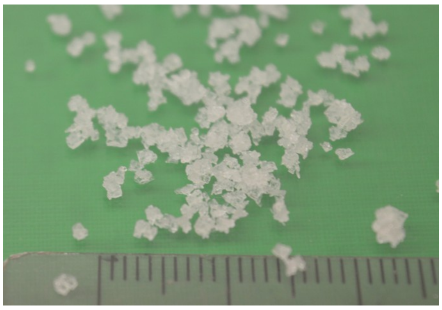
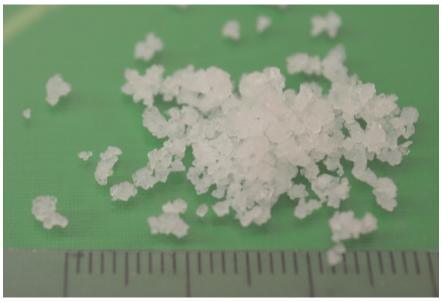
Appendix 2

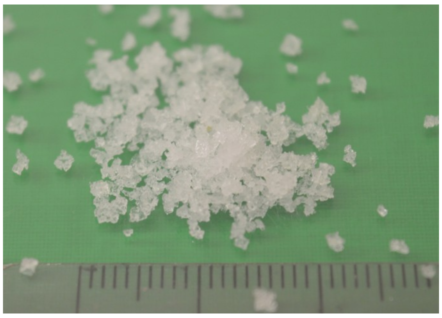
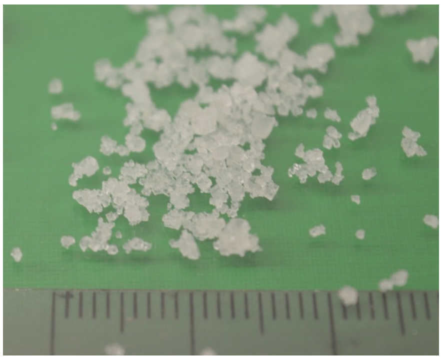
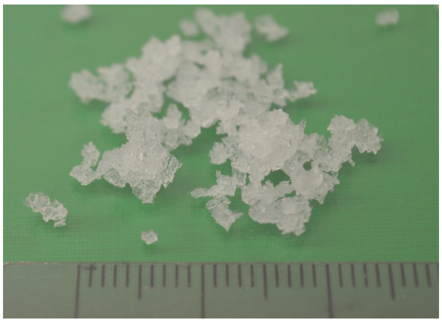
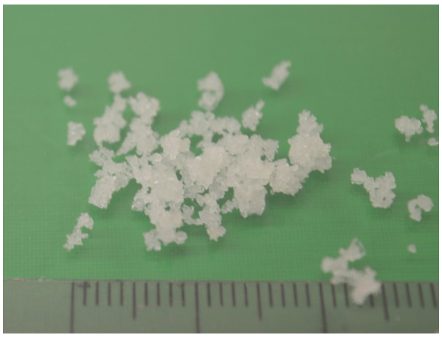
Table 4.S1 Photographic record of all halite samples used in this study. Scale shows mm increments.

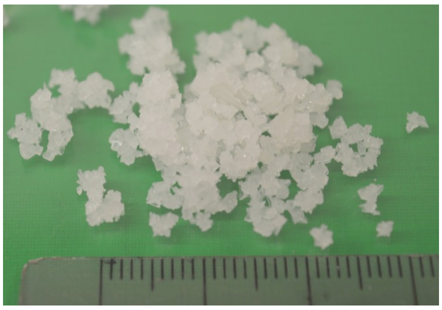
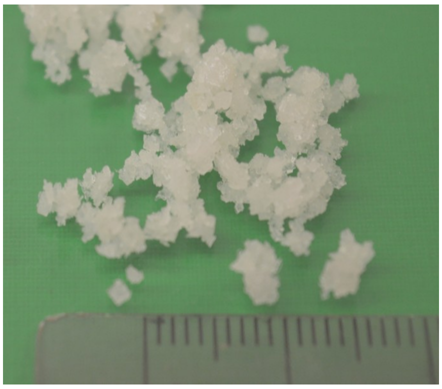
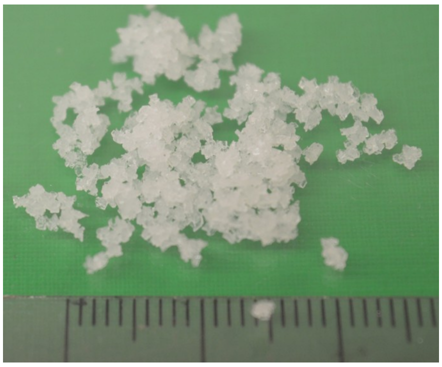
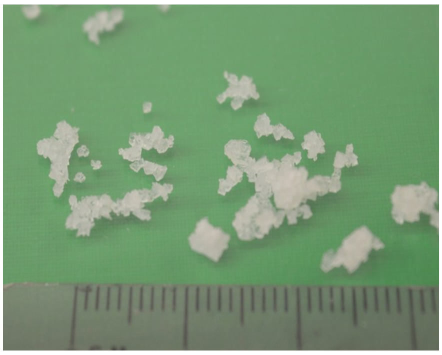
Sample photo	Notes	Location	Sample code
	Pure white, grain size ~ 1mm	Aigues Mortes, South France	AIGX
	Pure white, grain size 1-2mm	Algarve, Portugal	ALGX
	Slight grey hue, grain size 1-3mm	Bourgneuf, West France	BOUX

	Light grey hue, grain size 1-3mm	Cadiz, South Spain	CADX
	Pure white, grain size ~1mm	Camargue, South France	CAM1_X
	Pure white, grain size ~1mm	Camargue, South France	CAMX
	Pure white, 2-7mm	Cyprus	FALX

	Light grey hue, grain size 1-3mm	Fuencaliente, Canary Islands	FUENX
	White, grain size ~1mm	Guerande, West France	GUE1_X
	Grey-green, grain size 1-4mm	Guerande, West France	GUE2_X
	Light grey hue, grain size 1-2mm	Guerande, West France	GUE3_X

	Pure white, grain size ~ 1mm	Ibiza, Balearic Islands	IBIX
	Pure white, grain size ~1mm	Ifaty, Madagascar	IFA1_X
	Pure white, grain size ~1mm	Ifaty, Madagascar	IFAX
	White, grain size 1-2mm	St. Leu, Reunion Island	LEU1_X

	Pure white, grain size ~1mm	Mauritius	MAU1_X
	Pure white, grain size ~1mm	Mauritius	MAU2_X
	Pure white, grain size 1-2mm	Mauritius	MAU3_X
	White, grain size ~1mm	Mayotte	MAYX

	White, grain size ~1mm	Noirmoutier, West France	NOI1_X
	Slight grey hue, grain size 1-2mm	Noirmoutier, West France	NOI2_X
	White, grain size ~ 1mm	Noirmoutier, West France	NOI3_X
	White, grain size ~1mm	Aveiro, Portugal	PORX

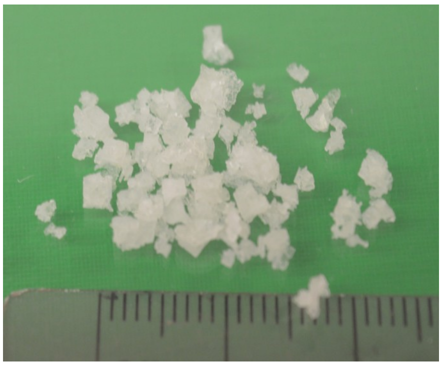
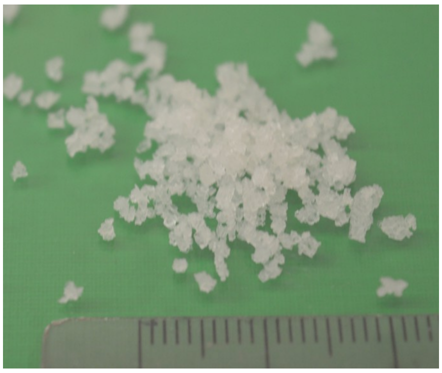
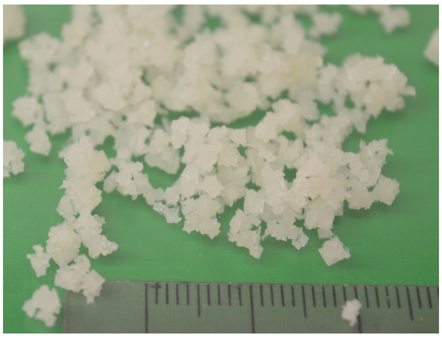
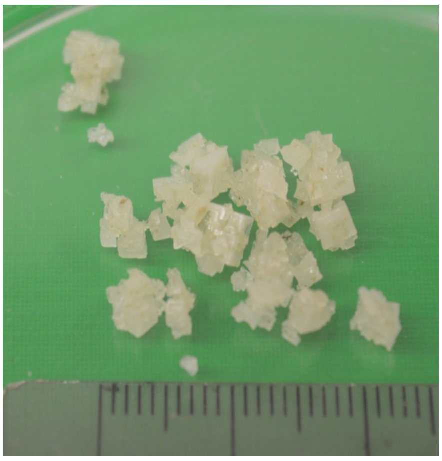
	White, grain size 1-2mm	Ile de Re, West France	SRE1_X
	White, grain size 1-2mm	Ile de Re, West France	SREX
	Grey-green hue, grain size 1-3mm	Ile de Re, West France	SRE2_X
	Strong grey hue, grain size 1-3mm	St. Armel, West France	STAX

Table 4.S2 The number of operational taxonomic units (OTUs) in each family identified to each genus. OTUs not identified to the genus level are binned under “unclassified” in the relevant family. OTUs not identified to family level are binned under the taxonomic group they were able to be identified to.

Archaeal family	Genus	Number of OTUs
Halobacteriaceae	*Haladaptatus*	4
	Halapricum	3
	Halarchaeum	5
	Haloarchaeobius	2
	Haloarcula	20
	Halobacterium	55
	Halococcus	13
	Halomarina	4
	Halomicroarcula	35
	Halomicrobium	9
	Halorhabdus	6
	Halorientalis	31
	Halorubellus	13
	Halorussus	10
	Halosimplex	5
	Halovenus	21
	Natronoarchaeum	4
	Natronomonas	63
	Salarchaeum	7
	Salinirubrum	1
	Unclassified Halobacteriaceae	545
Haloferacaceae	*Halobaculum*	8
	Halobellus	41
	Haloferax	1
	Halogeometricum	2
	Halogranum	3
	Halohasta	13
	Halolamina	64
	Halonotius	7

	Halopenitus	5
	Haloplanus	31
	Haloquadratum	10
	Halorubrum	58
	Salinigranum	4
	Unclassified Haloferacaceae	52
Methanosarcinaceae	*Methanohalobium*	1
Natrialbaceae	*Haloterrigena*	9
	Halovivax	3
	Natrinema	4
	unclassified_Natrialbaceae	14
Nitrososphaeraceae	*Nitrososphaera*	2
Unclassified Archaea	Unclassified	18
Unclassified Euryarchaeota	Unclassified	4
Unclassified Halobacteria	Unclassified	232
Unclassified Nanohaloarchaeota	*Candidatus Nanosalina*	137
Unclassified Woesarchaeota	Unclassified	2

Table 4.S3 The top ten archaeal genera that contribute to the accuracy of each classifier[a], as defined by node purity and classifier accuracy. Higher values for mean decrease in Gini index or accuracy indicate a greater contribution to the accuracy of the classifier.

Classifier	Node purity		Accuracy	
	Genus	Mean decrease in Gini index	Genus	Mean decrease in accuracy
Ocean	*Halarchaeum*	2.20	*Halarchaeum*	48.14
	Halohasta	1.86	*Halohasta*	42.22
	Halomicrobium	1.17	*Halomicrobium*	34.42
	Halovenus	1.02	*Halomicroarcula*	28.13
	Halapricum	0.99	*Halosimplex*	27.38
	Halosimplex	0.94	*Halovenus*	25.45
	Halorubrum	0.81	*Halorubrum*	25.00
	Halomicroarcula	0.78	*Halapricum*	24.74
	Halobaculum	0.65	*Halobacterium*	24.31
	Halobacterium	0.55	*Candidatus Nanosalina*	22.24
Geographic region	*Haloquadratum*	2.82	*Haloquadratum*	58.08
	Halapricum	2.52	*Halapricum*	55.39
	Halobaculum	1.84	*Halobaculum*	47.21
	Halarchaeum	1.48	*Halarchaeum*	42.38
	Halomicrobium	1.15	*Halohasta*	37.37
	Halohasta	1.13	*Salinigranum*	35.15
	Salinigranum	1.09	*Natrinema*	34.08
	Haloarcula	0.93	*Halomicrobium*	32.98
	Halorhabdus	0.92	*Halorubrum*	32.66
	Natrinema	0.91	*Halomicroarcula*	32.30
Biogeographic region	*Halarchaeum*	2.17	*Halarchaeum*	48.34
	Halohasta	1.89	*Halohasta*	44.58
	Halovenus	1.18	*Halovenus*	31.50
	Halapricum	1.15	*Halomicrobium*	30.34
	Halomicrobium	0.97	*Halosimplex*	30.23
	Halosimplex	0.92	*Halapricum*	28.92
	Halobaculum	0.84	*Halomicroarcula*	26.39
	Halorubrum	0.75	*Halorubrum*	25.69

	Halomicroarcula	0.64	*Halobacterium*	25.07
	Halorubellus	0.58	*Salarchaeum*	24.96

[a] Three classifiers were constructed to classify the nearest ocean, geographic region, and *a priori* defined biogeographic region of each community based on the relative abundances of haloarchaeal genera.

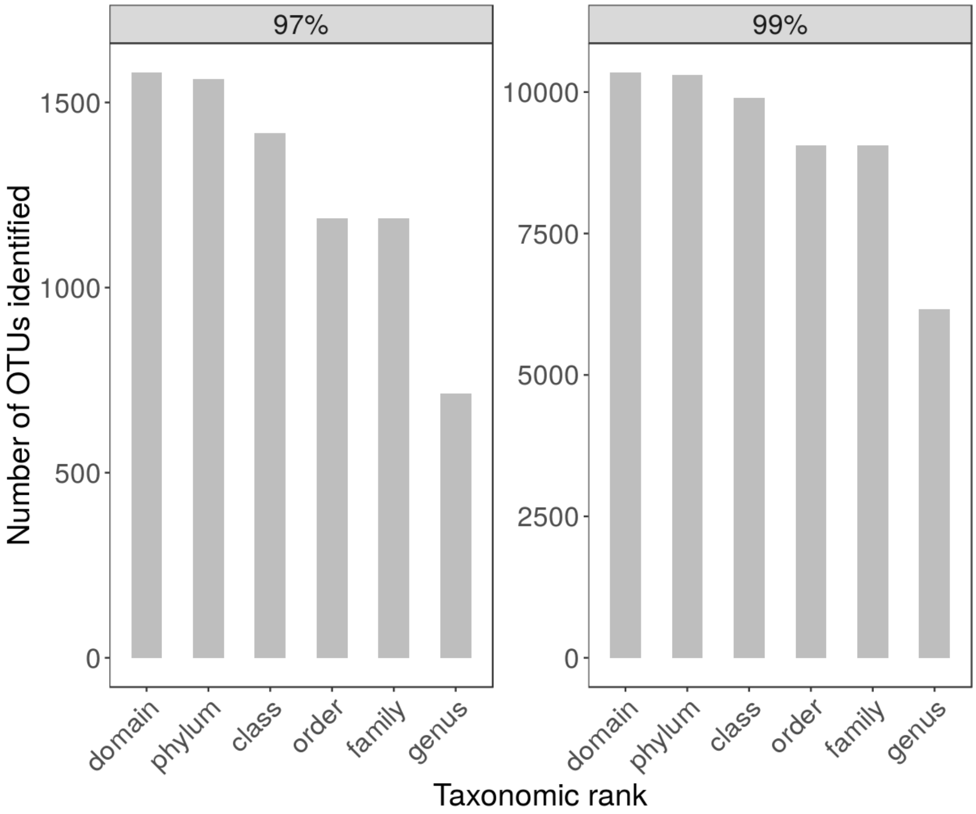


Figure 4.S1 The number of operational taxonomic units (OTUs) identified to each taxonomic level. The number of OTUs ranged from 1,581 in the 97% dataset to 10,346 in the 99% dataset.

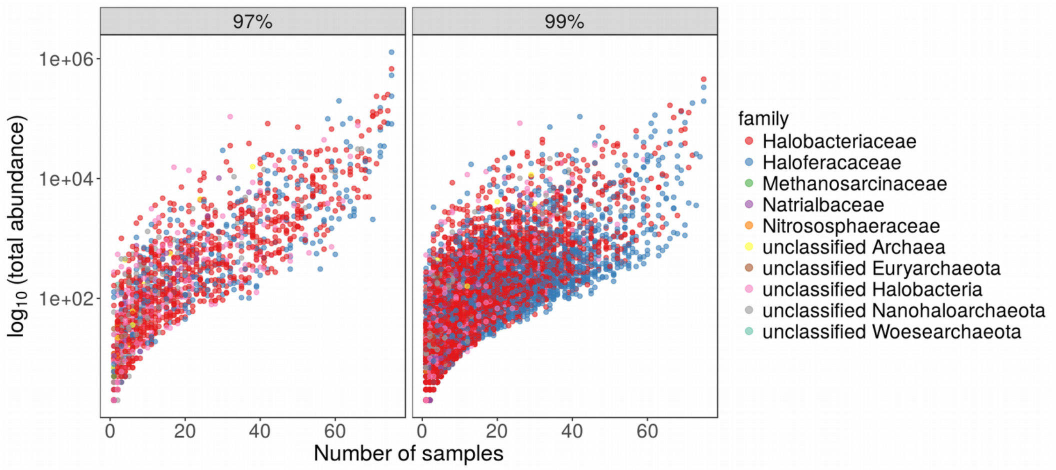


Figure 4.S2 The occupancy and abundance of all operational taxonomic units (OTUs). Total abundance is the abundance of each OTU in the entire dataset, prior to rarefaction. The total number of samples used in the study was 75.

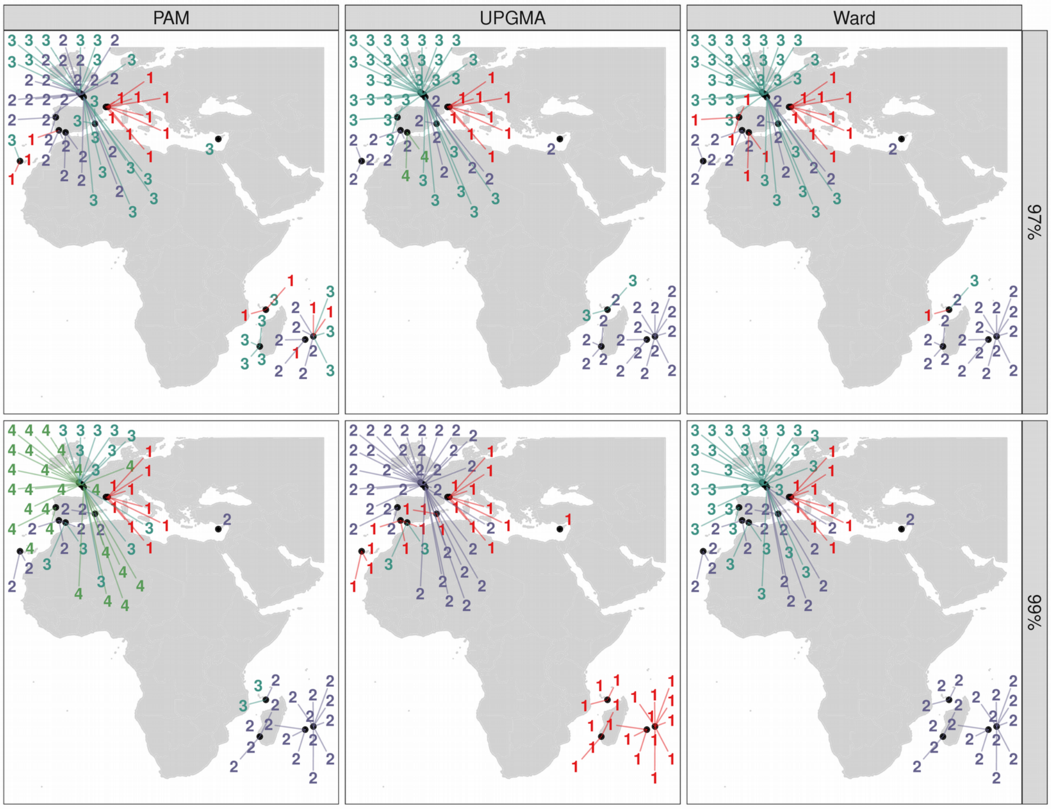


Figure 4.S3 The cluster memberships (indicated by colour and number of label) of communities for each clustering algorithm, based on the “knee

solution". This is the clustering solution that yielded the greatest increase in explained dissimilarity. Panel columns show results from the three different clustering algorithms used which were unweighted pair group method (UPGMA), partitioning around mediods (PAM), and Ward clustering method.

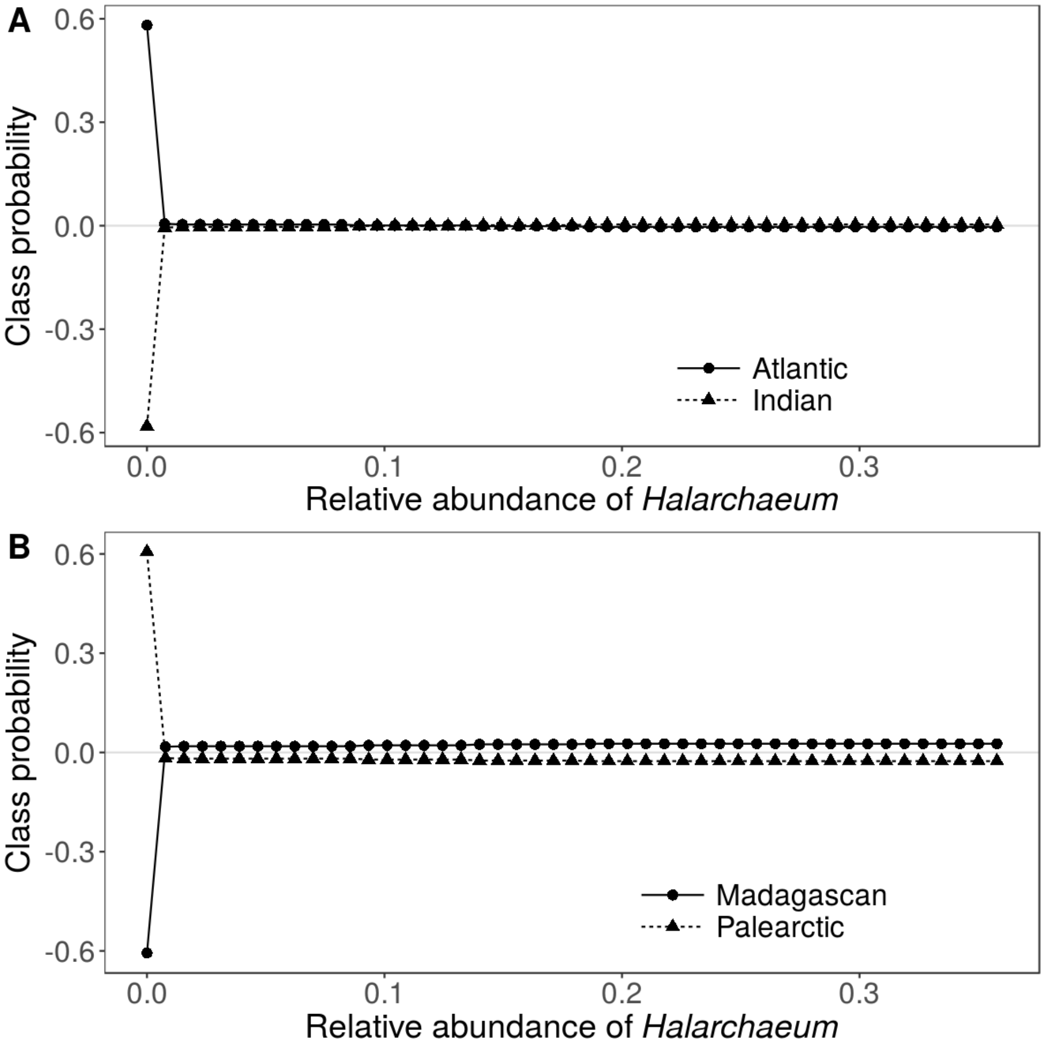


Figure 4.S4 Partial dependence plots showing how the relative abundance of the genus, *Halarchaeum*, influenced the class probabilities of the (A) oceanic and (B) biogeographic region random forest classifiers. Increased class probability indicates a higher probability that the random forest classifier will identify a community as belonging to a given class.

Appendix 3

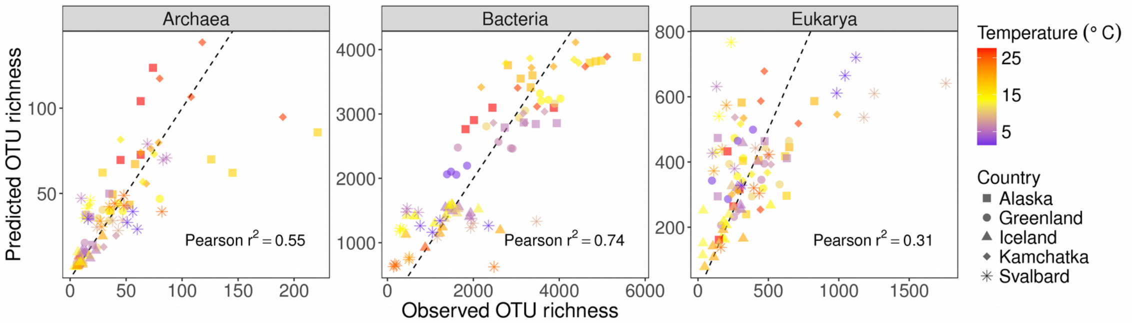


Figure 5.S1 The observed and predicted OTU richness for each of the taxonomic groups. Predictions are from the generalised linear mixed effects models as described in the manuscipt. The square of Pearson's correlation coefficient is included as an indicator of the predictive performance of each model. The dashed line indicates a 1:1 relationship between fitted and observed values.