

EVOLUTIONARY CONSEQUENCES OF GENE FLOW IN THE
ABSENCE OR INHIBITION OF DISPERSAL IN MICROBIAL
COMMUNITIES

by

Robert Mason Munro-Ehrlich

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

of

Doctor of Philosophy

in

Microbiology

MONTANA STATE UNIVERSITY
Bozeman, Montana

December 2023

©COPYRIGHT

by

Robert Mason Munro-Ehrlich

2023

All rights reserved

ACKNOWLEDGEMENTS

I would like to thank Dr. Jovanka Voyich for all her help during the thesis preparation and defense process. Along those lines, I would like to thank all my committee members, Drs. Alexis Templeton, Frank Stewart, Rachel Spietz, and Matt Taylor, for their continued support and guidance.

I would also like to thank my parents, as well as the past and present members of the Boyd lab: Drs. Eric Boyd, Dan Colman, Elizabeth Fones, Eric Dunham, Devon Payne, as well as Lisa Keller, Maria Fernandes-Martins, Ana Menchaca, and Su Cosar. I am also grateful to collaborators from the Templeton lab, the Rand lab, the Walk lab, and the DuBois lab.

TABLE OF CONTENTS

1. CHAPTER ONE: INTRODUCTION TO MICROBIAL GENE FLOW AND ITS IMPACT ON EVOLUTION	1
Gene Flow and Selection	1
Mechanisms of Microbial Gene Flow.....	3
Transformation.....	3
Transduction.....	4
Conjugation.....	5
Vesiduction.....	6
Relevance of Gene Flow to Microbial Evolution	7
Impact of the Local Environment on Dispersal and Gene Flow	8
Microbial Dispersal and its Barriers	8
Barriers to Dispersal Versus Barriers to Gene Flow	10
The Role of Gene Flow on Microbial Evolution	11
The Combined Impacts of Dispersal and Gene Flow on Microbial Evolution.....	15
Hypothesis.....	22
2. PARAPATRIC SPECIATION OF <i>MEIOTHERMUS</i> IN SERPENTINITE- HOSTED AQUIFERS IN OMAN	25
Contribution of Authors and Co-Authors.....	25
Manuscript Information.....	26
Abstract	27
Introduction.....	28
Materials and Methods.....	32
Site Description, Drilling, and Sampling.....	32
DNA Extraction and Shotgun Metagenomic Sequencing.....	33
Metagenomic Sequence Assembly, Binning, and Metagenome Assembled Genome (MAG) Metabolic Predictions	33
Compilation of <i>Meiothermus</i> Genomes and MAGs	34
Phylogenomic Analyses of <i>Meiothermus</i> -Affiliated MAGs.....	35
Estimation of <i>Meiothermus</i> Genome Replication Rates	36
Assignment and Characterization of <i>Meiothermus</i> Protein Clusters	36
Identification and Curation of Single Nucleotide Variants (SNVs).....	37
Characterization of SNVs	40
Evaluation of the Correlation Between Shared SNV Presence and Pairwise MAG Sequence Similarity	41
Estimating the Likelihood that Shared SNVs Arose Through Convergent Evolution	42

TABLE OF CONTENTS CONTINUED

Identification of Viral and Integrated Mobile Genetic Elements (IMGEs)	43
Results.....	44
Overview of Previous Hydrogeochemical Work Conducted at BA1A	44
BA1A Taxonomic Composition and Diversity	45
Inferred Oxygen Usage Among BA1A Populations	47
<i>Meiothermus</i> Phylogeny and Clade Designations	50
<i>Meiothermus</i> Inferred Genome Replication Rates.....	51
<i>Meiothermus</i> Genome Sizes	53
Protein Clustering of BA1A <i>Meiothermus</i> MAGs and their Inferred Metabolism	53
SNV Profiles of Individual BA1A <i>Meiothermus</i> MAGs	55
SNVs Shared Between MAGs	57
Correlation Between Shared SNVs and Pairwise MAG Genome Similarity.....	58
Likelihood of SNVs Arising Converently	61
Identification of Viral and Non-Viral Integrated Mobile Genetic Elements (IMGEs) in BA1A Communities	62
Discussion.....	65
Data availability statement.....	72
Author contributions	72
Funding	72
Acknowledgments.....	73
Conflict of interest	73
Publisher's note.....	74
 3. CHAPTER 3 CONCLUSIONS AND FUTURE STUDIES:.....	 75
Conclusions:.....	75
Future Directions	81
REFERENCES CITED.....	85

LIST OF TABLES

Table	Page
1. TABLE 1 Select geochemical measurements of waters recovered from discrete depth intervals in well BA1A in February 2019 and their inferred water type.....	45
2. TABLE 2 Name, estimated abundance, estimated genome size (mega basepairs, Mbp), estimated number of protein coding genes, strain level metagenomic estimation of growth rate (SMEG), and number of single nucleotide variants detected for metagenome assembled genomes (MAGs) recovered from discrete intervals in well BA1A..	52
3. TABLE 3 Single nucleotide variants (SNVs) identified between the specified pair or triplet of metagenome assembled genomes (MAGs).....	61
4. TABLE 4 Percent shared nucleic acid sequence identities between pairs of mobile genetic elements (MGE) retrieved from BA1A Meiothermus metagenome assembled genomes (MAGs).....	64
5. TABLE 5 Percent nucleic acid identities between pairs of transposon 125 (Tn125) sequence homologs retrieved from BA1A Meiothermus metagenome assembled genomes (MAGs).....	67

LIST OF FIGURES

Figure	Page
1. Figure 1.1 Schematic depicting common mechanisms of horizontal gene transfer.....	6
2. Figure 1.2 Schematics depicting microbial speciation regimes.....	14
3. Figure 1.3 Schematic depicting packer system employed in well BA1A. The schematic indicates the three depth intervals in which sampling takes place partitioned by the packer system. The sampling mechanism at the opening of the well is from a photograph taken during sampling season in 2019.	21
4. Figure 2.1. Phylum level taxonomic composition of MAGs	47
5. Figure 2.2 Maximum-likelihood phylogenomic reconstruction of a concatenation of 115 housekeeping genes from metagenome assembled genomes (MAGs).....	49
6. Figure 2.3 Average nucleotide identity of contigs between select pairs of metagenome assembled genomes (MAGs).....	57
7. Figure 3.1 visual abstract. Schematic depicting proposed evolutionary dynamics of the Type I surface (S) populations and the Type II deep (D) populations of <i>Meiothermus</i> in three discrete depth intervals	81

ABSTRACT

Much of our understanding of the evolutionary dynamics of microbial populations is derived from population level studies which focus on the immediately present populations and ignore the contributions of nearby communities. Microbial ecology studies typically do not distinguish between gene flow, i.e., the movement of genetic material between populations, and dispersal, i.e., the movement of those populations themselves. These two processes are indeed linked, but not identical. We have known for centuries that genetic material can be transferred between physically distant and taxonomically disparate microbial populations; molecular biology tools like cloning are dependent on this capability. In other words, gene flow can occur even without dispersal. However, our ecological and evolutionary studies of microbial populations typically fail to acknowledge the evolutionary impact and genetic contributions of outside populations. Unique evolutionary scenarios arise when dispersal between two or more populations is prevented or limited, but gene flow can still occur between them. We hypothesized that this scenario would impact microbial populations by facilitating speciation, selection, and local adaptation. We aimed to test this hypothesis by studying endemic *Meiothermus* populations inhabiting serpentinite rocks in the subsurface of the Samail ophiolite in Oman. Samail Ophiolite microbial communities, of which *Meiothermus* populations are a component, are dispersed across the subsurface and separated by meters of solid rock and by chemical and pH gradients spanning orders of magnitude. Despite barriers to dispersal that are significant enough to shape community structure, we found that gene flow still occurred between nearly all observed populations of *Meiothermus*. This gene flow is contributing to disruptive selection amongst cohabiting populations, and may also be contributing to local adaptation, both at the genetic and genomic level. We also identified potential mechanisms for this gene flow, including abundant viral elements. The sequence similarity of mobile genetic elements in these *Meiothermus* populations implies that this gene flow occurred after colonization by a common *Meiothermus* ancestor and that diversification is likely ongoing. To our knowledge, this is the first demonstration of gene flow across barriers to dispersal in an environmental microbial system. In conclusion, these results suggest that the capacity for microbial populations to undergo gene flow even in the absence or inhibition of dispersal is a natural process, has substantial consequences for the evolution of the effected population, and may also have consequences for the microbial and surrounding environment.

CHAPTER ONE: INTRODUCTION TO MICROBIAL GENE FLOW AND ITS IMPACT ON EVOLUTION

Gene Flow and Selection

The evolution of a population is impacted by two components: first, the total genomic content present in the population, which is affected by factors like mutation and gene flow, and secondly, how selection affects what genomic content remains to be inherited by subsequent generations. The quantity, content, and organization of the genetic material dictates what is ultimately available to be selected upon. Adjustments in the sequence or organization of genetic material can confer novel phenotypes. Organizational changes can adjust phenotypes through reorganizing multi-gene regulatory structures, moving genes into more highly expressed regions of the chromosome, or by disrupting the structure of genes themselves by moving or inserting into the middle of a gene. Genomic reorganization has empiric consequences for microbial evolution; genomic insertion and deletion events (indels) likely contributed to *Blochmannia chromaiodes* becoming an obligate symbiont of carpenter ants through the disruption of metabolic pathways which in turn led to its reliance on the host (Williams and Wernegreen). Adjustments to genetic sequences can induce changes in the amino acid structure or change regulatory regions. These processes have the potential to confer new phenotypes by adjusting the genomic content already present in the population, however, some of the most significant genomic shifts occur when new genomic content is introduced exogenously, i.e., from gene flow (Maddamsetti).

Gene flow in microbial populations takes two forms. The first is vertical transmission, i.e., the transfer of genes from one generation to the next after asexual reproduction. Vertical

transmission is well studied in microbial systems and many of the paradigms applied to non-microbial evolution via vertical inheritance are also applicable (Neu 2021). The second form of gene flow is through horizontal transmission which will be explored in this dissertation.

Horizontal gene flow, hence referred to just as gene flow, is best conceptualized as horizontal gene transfer from an exogenous source or recently immigrated population (Maddamsetti). Many of our assumptions regarding evolutionary outcomes in microbial communities are population level investigations that assume the only environmental and evolutionarily relevant organisms are those already present in the community. Merino et al 22 (Merino et al. 2022), for example, amplified 16s sequences to reconstruct microbial communities across different sites in Death Valley subsurface aquifers. The inferred community composition was used to infer the degree of connectedness between these sites; the rationale being more sites with similar communities are more likely to be connected. If we consider that microbial communities can be interconnected in ways that cannot be identified by 16s community reconstruction, i.e., movement of exogenous genetic material, or even populations too similar to be distinguished by 16s similarity, then the degree of interconnectedness might be much higher. In fact, the incorporation of exogenous genetic material, i.e., horizontal gene transfer, occurs frequently, has significant effects on population genotypes and phenotypes, and depending on the mechanism, does not require direct contact between populations (Koonin and Wolf 2012).

The aforementioned mechanisms determine what genetic material is present in a population, but selective pressures influence which of those phenotypes will persist. As gene flow impacts selection by dictating the available phenotypes to select against, selection pressures can also inversely impact gene flow. Populations which are well adapted to their specific niche

often experience high levels of purifying selection; there is little room for phenotypic experimentation, as any change will likely pull the organism out of its narrow niche. These organisms may even adopt genetic and genomic adaptations to help prevent this, like more stringent replication error proteins. The resultant impact is that the purifying selection pressure can reduce the impact of gene flow, and in some extreme cases, prevent certain gene flow mechanisms from occurring entirely (Maddamsetti).

Naturally, these two factors; gene flow and selection, are substantially impacted by the environmental factors surrounding the population. We have little understanding as to how environmental barriers to gene flow and microbial dispersal, i.e., movement of microbial populations, affect the evolution of microbial populations. This dissertation will explore the evolutionary consequences of those limitations to gene flow, including how that impacts selection.

Mechanisms of Microbial Gene Flow

Transformation

The ability for exogenous DNA to be incorporated even when the donor population is absent is impactful in that allows for gene flow even in the absence of dispersal. This process can occur in multiple ways: the first is transformation, in which environmental DNA (eDNA) is taken up by a specialized cellular apparatus called a transformasome (Hall et al. 2020). EDNA is estimated to be able to persist in the environment for up to a month in aquatic environments (Barnes et al.). Under more strenuous conditions, eDNA is less stable, but is regardless able to persist long enough to facilitate gene transfer between disconnected populations. DNA is able to move through permeable materials via pore sizes much too small for microbial life

(Muthukumar). While this is frequently exploited in biotechnology, its relevance in environmental systems is not well known. Given the known persistence of eDNA in most environments, and the capacity for DNA to cross barriers that cells cannot, at least in a laboratory setting, it seems reasonable that environmental DNA may cross barriers to dispersal, and consequently complicate microbial evolution. In some systems, up to 80% of the eDNA originates from absent populations. This eDNA is either relict DNA from former, now deceased populations, or DNA from distant populations occupying other habitats that was introduced to the community (Lennon et al.) (**Figure 1**).

Transduction

Transduction is another mechanism by which DNA can cross barriers to microbial dispersal. While transduction in laboratory settings refers to the facilitated transfer of genomic material by many mechanisms like physical or electrically induced competence, in environmental settings and in this work, transduction refers to the conveyance of DNA between bacterial hosts by a viral vector. Evidence for transduction also exists in archaeal genomes, but since archaeal viruses are less well known than their bacteriophage, i.e., bacteria infecting, kin, the mechanisms entailing archaeal transduction are similarly less well known. Viral particles have been identified as a significant component of nearly every microbial ecosystem studied (Weitz and Wilhelm 2012; Pan et al. 2017). They exert substantial selective pressures on their host populations, are responsible for nutrient cycling, and most importantly to this study, act as a mechanism for horizontal gene transfer (Volz, Koelle, and Bedford 2013; Weitz and Wilhelm 2012; Warwick-Dugdale et al. 2019; Koonin and Wolf 2012). Transduction allows for gene flow between two separated populations, provided that viral particles are still able to transit between

them. Viral particles range in size from dozens of nanometers to nearly a micron (Volz, Koelle, and Bedford 2013). Many viral particles can persist in the environment, with some taxon exhibiting the capacity to remain infectious nearly indefinitely under ideal conditions (Legendre et al.). In total, viral particles persist in the environment after their host population disappears and cross barriers their hosts cannot, both of which allow for gene flow to occur between separated populations (**Figure 1**).

Conjugation

Conjugation is a process allowing for the exchange of genetic material between two compatible cells. The process entails the exchange of mobile genetic elements, often plasmids, through a sex pilus connecting two adjacent cells (Virolle C Fau - Goldlust et al.). Conjugation can occur between microbial populations more distant than even the order level (Alderliesten et al. 2020), and although the frequency of successful conjugation decreases with taxonomic distance, there is some evidence that the process may even be possible between bacteria and archaea (Dodsworth et al. 2010). Conjugation can facilitate gene flow across a chemical barrier which separates two populations and prevents dispersal: Two closely related populations may be separated by some chemical barrier in which those cells cannot survive, but by conjugating with some intermediary third microbial population that can survive that barrier, gene flow can occur even when dispersal cannot. This process can occur when the horizontally transferred genes confer phenotypes that benefit all three populations or can occur if the conferred phenotype is neutral or even negative, assuming enough members of the population persist. This complicates our understanding of how barriers to dispersal can affect evolution, providing yet another mechanism by which gene flow can occur when dispersal cannot (**Figure 1**).

Vesiduction

Vesiduction is a process akin to transformation by which the horizontal gene transfer is mediated by an extracellular vesicle containing the exogenous DNA. Typically, the vesicle fuses with the host cell, but the possibility for more complicated integration mechanisms remain (Soler and Forterre). From an evolutionary and ecological perspective, the process acts similarly to transformation or transduction, in that it allows for movement of DNA through barriers that the microbial donors or recipients cannot themselves pass through.

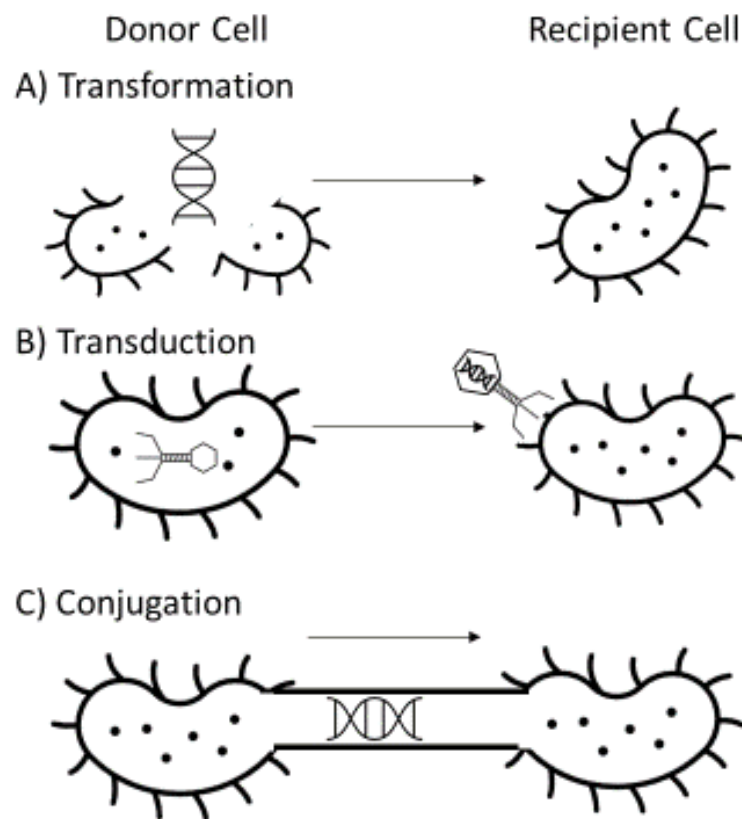


Figure 1. Schematic depicting common mechanisms of horizontal gene transfer. **A)** Schematic depicting transformation, in which exogenous DNA released from a lysed cell is subsequently taken up by a competent recipient cell. **B)** Schematic depicting transduction, in which a viral particle facilitates horizontal gene transfer by infecting the donor cell, incorporating a portion of

the donor cells genome, then subsequently infecting a recipient cell, which then incorporates the portion of the donor genome through recombination or insertion. C) Schematic depicting conjugation, in which competent cells directly transfer DNA through a pilus (Zamani Dahaj 2015).

Relevance of Gene Flow to Microbial Evolution

The horizontal gene transfer (HGT) processes discussed above can facilitate exogenous gene flow and significantly impact evolution (Hall et al. 2020). In fact, HGT events have had significant impact on the evolution of prokaryotes across history (Custer, Bresciani, and Dini-Andreote 2022). These events not only significantly impact the evolution of the recipient organism but have often had substantial consequences on the surrounding environment. For example, HGT events from acetoclastic bacteria to methanogens is proposed to have resulted in the development of acetoclastic methanogenesis (Fournier and Gogarten 2008). This presumably would have expanded the environment and substrate range available to methanogens, reshaping the environments in which methanogens could now inhabit, changing those environments and reshaping the global carbon cycle. Barriers to gene flow also impact evolution. As previously discussed, with unconstrained gene flow, multiple populations inhabiting the same niche will experience genetic homogenization over time, presuming they remain under the same selective pressures. The converse is also true; barriers to gene flow can accelerate evolution even if those barriers are later removed; as is the case with *Laspinema*, a genus of cyanobacteria. In this example, cryptic lineages arose in the past when there was some presumed barrier to gene flow. These lineages persist today as they have evolved to occupy a slightly different niche and avoid introgression into the larger sympatric populations of *Laspinema* (Stanojković et al. 2022).

Impact of the Local Environment on Dispersal and Gene Flow

The environment impacts gene flow in two ways. First, it impacts dispersal, i.e., the physical movement of microbial populations which are then able to exchange genetic material. In the subsurface of Samail Ophiolite for example, microbial populations face environmental barriers in the form of impermeable rock, large differences in pH, and different ion concentrations that can span orders of magnitude (Rempfert et al. 2017). These factors are assumed to strictly impede dispersal; microbial populations are separated by solid rock. In the few areas where there may be a rock fracture or path connecting two populations, dispersing microbes may encounter fluids at a pH or conductivity orders of magnitude different than what they are adapted to. Second, and uniquely to microbial populations, the environment affects the movement of exogenous DNA, whether that DNA is loose nucleic material, encased in a viral particle, or otherwise (Custer, Bresciani, and Dini-Andreote 2022). The myriad barriers to dispersal of microbial life also likely impact the dispersal of exogenous DNA, but the degree and consequences of this are a significant gap in microbial evolution and ecology that this dissertation aims to address.

Microbial Dispersal and its Barriers

The process by which microbial dispersal occurs has not been broadly studied for most environments. Dispersal is a critical factor in both community structuring and population level evolution (Custer, Bresciani, and Dini-Andreote 2022), as the capacity for horizontal gene transfer between distantly related populations (Alderliesten et al. 2020) means dispersal of one population can have significant consequences on the genetic material available for many other

populations. The Baas-Becking hypothesis regarding microbial dispersal is ‘everything is everywhere and the environment selects’, implying that there are no barriers to microbial dispersal, and that community structuring is a selective process only (O'Malley 2008). As the realm of microbial ecology develops, this dogma is being challenged (Burns et al. 2017). Application of phylogenetic analyses indicates that microbial communities are affected by microbial dispersal to different degrees depending on the environment in question.

In tropical reservoirs for example, the degree of dispersal seems strongly correlated with local rainfall, as the increased precipitation allows for dispersal of microbial populations across physically separated habitats. Consequently, during the wet season, community structuring is strongly impacted by dispersal; community turn-over increases in frequency, and communities occupying physically disparate but chemically similar habitats converge on similar community structure. During the dry season, when precipitation is reduced and dispersal occurs less frequently, community heterogeneity increases. Local selection pressures arising from inter-community selection cause each community to shift and become different from its neighbors. Turn-over decreases, and selection causes extant populations to adapt to local characteristics, rather than be excluded by the immigration of a more adapted competitor (Vieira et al. 2022). In a somewhat contrasting situation, subsurface microbial populations appear to have very limited dispersal (Merino et al. 2022). Based on the phylogenetic community structure, as inferred by the proportions of amplified 16s genes, subsurface communities in the Death Valley regional flow system subsurface experience very limited turn-over. Consequently, community structuring in this system seems to be driven largely by inter-site competition, and local adaptation to the conditions of the surrounding micro-environment. As a result, communities can

arrive at different community structures despite having similar environmental characteristics due to the stochastic and selective processes acting on a different, ‘locked-in’, founder community. In total, microbial dispersal is likely as complicated as animal or plant dispersal and is substantially influenced by local environmental characteristics. It is evident that barriers to dispersal differ from environment to environment, and even temporally in the same environment.

At a glance, this would suggest that the ecological assumptions regarding the dispersal of non-microbial communities, whose evolution is substantially impacted by dispersal mechanics, may be applicable for microbial communities. As previously discussed, gene flow without direct contact between populations is much more facile in microbial populations than in animal populations for example. Consequently, understanding the limiting factors to gene flow, rather than solely dispersal, is key to understanding microbial evolution.

However, there is a problem in the difficulty of measuring true dispersal. Measuring dispersal can be difficult when only looking at a single time point, and typically entails measuring the phylogenetic or taxonomic dissimilarity of two communities, and comparing the degree of similarity against a null model in which no dispersal is possible. More concrete methods of evaluating dispersal require multiple timepoints; dispersal is much more noticeable when there are multiple investigated timepoints allowing for the measurement in community change over time. Regardless of the method utilized, it is difficult to say with certainty whether or not dispersal is occurring.

Barriers to Dispersal Versus Barriers to Gene Flow

Critically, in microbial populations, barriers to dispersal are not always barriers to gene flow. As previously discussed, relict eDNA can persist in the environment for months and transit

through structures too solid for microbial dispersal (Lennon et al. ; Hall et al. 2020; Barnes et al.). It is very difficult to determine what constitutes a barrier to microbial gene flow, owing to the multiple mechanisms, e.g., transformation, transduction, or conjugation that can facilitate it. Indeed a barrier capable of blocking transformation by degrading eDNA may not constitute a barrier to transduction facilitated by a hardier enveloped virus (Volz, Koelle, and Bedford 2013). The prevalence of horizontal gene transfer events (Hall et al. 2020) and extensive introgression (Diop et al. 2022) amongst sequenced microbial isolates implies gene flow is substantial across prokaryotes, to the point where the evolutionary dynamics used to model animal populations begin to be applicable to these microbial populations as well (Diop et al. 2022). However, it is near impossible to determine if historic gene flow events occurred across a barrier. Studies of contemporary gene flow events are limited, but there have been documented instances of gene flow occurring between separated populations on a continental scale, as is the case with *Polynucleobacter paneuropaeus*, a pelagic freshwater bacterium (Hoetzing et al. 2021). Similar studies investigating *Sulfolobus* sp. identify gene flow across kilometer scales (Guo et al. 2011). Critically however, it is not known whether these gene flow events were accompanied by the dispersal of the organism or occurred in its absence. As such, it is broadly unknown what constitutes a true barrier to gene flow.

The Role of Gene Flow on Microbial Evolution

The microbial specific capacity for gene flow to occur in the absence of dispersal is important for speciation concepts. Critically, barriers to gene flow may be necessary for speciation to occur, as posited in Mallet et al 2008 (Mallet 2008). Without barriers to gene flow (**Figure 2**), niche level homogenization occurs by which microbes occupying similar guilds, i.e.,

under the same selective pressures, and capable of horizontal gene flow will converge on a single homogenized species over time as they essentially have access to the same gene pool from which the optimal alleles are selected from. This is inevitable over long enough timescales, assuming selective pressures remain the same, unless there is some barrier to gene flow. Thus, barriers to gene flow constitute catalysts to speciation in that they slow the speed at which genomic homogenization can occur.

Classical speciation designations are questionably applicable to microbial populations. Allopatric speciation is the term used to describe the speciation that occurs when a population becomes physically split, and the two resultant populations are physically separated, and consequently begin to speciate due to the differential selection pressures in their new respective environments. This is distinct from sympatric speciation, in which a population begins to diverge into two new populations whilst maintaining physical contact, oftentimes a result of behavioral differences, reproductive barriers, or specialization into concurrent micro-niches. These definitions are useful when barriers to dispersal are also barriers to gene flow. However, in microbial populations, two physically separated populations may still undergo gene flow by the movement of exogenous DNA. In animal populations, parapatric speciation typically refers to two physically separated populations that are able to maintain some degree of gene flow through the rare interactions between members, often during migration. In Munro-Ehrlich et al 2023 (Munro-Ehrlich et al.), we adapted this term to refer to speciation occurring between physically separated populations that still experience gene flow, with the key difference that dispersal, or physical contact, between populations is not required for this process to occur. ***However, it is as yet unknown how frequently this mode of speciation occurs in microbial populations.*** It is

possible that all microbial speciation is parapatric in some form, due to the ready movement and exchange of exogenous DNA between populations (Custer, Bresciani, and Dini-Andreote 2022).

Moreover, the evolutionary impacts of gene flow may be falsely attributed to dispersal, since we still do not have a clear idea of what constitutes a barrier to either process. For example, subsurface populations like those in the Death Valley subsurface reservoir (Merino et al. 2022) seem to be locked in, i.e., unable to disperse, implying that the solid rock surrounding these communities constitutes a true barrier to dispersal. The evolutionary facets of these communities might be erroneously attributed to the lack of dispersal; evolution of that community would be assumed to be driven by solely endogenous mutation, or founder effects, when in reality, significant gene flow could be occurring, introducing vast quantities of novel genetic material, and substantially impacting the evolution of the community. In total, our understanding of microbial evolution is hindered by our inability to identify barriers to dispersal, barriers to gene flow, and assuming that they are one and the same.

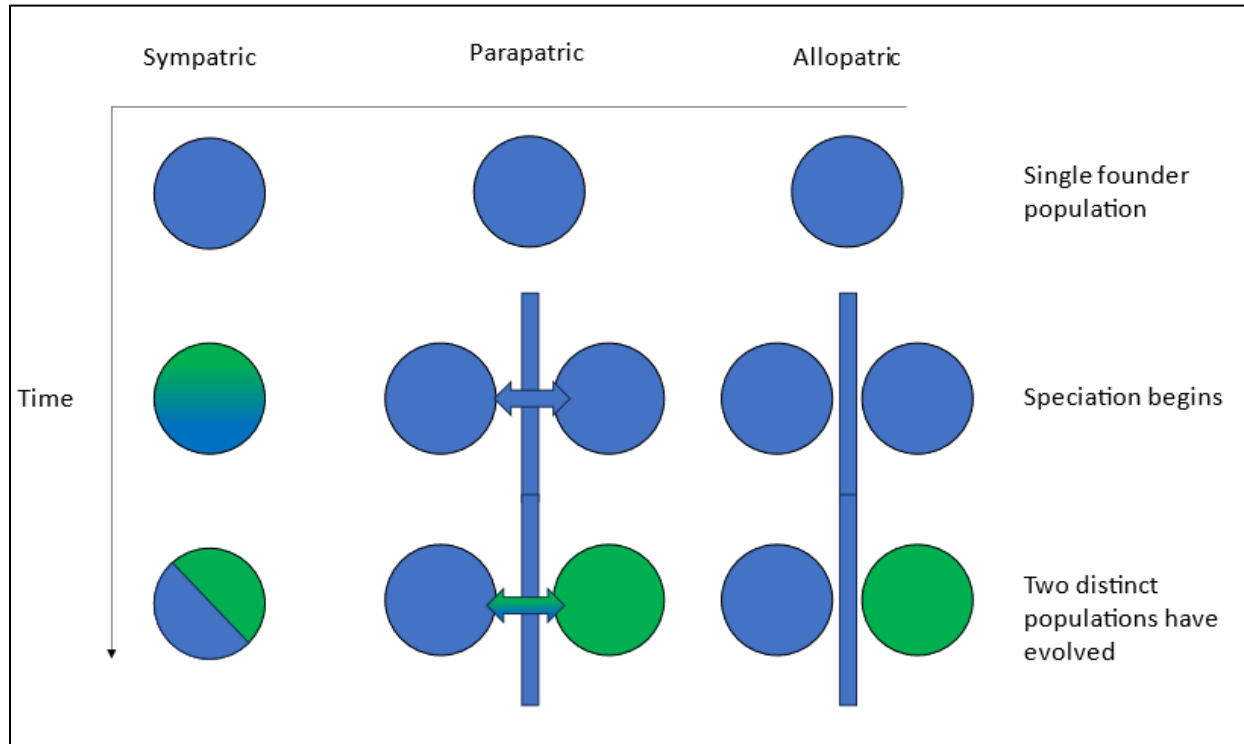


Figure 2. Schematics depicting microbial speciation regimes. Columns indicate different speciation modes, ordered from left to right broadly based on the degree of gene flow and / or physical separation between populations during speciation, with the left most mode pertaining to the least amount of gene flow and / or physical separations. Rows depict the temporal process of speciation, initiated at the top and proceeding to completion at the bottom row. Left: Sympatric speciation, in which a single founder population splits into two descendent populations, both of which inhabit the same area and are in frequent contact. Middle: Parapatric speciation, in which one founder population is geographically split in two. Some degree of gene flow, as indicated by an arrow, either through dispersing microbes or the movement of exogenous DNA persists across the physical barrier connecting the populations even as they diverge. Right: Allopatric speciation, in which a single founder species is split in two by a barrier preventing any sort of gene flow or dispersal. Reproductive isolation and differential selection pressures culminate in the founder population splitting into two descendant populations.

The Combined Impacts of Dispersal and Gene Flow on Microbial Evolution

It has become apparent that dispersal can substantially affect microbial community structure and consequently, microbial evolution. The degree to which dispersal is a relevant force in shaping microbial evolution depends on the environment, and often, the temporally variable conditions surrounding that environment. Moreover, gene flow has obvious and immediate impacts on microbial evolution. The problem lies in assuming that these two factors are inextricably linked. In many circumstances, like the one this thesis will discuss, the movement of microbial life, dispersal, and the movement of DNA, gene flow, is affected by different factors. When this disjunction occurs, unique evolutionary scenarios ensue. There are a few specific gaps that should be addressed to provide the groundwork to further investigate the role of dispersal and gene flow in microbial evolution, particularly in the scenarios in which the two processes are disconnected.

Of specific interest are systems where dispersal is limited but gene flow is still possible, for example, subsurface systems in which microbial communities are separated by meters of solid rock which may inhibit microbial dispersal but allow for movement of fluids and smaller biological components like exogenous DNA or viral particles. These systems are uniquely important because they constitute a population genetic structure that only exists in microbial communities, and therefore, studying them will grant us a unique understanding of microbial evolution that is distinct from the evolution of non-microbial communities. The final gap to address is how do microbes evolve in these unique circumstances; what is the quantity of gene flow in the presence of barriers to microbial dispersal, what is the content of that gene flow,

which mechanisms of gene flow, e.g., transduction, transformation, are present and relevant, and finally, what consequences does this gene flow have on selection and local adaptation.

The work presented in this dissertation seeks to answer the above questions and evaluate how barriers to dispersal and gene flow, both physical and chemical, impact microbial evolution. Previous studies have focused on how barriers to dispersal may impact overall community structuring (Merino et al. 2022; Vieira et al. 2022), but have not addressed how individual populations separated by those barriers may evolve. Additionally, many of these studies inferred community composition by amplifying the 16s small ribosomal subunit, then inferring community composition based on the ratios of different retrieved amplicons. This approach precludes any sort of analysis of gene flow and reduces the fidelity of community analysis to the genus or species level. The studies that have sequenced the entirety of the metagenome did not investigate the potential connectedness of the investigated genomes. Briefly, most investigations of microbial community structure have focused on the distribution of microbial populations without investigating the potential interconnectedness of those populations.

Our understanding of how microbial communities evolve has consequences outside of microbiology. For example, the rapid dissemination of antibiotic resistance is of immediate pertinence. Antibiotic resistance is frequently spread through horizontal gene transfer (Hu, Gao, and Zhu 2017). While the biological mechanisms facilitating this HGT, e.g., transformation, are well understood, the evolutionary dynamics that dictate when this HGT is possible, as well as how likely transferred genes are likely to remain incorporated is less well known and is contingent on our understanding of microbial gene flow and evolution. The impacts of microbial evolution extend beyond medical settings. In Merino et al 2022 (Merino et al. 2022), the

evidence for limited microbial dispersal is used as evidence that there is minimal fluid exchange between subsurface groundwater sites in the Death Valley basin, and it is proposed that the degree of connectedness between these microbial communities could even be used as a proxy measure for fluid residence times.

Critically however, this study neglects to investigate the degree of non-dispersal related gene flow, i.e., the quantity of DNA exchanged between these communities without a bacterial or archaeal vector. It is possible that they might find significant exchange of exogenous DNA between these communities, which would indicate that there is indeed significant fluid exchange between these sites, albeit through pores too small for most microbes to transit. Paukert et al 2019 (Paukert Vankeuren et al. 2019) attempted to constrain groundwater residence times in the Samail Ophiolite, in order to better quantify rates of CO₂ sequestration via mineralization. An investigation into gene flow and dispersal between the microbial communities of this system would consequently provide additional perspective on estimates of groundwater residence times. The high quantity of contemporary gene flow indicates a connectedness of these communities, and thus presumably a connectedness of the surrounding fluids as well.

To address these gaps in our understanding of microbial evolution, Chapter 2 of this dissertation investigates the quantity and content of gene flow between spatially and chemically segregated microbial communities, and subsequently investigates the consequences of this gene flow on selection and speciation. To do this, three microbial communities from a single subsurface sampling site in the Samail Ophiolite were investigated. The Samail Ophiolite is the world's largest terrestrial serpentinite and is one of the most frequented sites for studying how serpentinization supports and impacts microbial communities (Rempfert et al. 2017). In this

system, serpentinization reactions, i.e., the hydration of rock bearing high proportions of magnesium and iron bearing minerals, occurring in the subsurface support expansive microbial communities through the generation of H_2 (Russell, Hall, and Martin 2010b). Fractures in the rock and the fluid intercalating through those fractures create a network of fluid filled spaces which support diverse microbial communities (Rempfert et al. 2017). The rocks surrounding these communities have varying porosity, but fluid residence times tend to be long (Paukert Vankeuren et al. 2019), indicating that while there are many fluid filled spaces for microbial communities to inhabit, these spaces are segregated and mixing and fluid refill occurs on slow timeframes.

While the Samail Ophiolite is frequently used for microbial community studies because of the wealth of empiric data and its applicability to extraterrestrial habitats, it is also the ideal study site for an investigation into microbial dispersal and gene flow. This is because the area hosts many microbial communities, some of which inhabit similar local environments, and are distributed across the subsurface and separated by meters to kilometers of solid rock. Samail ophiolite microbial communities are expected to have limited dispersal for multiple reasons. First, there are substantial physical barriers between communities in the form of solid, limited porosity rock. There are also chemical barriers as well; adjacent fluids hosting microbial communities may differ from each other in pH, the concentration of dissolved ions, and the conductivity. As serpentinization progresses, the pH of the surrounding fluids increases as well, through mechanisms including the equilibration of hydroxide ions with the cations like Ca^{2+} which are liberated from the surrounding rock (Russell, Hall, and Martin 2010a). The pH of fluids in the ophiolite span from circumneutral to 12 (Rempfert et al. 2017). Conductivity also

shifts significantly, which is the result of large shifts in ion concentrations, including Ca^{2+} , across different fluids. These chemical gradients likely also constitute a barrier to dispersal, as microbial communities inhabiting the subsurface may be surrounded by fluids with significantly different pH, conductivity, or ion concentration than their realized niche. Because fluid residence times are long (Paukert Vankeuren et al. 2019), implying limited fluid movement between microbial communities, it is expected that dispersal of microbial life is limited. Previous investigations including Rempfert et al 17 (Rempfert et al. 2017) hint at dispersal limitations; Different sample sites host very different microbial communities, even when the local environmental conditions are similar. As expected in a system with limited dispersal, different communities exhibit different characteristics despite having similar environmental conditions (Rempfert et al. 2017). This community heterogeneity is characteristic of meta-populations where dispersal is limited in some way (Rempfert et al. 2017; Munro-Ehrlich et al.). Importantly, contemporary investigations have revealed the presence of cells inside the solid rock (Templeton et al. 2021). These endolithic microbial communities have demonstrated some activity, in the form of biotically mediated transformation of carbon compounds (Fones et al. 2022). However, it is yet unknown whether those endolithic communities are able to move through the rock to engage in dispersal, or if they are trapped. Regardless, it is apparent that the subsurface of the Samail Ophiolite constitutes a system harboring multiple microbial communities in which dispersal is expected to be limited, but recent developments like the identification of cells in rock interiors challenge this assessment.

In 2017, the Oman drilling project drilled an additional 6 wells. One of these wells, BA1A, was fitted with a packer system (Lods, Roubinet, Matter, Leprovost, Gouze, et al. 2020)

(Figure 3). This system effectively partitions off four different depth intervals: 0 – 30 meters, 41 – 65 m, 108 -132 m, and 132 – 400m. After packing off these intervals, fluids were completely drained, and then allowed to recharge. After fluid recharge, the top three depth intervals, 0 – 30, 41 – 65, and 108 – 132 m were pumped and filtered to retrieve biomass. Fluids were completely drained and then recharged prior to sampling to ensure that each sample was comprised solely of fluids from fractures of that partitioned depth interval, rather than a mixture of multiple intervals. This allowed for a depth dependent investigation of microbial communities in the region; previous sampling methods had required some degree of fluid mixing which would presumably also mix the microbial communities from different depths. Fractures in these three intervals hosted microbial communities which were not only spatially segregated, but also likely chemically segregated; the pH of the top two intervals was 8.1 and 8.2, whereas the pH of the fluids from the bottom interval was 10.6. Using the metagenomes assembled from DNA retrieved from these samples, we aimed to determine how these physical and chemical barriers to gene flow and dispersal impact gene flow and selection in these communities.

Rather than investigate gene flow in every community member distributed across BA1A, we instead opted to investigate gene flow in *Meiothermus*. *Meiothermus* is an abundantly distributed aerobic heterotroph, commonly found in alkaline hot springs, but is also readily distributed throughout sample sites in the Samail Ophiolite (Rempfert et al. 2017). There are two clades of *Meiothermus* present in Samail Ophiolite microbial communities, both of which are endemic to the ophiolite. There are populations of both clades in each depth interval in BA1A, and a high quality metagenome assembled genome (MAG) was obtained for each of these six populations. This unique population structure, combined with high quality sequence data, makes

Meiothermus the ideal target organism to investigate gene flow in subsurface microbial communities; the population structure allows for investigations into gene flow within a single community, between communities, between closely related populations, and between distantly related populations. As such, our findings into gene flow in the system are based off of *Meiothermus* but are expected to be applicable to other organisms within the system.

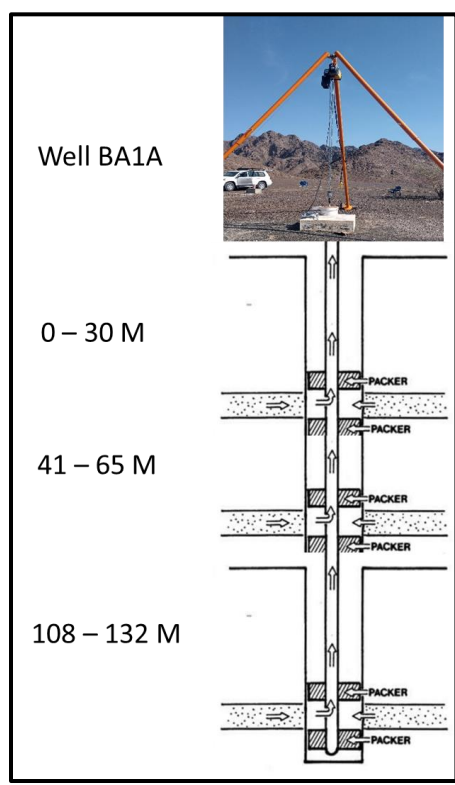


Figure 3. Schematic depicting packer system employed in well BA1A. The schematic indicates the three depth intervals in which sampling takes place partitioned by the packer system. The sampling mechanism at the opening of the well is from a photograph taken during sampling season in 2019.

Hypothesis

Chapter 2 of this thesis tests the hypothesis that **gene flow should accelerate evolution and speciation in dispersal limited communities**. For this hypothesis to be true, four subsequent hypothesis must be confirmed. First, there must be some degree of dispersal limitation between microbial populations inhabiting different depths in BA1A. We first tested the hypothesis that the physical barriers, i.e., solid rock with limited porosity, and chemical barriers, i.e., shifts in pH and conductivity, would significantly restrict microbial dispersal. If true, this should create distinct microbial communities in which the community structure reflects the environmental challenges inherent to each sample site. To evaluate if barriers to dispersal impacted microbial community structure, we assembled metagenomes from the three depth intervals in BA1A and compared their community composition.

The following hypothesis that these separated populations must be experiencing gene flow, in spite of the barriers to dispersal. We then hypothesized that gene flow would still be possible despite the physical and chemical barriers via the movement of exogenous DNA, either through viral particles, environmental DNA or some other mechanism. This would allow for the exchange of genetic material between closely related, but physically separated populations. To evaluate the quantity, mechanisms, and content of gene flow between these separated populations, we specifically investigated six *Meiothermus* populations. *Meiothermus*, while characterized as an aerobe, is present in all three depth intervals in BA1A. Specially, there are two endemic clades of *Meiothermus* in BA1A, which share around 80% average nucleotide identity (ANI) and have some abundance, as well as high quality genomes, in each of the three sample sites, making them ideal for a population evolution study. We developed a single

nucleotide variant (SNV) dependent approach to determine the quantity and content of gene flow between populations either inhabiting the same space, or in different sample sites.

Thirdly, we hypothesized that there should also be some evidence indicating the presence of horizontal gene flow mechanisms, e.g., evidence for potential transduction in the form of recovered viral sequences. We investigated this by identifying mobile genetic elements (MGEs) and viral sequences. These sequences were evaluated to determine if they could mechanistically facilitate horizontal gene flow. In the case of the MGEs, that was done by evaluating whether or not those MGEs were still functional, and in the case of viral sequences, by evaluating whether these sequences constituted a relictual viral sequence or part of an extant virus. We also evaluated the phylogenetic range of viral sequences, as well as their presumed host range to determine if they would constitute an ecologically active part of the BA1A community. By comparing the sequence similarity of MGEs to their encoding population we were also able to determine if these MGEs represented contemporary gene flow events.

We finally hypothesized that if gene flow was occurring even in the absence of microbial dispersal, then evolution and speciation would be impacted; rates of evolution should increase since the import of exogenous DNA increases the genetic material available to select against, in sense increasing the functional population size of each group. We expect this accelerated evolution to be visible through disruptive selection between closely related populations that, while able to exchange genetic material, are now able to diversify to fit different niches, resulting in parapatric speciation. To test this, we looked at the patterns of gene flow, as well as the contents, to infer potential selection pressures affecting these populations. We then aimed to identify mechanisms of gene flow by looking at the mobile genetic elements present in our

metagenomes. Finally, we identified potential genomic and genetic adaptations to the different environmental conditions of the three sample sites to determine if gene flow may be contributing to local adaptation and potentially speciation. Without the import of exogenous DNA, this process should be significantly slower, and perhaps even reversed; genetic pressure from the larger population would even cause the introgression of the smaller population, who without the import of novel DNA, would not be able to evolve fast enough to diversify into a distinct niche.

Chapter 3 discusses the conclusions and outside implications of this work; the influence of gene flow in a dispersal limited microbial environment not only revises our understanding of microbial community structuring, evolution, and ecology, but also how microbial communities impact their environment. In the Samail ophiolite specifically, this potential greater interconnectedness between communities implies a potentially faster recharge rate and greater fluid connectivity in the subsurface there.

CHAPTER TWO

PARAPATRIC SPECIATION OF *MEIOTHERMUS* IN SERPENTINITE-HOSTED AQUIFERS
IN OMAN

Contribution of Authors and Co-Authors

Author: Mason Munro-Ehrlich

Contributions: performed bioinformatics analyses, wrote manuscript

Author: Daniel Nothaft

Contributions: designed experiment, wrote sections

Co-Author: Elizabeth Fones Contributions: designed experiment

Co-Author: Jeurg Matter

Contributions: designed experiment

Co-Author: Alexis Templeton

Contributions: designed experiment

Co-Author: Eric Boyd

Contributions: wrote, edited, and prepared manuscript

Manuscript Information

Mason Munro-Ehrlich, Daniel B. Nothaft, Elizabeth M. Fones, Juerg M. Matter, Alexis S.
Templeton, and Eric S. Boyd

Frontiers in Microbiology

Status of Manuscript:

☐ Prepared for submission to a peer-reviewed journal

☐ Officially submitted to a peer-reviewed journal

☐ Accepted by a peer-reviewed journal

☒ Published in a peer-reviewed journal

Frontiers

Submitted 5 Jan 2023

Published 13 April 2023

Volume 14 - 2023

<https://10.3389/fmicb.2023.1138656>

Abstract

The factors that control the distribution and evolution of microbial life in subsurface environments remain enigmatic due to challenges associated with sampling fluids from discrete depth intervals via boreholes while avoiding mixing of fluids. Here, using an inflatable packer system, fracture waters were isolated and collected from three discrete depth intervals spanning >130 m in a borehole intersecting an ultramafic rock formation undergoing serpentinization in the Samail Ophiolite, Sultanate of Oman. Near surface aquifer waters were moderately reducing and had alkaline pH while deeper aquifer waters were reduced and had hyperalkaline pH, indicating extensive influence by serpentinization. Metagenomic sequencing and analysis of DNA from filtered biomass collected from discrete depth intervals revealed an abundance of aerobes in near surface waters and a greater proportion of anaerobes at depth. Yet the abundance of the putatively obligate aerobe, *Meiothermus*, increased with depth, providing an opportunity to evaluate the influence of chemical and spatial variation on its distribution and speciation. Two clades of *Meiothermus* metagenome assembled genomes (MAGs) were identified that correspond to surface and deep populations termed Types I (S) and II (D), respectively; both clades comprised an apparently Oman-specific lineage indicating a common ancestor. Type II (D) clade MAGs encoded fewer genes and were undergoing slower genome replication as inferred from read mapping. Further, single nucleotide variants (SNVs) and mobile genetic elements identified among MAGs revealed detectable, albeit limited, evidence for gene flow/recombination between spatially segregated Type I (S) and Type II (D) populations. Together, these observations indicate that chemical variation generated by serpentinization, combined with physical barriers that reduce/limit dispersal and gene flow, allowed for the parapatric speciation of *Meiothermus* in the

Samail Ophiolite or a geologic precursor. Further, *Meiothermus* genomic data suggest that deep and shallow aquifer fluids in the Samail Ophiolite may mix over shorter time scales than has been previously estimated from geochemical data.

Introduction

The terrestrial subsurface is host to an abundant and active microbial biosphere (McMahon and Parnell, 2014; Bar-On et al., 2018; Magnabosco et al., 2018) that comprises populations of cells inhabiting the pore spaces and (micro)fractures of rocks (Goordial et al., 2021; Fones et al., 2022). A large amount of the habitable subsurface is comprised of mafic and ultramafic igneous rocks. Mafic and ultramafic igneous rocks can undergo the geologic process of serpentinization that can generate H₂ through the oxidation of Fe(II)-containing minerals, such as olivine and brucite (Miller et al., 2017; Ellison et al., 2021). H₂, in turn, can react with CO₂ to generate formate (HCOO⁻), carbon monoxide (CO), or methane (CH₄) (Seewald et al., 2006; McCollom and Seewald, 2007) that can be used as electron donors and/or carbon sources in microbial metabolism.

Serpentinization also generates highly reducing conditions that can limit the availability of oxidants needed to support microbial metabolism. For example, fracture fluids from rock formations undergoing serpentinization can exhibit oxidation- reduction potentials (Eh) as low as -750 mV (Templeton et al., 2021). Available oxidants in waters infiltrating these environments are consumed either abiotically via reactions involving minerals or soluble reductants (e.g., ferrous iron or hydrogen) or biotically by microorganisms. Low oxygen fugacity and oxidant limitation is characteristic of fluids collected from environments undergoing serpentinization (Frost and Beard, 2007; Rempfert et al., 2017; Suzuki et al., 2017; Templeton et al., 2021). In

these systems, dissolved O₂ concentrations rarely exceed 40 µM even in the least reacted waters and are below detection in highly reacted waters (Rempfert et al., 2017; Suzuki et al., 2017; Piccoli et al., 2019). Yet, despite the apparent absence of O₂ in reacted waters, putatively obligate aerobes are often detected. For example, sequences related to the obligately aerobic genus of bacteria, *Meiothermus* (Deinococcota), have been detected in subsurface hyperalkaline and suboxic environments, oftentimes in high relative abundance (Chung et al., 1997; López-López et al., 2015; Woycheese et al., 2015). Likewise, fracture waters collected from a variety of depths in numerous wells have been shown to host *Meiothermus* DNA, including circumneutral near-surface waters and the most hyperalkaline and most highly reacted waters in the Samail Ophiolite, Sultanate of Oman (Rempfert et al., 2017; Nothaft et al., 2021).

Multiple reports of the presence of putative aerobes (e.g., *Meiothermus*) in what appear to be otherwise anoxic serpentinite fluids suggest the possibility that aerobes are widespread, can freely disperse among fractures in rocks, and are inactive in reduced portions of the serpentinites only to become active when favorable conditions are encountered. Depending on the time scales over which this occurs, the extent of dispersal could either allow for or fully restrict gene flow/recombination among otherwise isolated populations, allowing for localized speciation among closely related populations. Several recent studies offer differing perspectives on whether dispersal and/or gene flow/recombination are possible in subsurface habitats. A metagenomic analysis of low biomass subseafloor sediment communities in North Pond in the Atlantic Ocean sampled repeatedly ($n = 10$) over a period of nearly two years showed strain level shifts in the composition of populations comprising those communities (Anderson et al., 2022). Similarly, an analysis of 16S rRNA genes recovered from fluids in three boreholes intersecting a 1,400 m deep

hard rock aquifer sampled over a 10 month period revealed substantial changes in community compositions (Zhang et al., 2022). In both studies, the results were interpreted to reflect dispersal of cells through sediment porewaters or fracture networks. In contrast, a recent study of 16S rRNA genes collected from subsurface fluids in the Cedars, a hard rock serpentinite formation located in California, USA, over multiple years concluded that spatial barriers limit dispersal (Putman et al., 2021). Together, these observations raise questions as to the predominant mode of speciation that drives microbial evolution in serpentinite formations, as well as other hard rock subsurface environments.

In 2017, the Oman Drilling Project drilled six new wells to establish a multi-borehole observatory of the serpentinitizing subsurface in the Samail Ophiolite. In 2018 and again in 2019, a submersible down borehole packer and pump system was installed in several of the boreholes, including BA1A, a 400 m borehole that primarily intersects dunite nearer to the surface and harzburgite at depth (Lods et al., 2020; Malvoisin et al., 2020). The packer system uses inflatable nitrogen balloons to seal off permeable zones in the bedrock and to identify the primary fractures that allow for fluid flow. In BA1A, these were identified as 0–30 m, 41–65 m, and 108–132 m, with limited connectivity to aquifers below this depth (Lods et al., 2020). By sealing off these discrete intervals, fluids could be sampled while limiting the possibility for mixing of subsurface waters during sampling within the well. Previously reported chemical analyses reveal aquifer waters from the top two depth intervals in BA1A are circumneutral and moderately reducing whereas aquifer waters from the deeper interval are hyperalkaline and highly reducing, consistent with successful segregation of fluid types by the packer system (Nothaft et al., 2021). Furthermore, 16S rRNA gene sequencing of biomass collected from each of the three discrete

depth intervals revealed shifts in the composition of the communities from largely aerobic taxa nearer to the surface toward anaerobic, putatively sulfate reducing taxa at depth (Nothhaft et al., 2021). Notably, *Meiothermus* sequences were detected at all three sampled depths, but had an order of magnitude higher relative abundance in the 108–132 m interval. This suggests that the presence of *Meiothermus* is unlikely to be due solely to mixing of fluids during sampling and raises the question of how *Meiothermus* withstand the polyextremophilic conditions associated with heavily serpentinized deep subsurface fluids, including an apparent lack of O₂.

Here, we apply metagenomic sequencing to DNA extracted from biomass from each of the three isolated depth intervals (0–30 m, 41–65 m, and 108–132 m) in BA1A to further evaluate controls on the distribution and evolution of microbial life in subsurface environments undergoing serpentinization, with a specific focus on *Meiothermus*. Sequences were assembled, binned into metagenome assembled genomes (MAGs), and *Meiothermus*-affiliated MAGs were identified and curated. *Meiothermus* MAGs were then compared bioinformatically, phylogenetically, and metabolically to identify similarities and differences as a function of depth and extent of serpentinization in the fluids they inhabited. Estimated genome replication rates were calculated for *Meiothermus* MAGs to establish whether those populations are likely in an active state of replication. Single nucleotide variants (SNVs) were identified to generate high fidelity *Meiothermus* population structures to evaluate patterns in the diversification and gene flow/recombination among depth-resolved populations. Collectively, the results indicate that chemical and physical barriers that limit co-habitation and gene flow/recombination, in combination with gene loss and gain via mobile genetic elements, drove parapatric speciation of *Meiothermus* in the Samail Ophiolite. While gene flow/recombination among depth-resolved

populations was limited, it was detected, suggesting that the reduced, hyperalkaline subsurface aquifer waters in the Samail Ophiolite may not be as isolated from the surface aquifer as has been previously estimated based on bulk characterization of the chemistry of fluids [e.g., deep fluids isolated for >20,000 years (Paukert Vankeuren et al., 2019)]. The results are discussed in terms of processes controlling the distribution and evolution of microbial populations in subsurface environments undergoing serpentinization.

Materials and Methods

Site Description, Drilling, and Sampling

BA1A is a part of a multi-borehole observatory established by the Oman Drilling Project. Details on drilling are reported previously (Lods et al., 2020; Kelemen et al., 2021; Nothaft et al., 2021) and are further expanded upon in the Supplementary methods. Briefly, BA1A is a 400 m deep, 0.152 m diameter well, intersecting fully serpentinized dunite in the upper 250 m and partially serpentinized harzburgite at greater depth (Lods et al., 2020; Kelemen et al., 2021). A Solexperts packer system (Zurich, Switzerland) that included two inflatable bladders (“packers”) and a Grundfos (Bjerringbro, Denmark) model SQE 1–140 submersible pump was installed in BA1A in February 2019, enabling the isolation of discrete depth intervals for hydrological testing (Lods et al., 2020) and microbiological and chemical characterization of groundwaters (Nothaft et al., 2021). Samples of planktonic biomass were collected from discrete depth intervals in February 2019 using the packer system, as described previously (Nothaft et al., 2021). Controls for contamination, both in the form of remnant drilling fluid as well as laboratory contamination and DNA extraction controls, are described in previous studies (Nothaft et al., 2021; Templeton et al., 2021).

DNA Extraction and Shotgun Metagenomic Sequencing

Genomic DNA was extracted from filtered biomass with the Qiagen PowerSoil kit (Germantown, MD, USA) and was submitted to the University of Wisconsin Biotechnology Center for library preparation following the Illumina (San Diego, CA, USA) regular fragment (~300 bp) kit and these libraries were shotgun sequenced via the Illumina NovaSeq 6000 (2 × 150 bp) platform. Information on the depth and quality of sequences obtained from the three libraries are reported in Supplementary Table 1. Additional details on the DNA extraction can be found in Nothhaft et al. (2021) and in the Supplementary methods.

Metagenomic Sequence Assembly, Binning, and Metagenome Assembled Genome (MAG) Metabolic Predictions

DNA was shotgun sequenced, curated, assembled, and binned into MAGs using the same pipeline outlined previously (Fernandes-Martins et al., 2021). Further details regarding assembly and binning can be found in the Supplementary methods. MAGs are available from National Center for Biotechnology Information (NCBI) under the BioProject identification number PRJNA918706. The taxonomic affiliations, relative abundances, and completeness/quality of MAGs in each of the three communities are reported in Supplementary Table 2.

Metagenome assembled genomes were characterized as corresponding to putatively aerobic or anaerobic cells using the Basic Local Alignment Search Tool (BLASTp) to query MAGs first for homologs of cytochrome c oxidase [Cox I and II; Enzyme Category (EC) 7.1.1.9] and cytochrome bd complex (CydABX; EC 7.1.1.7), two proteins necessary for aerobic metabolism. Following this, MAGs were uploaded to the Kyoto Encyclopedia of Genes and Genomes (KEGG) server for annotation and pathway prediction. Specifically, orthologs

predicted by KEGG to be involved in other (i.e., those not involving O₂) aspects of putative electron transport chains were examined, specifically orthologs of terminal oxidases (e.g., dissimilatory nitrate reductase, dissimilatory bisulfite reductase). The potential for a MAG to correspond to an aerobe or an anaerobe was then cross checked against the metabolism of its closest cultivated relative, as assessed using the Genome Taxonomy Database-Toolkit (GTDB-Tk) (Chaumeil et al., 2019). Further details regarding this approach can be found in the Supplementary methods. We also evaluated the potential for alternative O₂-producing biochemical mechanisms in BA1A populations. MAGs were examined for genes encoding enzymes known to produce O₂, including nitric oxide dismutase (Nod; Zhu et al., 2019), superoxide dismutase (SOD; Imlay, 2002), chlorite dismutase (Cld; Hofbauer et al., 2014), and peroxidases/catalases (Cat; Singh et al., 2008) using BLASTp, with characterized proteins as query sequences (Supplementary Table 3).

Compilation of *Meiothermus* Genomes and MAGs

Metagenome assembled genomes were assigned taxonomy using the GTDB-Tk (Chaumeil et al., 2019). Using this approach, MAGs that showed close affiliation to *Meiothermus* were compiled from BA1A metagenomes, as well as fracture fluid metagenomes from other subsurface locations in the Samail Ophiolite (Fones et al., 2019, 2021). Taxonomic assignments were verified manually via BLASTp analysis of housekeeping genes (e.g., RNA polymerase) against the NCBI non-redundant database.

The size of *Meiothermus* MAGs was estimated by normalizing MAG size to estimated completeness, where completeness was determined as the proportion of housekeeping genes present as determined by Metawrap (version 1.3) (Uritskiy et al., 2018). The housekeeping genes

assumed to be present in a complete genome were expected to be the same as those present in *Meiothermus hypogaeus* the closest relative (~85% ANI) to Oman *Meiothermus* strains. Only MAGs that contained >60 full length housekeeping genes were retained for downstream analysis. This included six *Meiothermus* MAGs from BA1A and twelve *Meiothermus* MAGs from our database of Oman fracture fluid metagenomes (Fones et al., 2019, 2021).

Phylogenomic Analyses of *Meiothermus*-Affiliated MAGs

The database of Oman *Meiothermus* MAGs was subjected to alignment of housekeeping genes using the GTDB-Tk (Chaumeil et al., 2019). Also included in the alignments were housekeeping genes from eight *Meiothermus* lineages not from Oman (*Meiothermus taiwanensis* GCA_000482765.1, *Meiothermus ruber* GCA_015478585.1, *Meiothermus silvanus* GCF_000092125.1, *Meiothermus rufus* GCF_00042325.1, *Meiothermus cerbereus* GCF_000620065.1, *Meiothermus* sp. Pnk-1 GCF_003226535.1, *Meiothermus* sp. QL-1 GCF_003351145.1, and *Meiothermus hypogaeus* GCF_003574035.1) to be used as an outgroup. The alignment was generated using Clustal Omega (version 1.2.4) (Sievers et al., 2011), and IQtree (version 1.6.12) (Minh et al., 2020) was used to generate the tree specifying the LG substitution model and 1,000 bootstraps. The tree was visualized using the Interactive Tree of Life (iTOL) web platform (version 6) (Letunic and Bork, 2021), which identified two clades of *Meiothermus* that are herein referred to as “Type I surface or (S)” and “Type II deep or (D).”

Additionally, since housekeeping genes could not be retrieved for many other *Meiothermus* MAGs or metagenomic assemblies, RpoB sequences from MAGs closely related to *Meiothermus* and from unbinned metagenomic sequence data from other environmental samples were retrieved (Supplementary Table 4), including those from the NCBI non-redundant (NR)

database and the sequence read archive (SRA) database. RpoB from *Thermus thermophilus*, a member of the sister genus to *Meiothermus* (Henne et al., 2004), was used to root the phylogeny. The tree was generated and visualized as described above.

Estimation of *Meiothermus* Genome Replication Rates

The Strain level Metagenomic Estimation of Growth (SMEG) program (ver. 1.1.1) was used to infer genome replication rates for MAGs that showed closest affiliation to *Meiothermus* (Emiola et al., 2020). Mauve (ver. 2.4) (Darling et al., 2004) was used to reorder contigs of *Meiothermus*-affiliated MAGs through alignment to the genome of *M. hypogaeus*. The quality of contig realignment was examined by evaluating the depth of mapped raw reads to ensure that they were generally reordered from highest read depth (origin of replication) to lowest read depth (terminus of replication). SMEG then aligned these reordered contigs to determine single nucleotide variants (SNVs) and assigned MAGs into phylogenetic subclusters based on SNV sites. Read coverage for each contig within each MAG at each SNV site was then determined, and the resultant read map coverage ratios were used to estimate genome replication rates.

Assignment and Characterization of *Meiothermus* Protein Clusters

Protein clustering was used to identify protein encoding genes unique to each *Meiothermus* MAG and to facilitate downstream evaluation of SNV profiles. All called proteins for each MAG were subjected to protein clustering using CD-HIT (ver. 4.8.1) (Fu et al., 2012) resulting in clusters of homologous proteins. Additional details regarding the protein clustering approach used herein can be found in the Supplementary methods.

Representative sequences from each protein cluster among BA1A *Meiothermus* MAGs were uploaded to KEGG and analyzed using the BlastKOALA annotation pipeline (Kanehisa et al., 2020). This pipeline assigns hierarchical annotations to protein sequences, wherein each uploaded protein sequence is annotated with the closest database protein, and then is placed into a hierarchy of biological pathways. These annotations were collated and compared across the six *Meiothermus* MAGs from BA1A to identify which KEGG orthologs, protein families, and biological pathways were overrepresented in certain MAGs or MAG groupings. Differences between the Type I (S) and Type II (D) clades were investigated, as were differences between more abundant and less abundant (i.e., rare) MAGs.

Identification and Curation of Single Nucleotide Variants (SNVs)

Single nucleotide variants were identified in the six high quality BA1A *Meiothermus* MAGs using Instrain (ver. 1.5.5) (Olm et al., 2021). Instrain identifies SNVs in metagenomic data by mapping the trimmed metagenomic reads against a provided reference genome. Since the BA1A *Meiothermus* MAGs are not closely related (<90% ANI) to any published reference genome, each of the six MAGs was used as a reference genome for SNV identification. SNVs were identified among the six MAGs, and these were then curated to identify and remove those that likely result from sequence read errors. Specifically, putative sequence errors were identified by calculating the ANI between sequence read pairs and the reference genome. Read pairs exhibiting pairwise ANI to the reference genome of <90% were discarded. This ensured that read pairs bearing a significant mismatch were not used for SNV calling, as this would likely result in the retrieval of false positive SNVs. The minimum consensus score was set to 90% rather than

95% to account for lower overall quality of metagenomic data relative to genomes from pure cultures.

Single nucleotide variants identified as a consensus SNV [i.e., SNVs where the variant base frequency was present in 40% of mapped reads or more (see below)] were also identified. Consensus SNVs were rare in the BA1A Oman *Meiothermus* MAGs and occurred at a frequency of about 1:25,000 base pairs (bp). Consensus SNVs typically occurred when two or more alleles were present in the population and represent either intra-strain level genetic heterogeneity or a sequencing artifact. Since there is no definitive way to determine whether these correspond to true allelic heterogeneity or are artifactual, they were discarded without further consideration. Nonetheless, since they are rare, they would have a minimal influence on the outcome of the study.

Single nucleotide variants were further curated by only using trimmed reads that had a mapQ score equal to or exceeding 1.0. The mapQ score is a read quality measurement determined by bowtie during the read mapping step of the pipeline (Langmead and Salzberg, 2012). The mapQ score is also used to validate the consensus base or the most frequent allele at a variable nucleotide site by verifying that it was also the base with the highest average mapQ score across reads exhibiting that base. Finally, a minimum and maximum sequence insert size is used to further curate SNVs. Here, if sequence inserted into the middle of an ORF alignment is too large, the SNVs in that alignment are discarded since the insertion may indicate incorrect alignment of sequences that are separated on the genome. Default parameters of 50 bp and 3× median insert size were used for minimum and maximum sequence insert size, respectively. The total number of SNVs, both inside and outside of ORFs, was determined for each MAG for

comparative purposes. However, only SNVs that were within ORFs, as identified by PROKKA, were used for downstream analyses, since these would allow for potential annotation of the encoded proteins. This secondary filtering step was performed to reduce the frequency of false positive SNVs, as the ORFs tended to have much higher read quality than non-coding sequences. Further, SNVs on ORFs are expected to be more informative regarding the functional dynamics of populations and thus more indicative of where selective pressures are acting on the population level genomes.

Curated SNVs among BA1A *Meiothermus* MAGs were collated using the R base package to facilitate downstream comparison. Specifically, a table was uploaded to R that contained (1) each SNV, (2) the nucleotide frequencies of that SNV, (3) the location of that SNV on its ORF, (4) which of the six MAGs contained that SNV, (5) the full nucleotide sequence of the ORF containing that SNV, (6) the inferred amino acid sequence encoded by that ORF, (7) the assigned KEGG ortholog of the inferred protein encoded by that region (where possible), (8) the protein cluster that protein was assigned to, and (9) the KEGG pathways that the protein identified in the preceding step presumably belongs to. This allowed for identification of SNVs shared by multiple MAGs, statistical evaluation of those SNVs, and other relevant analyses within the R statistical platform.

The overall SNV profile was determined for each MAG. Specifically, the total number of SNVs in the MAG was identified, regardless of whether they were on an ORF. Similarly, the number of SNVs in ORFs in those MAGs, the proportion of ORFs in the MAG containing at least one SNV, and the mean and variance of the SNV frequency across the length of the MAG were identified. Student's T-tests were performed to compare total SNVs among pairs of MAGs,

SNVs in ORFs among pairs of MAGs, variance in SNVs per ORF among pairs of MAGs, and the total number of ORFs containing at least one SNV across the designations of community (depth from which it was recovered), abundance (rare versus abundant), and phylogeny (shallow clade versus deep clade, as described below).

Characterization of SNVs

The six BA1A *Meiothermus* MAGs were compared to identify SNVs that were shared at the same position within an ORF. Shared SNVs identified in all six MAG populations were presumed to be ancestral SNVs that were likely present in the population of founder cells, or those that first colonized the Samail Ophiolite or its geologic precursor, and whose genomes gave rise to the descendent MAGs identified herein. The distribution of SNVs was examined between Type I (S) and Type II (D) MAGs using a hierarchical approach. First, shared SNVs that are present in more than one MAG were identified. This approach was more facile than alignment-based approaches that are more time and computationally intensive. While the latter would identify many more potential SNVs, it would also increase the identification of false positive SNVs. Consequently, SNVs identified using the approaches reported herein should be regarded as a conservative estimate of the total number. Secondly, shared SNVs with the same substitution (i.e., the same consensus base and same variable base) at the same site on the same protein encoding ORF were identified. The number of shared SNVs across these hierarchical categories was subjected to statistical analysis using the R base package. Student's t-tests were used to compare the number of shared SNVs by different MAG pairs or MAG triplets.

Evaluation of the Correlation Between Shared SNV Presence and Pairwise MAG Sequence Similarity

To determine if the presence of shared SNVs corresponded to areas of genomes with high similarity, the ANI of pairs of contigs in pairs of MAGs that contained shared SNVs was determined and compared to the average ANI of pairs of contigs of those same pairs of MAGs that lacked shared SNVs. FastANI (version 1.3.2) (Jain et al., 2018) was used to determine the pairwise ANI of every contig pair between MAGs. Contig pairs containing a shared SNV were then identified. The average ANI for contig pairs containing a shared SNV was then determined for each MAG pair and was compared to the average ANI for contig pairs in that MAG pair that did not contain a shared ANI. The length and coverage of contigs containing shared SNVs and those contigs lacking shared SNVs were also compared to ensure that sequence quality and read depth was not a confounding factor in this analysis. No significant difference was observed between contig coverage (i.e., read depth) between the two groups ($P = 0.12$; two tailed t-test). There was a significant difference in length between contig groups ($P < 0.01$; two tailed t-test), but this is likely a consequence of large sample sizes rather than a true difference, as mean contig lengths were similar (84,325 bp for contigs containing shared SNVs, 50,007 bp for those without), as were the standard deviations (73,698 bp for contigs containing shared SNVs, 63,906 bp for those without). However, since contigs containing less than 80% ANI cannot reliably be aligned, the reported similarity values are intended to be comparative, not an absolute measurement of the actual sequence similarity between specific portions of the genome. Nonetheless, due to the large number of total calculated alignments, the difference in similarity between genomic regions with shared SNVs and those without is expected to be real.

Estimating the Likelihood that Shared SNVs Arose Through Convergent Evolution

To evaluate the likelihood that SNVs arose in *Meiothermus* MAGs convergently rather than being inherited from a common ancestor (allopatry) or through gene flow/recombination between MAGs, a simple site substitution model was developed and employed. For each MAG pair, the number of SNVs occurring at the same site (i.e., on the same residue on the same ORF) was determined by comparing the retrieved SNVs from InStrain for each MAG. The previously discussed protein clustering was used to identify orthologous proteins and the genes encoding them. This protein clustering approach was utilized in place of a much more computationally intense and error prone whole genome alignment-based approach. The number of shared SNVs that would be expected to have the same base substitution (i.e., the same consensus and variant bases) through convergent evolution (i.e., random mutations and not through horizontal or vertical gene flow) was calculated. Here, roughly 12/256 of SNVs are expected to converge on the same substitution at the same site through random mutation alone. The denominator of 256 (i.e., the 256 total possible pairs of base changes) comes from a base at a given position in the MAG (4 bases are possible) and the possibility of a variant base (4 bases are possible) at this position (4 bases $\times$ 4 bases or 16 base combinations). The same applies to the second MAG pair, resulting in 16 combinations $\times$ 16 combinations or 256 total combinations for the identity of the original base and what it could mutate to at a single position in both MAGs. The numerator is based on the 16 total pairings where both SNVs start and end with the same base. Thus, the likelihood two SNVs arose from the same substitution is (4 bases choose 2 bases) 2 or $(4)^2$ or 16 minus 4 (the four instances where no base change occurs are subtracted out) is 12. The combined likelihood is therefore 12/256.

For triplet sets of MAGs, the number of possible variations is now based on three mutations rather than two. The likelihood that the three MAGs convergently develop the same SNV at the same position is the combined likelihood that the three MAGs experience a mutation from one nucleotide to another nucleotide (i.e., three MAGs experience the same substitution at the same position in an ORF). Consequently, in the absence of gene flow, it is expected that roughly 12/4,096 of SNVs will converge on the same substitution at the same site. The 4,096 denominator comes from a MAG having an original base at a given position and the possibility of a variant base at this position (4 bases $\times$ 4 bases or 16 base combinations). The same is true for the second and third MAGs, resulting in a total of 4,096 combinations (16 $\times$ 16 $\times$ 16). Thus, the likelihood two SNVs arose from the same substitution is (4 bases choose 2 bases)² or (4)² or 16 minus 4 (the four instances where no base change occurs) or 12. The combined likelihood is therefore 12/4,096.

Identification of Viral and Integrated Mobile Genetic Elements (IMGEs)

To identify the quantity and type of viral elements present in the three BA1A subsurface microbial communities, metagenomes were probed for viral nucleotide sequences using VIBRANT (version 1.2.1) (Kieft et al., 2020), which identifies viral sequences by comparing retrieved metagenomic sequences to the KEGG, protein Family (PFAM), and Virus Orthologous Groups (VOG) databases (Thannesberger et al., 2017; El- Gebali et al., 2018; Kanehisa et al., 2020). Contigs from the six BA1A *Meiothermus* MAGs were submitted to the Mobile Genetic Element finder tool (version 1.0.3) (Johansson et al., 2021) to identify the quantity, location, and type of IMGEs present.

Results

Overview of Previous Hydrogeochemical Work Conducted at BA1A

Hydrological experiments, including flowmeter tests under ambient and forced hydraulic conditions, revealed three discrete depth intervals with connectivity to local aquifers in the 400 m deep BA1A well and these intersected different bedrock types, as reported previously (Lods et al., 2020; Kelemen et al., 2021). Briefly, the 0–30 m depth interval (BA1A30) intersected alluvium (well cased to 21 m to isolate this section) and weathered dunite to at least 30 m, whereas the 41–65 m (BA1A65) and the 108–132 m (BA1A132) depth intervals intersected intact dunite. There was no significant connectivity of upper aquifers to aquifers between 132–400 m depth, as assessed in field with a flowmeter (limit of detection of 0.1 L min⁻¹). Field hydraulic tests suggested that waters from the 0–30 m interval may have limited connectivity with those from 41–65 m interval (Lods et al., 2020). Further, it is possible that connectivity exists to the deeper 108–132 m aquifer, although the flow is expected to be substantively less than from the weathered dunite aquifer (0–30 m interval) to the upper (41–65 m) dunite-hosted aquifer (Lods et al., 2020).

The different depth intervals sampled generally corresponded to different geochemical regimes (Lods et al., 2020; Nothaft et al., 2021), with the 0–30 m depth water having a temperature of 34.9°C, a pH of 8.10, a conductivity of 0.458 mS cm⁻¹, and an oxidation reduction potential (Eh) of 128 mV (Table 1). Similarly, water from the 41–65 m depth interval had a temperature of 35.0°C, a pH of 8.21, a conductivity of 0.402 mS cm⁻¹, and an Eh of 120 mV. In contrast, water from the 108–132 m depth interval had a temperature of 36.5°C, a pH of 10.67, a conductivity of 0.871 mS cm⁻¹, and an Eh of -249 mV. Thus, waters from 0–30 m and

41–65 m are classified as Mg–HCO₃-type waters (Type I) while waters from 108 to 132 m are classified as Ca–OH-type waters (Type II), according to a previously reported classification scheme (Rempfert et al., 2017). Type II waters have extensively reacted with rock (i.e., dunite) in regions of the subsurface thought to be closed to atmospheric inputs whereas Type I waters may include atmospheric inputs (Rempfert et al., 2017).

Isolated depth interval (meters below surface)	Water type	pH	Temperature (°C)	Conductivity (μS cm ⁻¹)	Redox potential (mV)
0–30	Mg_HCO ₃	8.1	34.9	458	127.7
41–65	Mg_HCO ₃	8.2	35.0	402	120.3
108–132	Ca_OH	10.6	36.4	950	–310.0

Table 1. Select geochemical measurements of waters recovered from discrete depth intervals in well BA1A in February 2019 and their inferred water type. Mg-HCO₃ type fluids are also termed Type I fluids whereas Ca-OH fluids are also termed Type II fluids, based on prior classification schemes (Barnes and O’Neil, 1969; Neal and Stanger, 1985). Values were previously reported in Nothaft et al. (2021).

BA1A Taxonomic Composition and Diversity

A total of 447, 626, and 229 giga base pairs (Gbp) of metagenomic sequence from depth intervals of 0–30 m (BA1A30), 41–65 m (BA1A65), and 108–132 m (BA1A132), respectively, was generated and subjected to assembly and binning (Supplementary Table 1). A similar proportion of each of the BA1A30, BA1A65, and BA1A132 communities was binned (77, 80, and 79%, respectively). In BA1A30, a total of 11 high quality MAGs (i.e., MAGs that had at least 90% completeness and less than 5% contamination) were identified. Thirteen high quality MAGs were recovered from BA1A65, and ten high quality MAGs were recovered

from BA1A132. The taxonomic compositions of communities were consistent with those of Nothaft et al. (2021) and showed a decreasing proportion of Proteobacteria with depth and an increasing proportion of Nitrospirae and Deinococcota with depth (Figure 1 and Supplementary Table 2). The most abundant MAGs in BA1A30 were affiliated with Parvibaculum (Proteobacteria; 25% of the binned community), Burkholderia (Proteobacteria; 8%), Ignavibacteria (5%), Patescibacteria (3%), and Thermodesulfovibrio (3%). The composition of BA1A65 was similar to BA1A30 at the phylum level and was dominated by Proteobacteria, including an unclassified Acidiferrobacterales population (9% of the binned community), Burkholderia (7%), and Lysobacter (6%). Like BA1A30, Ignavibacteria (5% of the binned community) and Thermodesulfovibrio (5%) were also detected in the BA1A65 community. A major shift in taxonomic composition was observed in BA1A132, where Thermodesulfovibrio comprised nearly 50% of the binned community. *Meiothermus* was also abundant in BA1A132, comprising roughly 10% of the binned community.

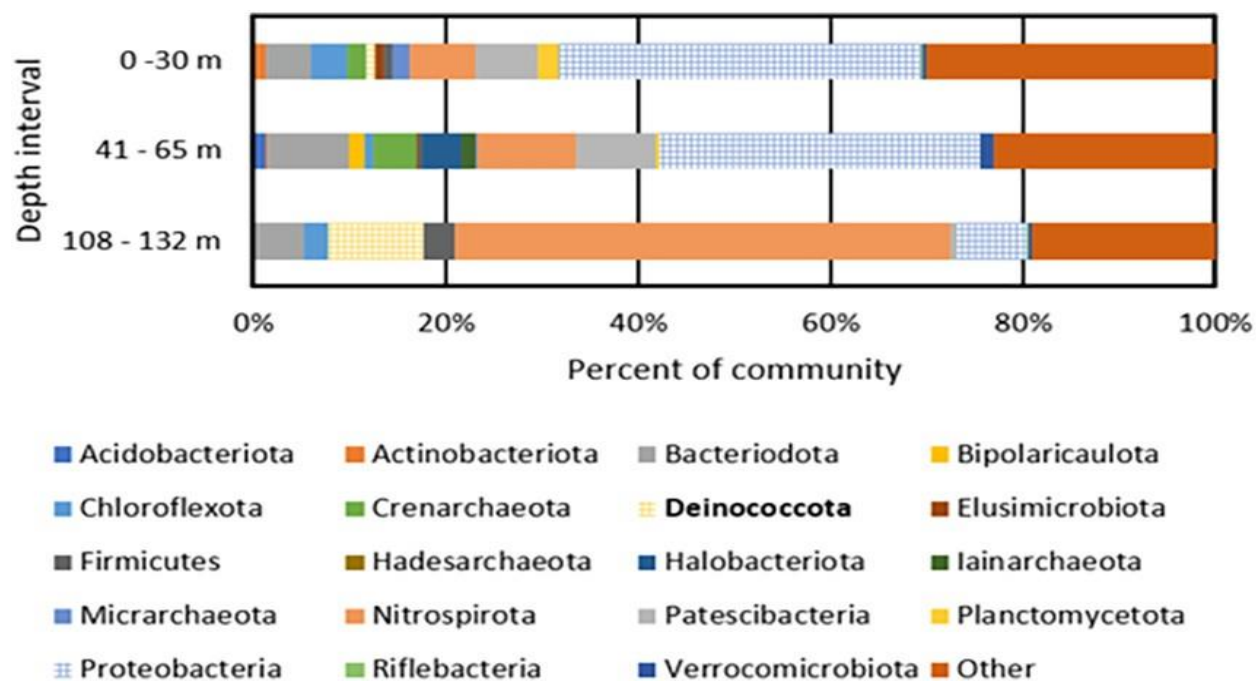


Figure 1. Phylum level taxonomic composition of metagenome assembled genomes (MAGs) recovered from BA1A filtered biomass from discrete depth intervals. MAGs that encode homologs of enzymes or enzyme complexes that allow for the respiration of oxygen (Cox I and II, CydABX) are depicted with a grid pattern. Deinococcota, the phylum containing *Meiothermus*, is bolded in the legend.

Inferred Oxygen Usage Among BA1A Populations

Metagenome assembled genomes from the discrete depth intervals sampled in BA1A were examined for protein homologs of cytochrome c oxidase [Cox I and II; Enzyme Category (EC) 7.1.1.9] and cytochrome bd complex (CydABX; EC 7.1.1.7), known protein complexes involved in the respiration of O₂ (Ludwig, 1987; Jünemann, 1997). MAGs that encoded homologs of one or both complexes were considered as capable of integrating O₂ into their energy metabolism (i.e., obligate aerobes or facultative anaerobes). MAGs that did not encode homologs of these complexes were considered incapable of integrating O₂ in their cellular energy metabolism (i.e., obligate anaerobes). Forty-three and 37% of the binned BA1A30 and BA1A65 populations were determined to be aerobic/facultatively anaerobic, respectively, with

the remainder inferred to be strict anaerobes (Figure 1). Surprisingly, 20% of the BA1A132 community was determined to be aerobic/facultatively anaerobic based on MAGs encoding Cox; *Meiothermus* comprised 10% of this community. Homologs of proteins that would putatively allow for respiration of additional oxidants (e.g., SO_4^{2-} , S° , Fe(III), NO_3^-) were not identified among BA1A *Meiothermus* MAGs (data not shown). In addition to encoding for Cox, BA1A *Meiothermus* MAGs are closely related (83–85% ANI) to the obligately aerobic cultivar, *Meiothermus hypogaeus* (NCBI taxonomy ID: 884155) (Mori et al., 2012), suggesting a similar O_2 -dependent energy metabolism.

To investigate if putative aerobes in BA1A may be capable of generating endogenous O_2 thereby allowing for a high abundance of putative aerobes in an otherwise anoxic environment, MAGs were probed for genes that encode oxygenic proteins including Nod, Cld, SOD, and Cat (see section “Materials and methods”). Homologs of Cld, including both the pentameric and dimeric forms, were the only such proteins identified in BA1A MAGs. One hundred and fourteen Cld sequences among the BA1A MAGs passed the alignment and conserved residue cutoff used to curate homologs (Supplementary Table 3). All six *Meiothermus* MAGs from BA1A (discussed below) encoded a homolog of Cld, regardless of the depth from which they were recovered.

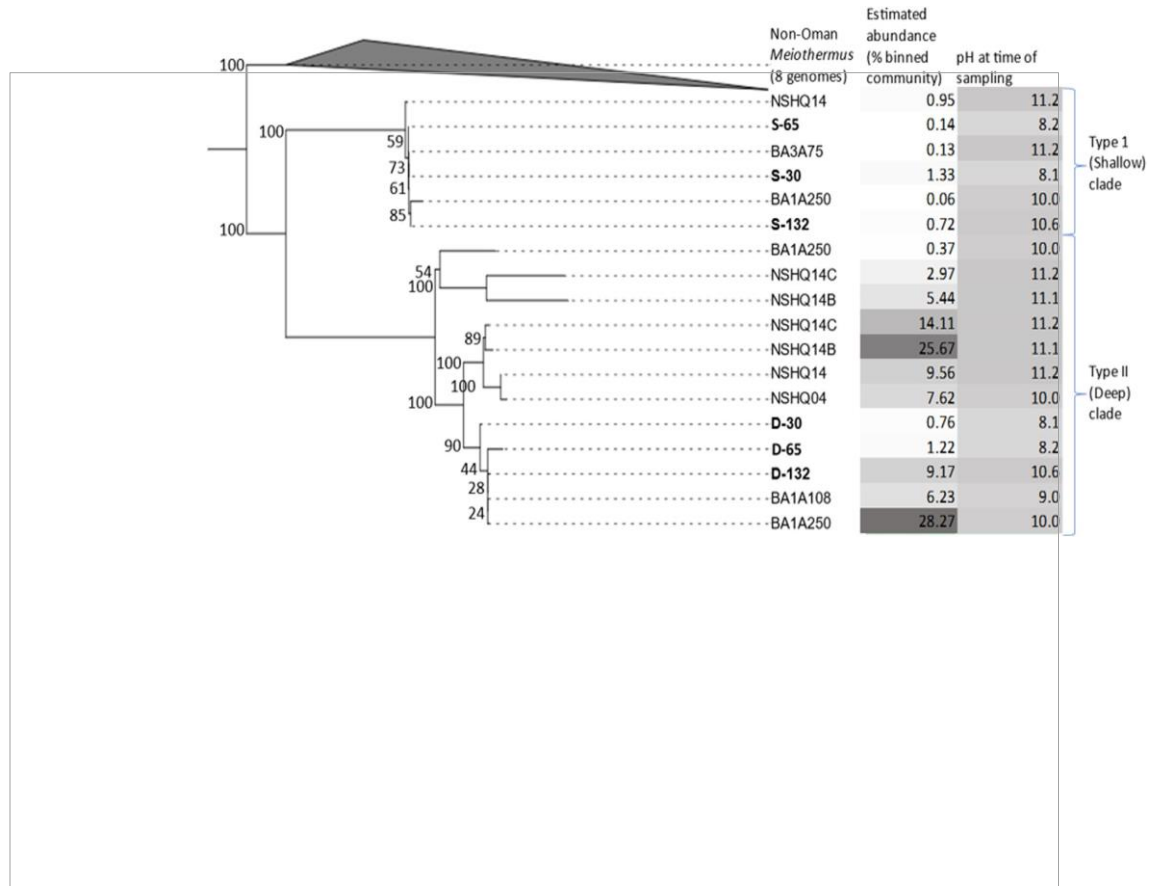


Figure 2. Maximum-likelihood phylogenomic reconstruction of a concatenation of 115 housekeeping genes from metagenome assembled genomes (MAGs) recovered from discrete depth intervals in wells BA1A (sampled with packer system), BA1A (sampled using niskin bottles), BA3A (sampled using niskin bottles), or from specified depths (open well pumping) in NSHQ14B (50 m) and NSHQ14C (85 m). Eight genomes of non-Oman *Meiothermus* served as the outgroup. For BA1A samples collected with the packer system, MAGs are labeled as belonging to the Type I surface phylogenetic clade (S) or the Type II deep phylogenetic clade (D) followed by the maximum depth interval (m) from where they were recovered (30, 65, 132). For BA1A and BA3A samples collected with niskin bottles, MAGs are labeled by well followed by depth of recovery. Only MAGs that were most closely related to *Meiothermus* were considered. MAGs obtained from discrete depth intervals isolated by packers in well BA1A that were used for further genomic analyses are in bold. Bootstrap scores are indicated at each node. The relative abundance of MAGs (% of mapped reads) and the pH of fluids at the time of sampling are indicated, with darker colors corresponding to higher values for both parameters.

Meiothermus Phylogeny and Clade Designations

Phylogenomic reconstruction of *Meiothermus* MAGs from BA1A, *Meiothermus* from other wells in the Samail Ophiolite study site, and all *Meiothermus* reference genomes available in the Genome taxonomy database (Figure 2) revealed that MAGs from Oman formed a monophyletic clade, with *M. hypogaeus* as the closest non-Oman outgroup. However, *Meiothermus* MAGs from other serpentinizing systems were not yet available to compare phylogenetically to the *Meiothermus* from the Samail Ophiolite. Yet, a single RNA polymerase subunit B (RpoB) sequence from *Meiothermus* was available in the NCBI NR and SRA databases from a metagenome from the Zambales ophiolite (Woycheese et al., 2015). Phylogenetic reconstruction of RpoB homologs from *Meiothermus* MAGs from BA1A, *Meiothermus*-affiliated RpoB from the Zambales ophiolite, and *Meiothermus* RpoB sequences from the NCBI NR and SRA databases revealed the same pattern whereby *Meiothermus* RpoB from the Samail Ophiolite formed a monophyletic lineage, referred to herein as the Oman-specific clade (Supplementary Figure 1).

Within the Oman-specific *Meiothermus* clade, MAGs from BA1A partitioned into two distinct sub-clades. Both *Meiothermus* sub-clades were present in all three depth intervals, but they exhibited opposing abundance trends (Table 2). Specifically, one *Meiothermus* clade, hence referred to as Type II (D) from the BA1A30 (D-30), BA1A65 (D-65), and BA1A132 (D-132) intervals, increased in abundance as depth increased, whereas the other *Meiothermus* clade, hence referred to as the Type I (S) clade comprising BA1A30 (S-30), BA1A65 (S-65), and BA1A132 (S-132), is most abundant in the BA1A30 community and decreased in abundance with depth. Within clade average nucleotide identity (ANI) was 99.99% for each clade, indicating a single population. However, ANI between MAGs forming Type I (S) and Type II

(D) clades averaged 84%, whilst the ANI between Oman-specific *Meiothermus* MAGs and *Meiothermus hypogaeus* varied between 78 and 81%, depending on the MAG.

Meiothermus Inferred Genome Replication Rates

To begin to assess whether *Meiothermus* MAGs detected in each depth interval correspond to cells that are likely active or inactive, genome replication rates were estimated using metagenomic sequence data and SMEG calculations. SMEG scores for five of the six high quality *Meiothermus*-affiliated MAGs in BA1A indicate active genome replication, as indicated by a SMEG score exceeding 1.0 (Table 2). The sole exception was *Meiothermus* S-132, which belongs to the Type I (S) clade but that was recovered from the BA1A132 community. The other two MAGs that belong to the Type I (S) clade, S-30 and S-65, had SMEG values of 3.019 and 2.722, respectively. In contrast, MAGs that belong to the Type II (D) clade that were detected in BA1A30 (D-30), BA1A65 (D-65), and BA1A132 (D-132) communities had SMEG values of 1.060, 1.042, and 1.013, respectively. The 95% confidence intervals for all three Type II (D) clades are above one, indicating that these values are significantly different than 1.000 (non-replicating genome). Thus, while the MAGs corresponding to the Type II (D) clade are in a state of active genome replication, they are doing so more slowly than the Type I (S) clade at the 0–30 m and 41–65 m intervals. Further, the Type I (S) *Meiothermus* MAG S-132, recovered from Type II waters, is not in a state of active genome replication.

Metagenome assembled genome (MAG)	Estimated abundance (% binned community)	Abundance designation	Estimated genome size (Mbp)	Estimated number of protein coding genes	Inferred genome replication rate (SMEG)	Single nucleotide variants
S-30	1.33	Abundant	3.56	3,278	3.02	1,473
S-65	0.14	Rare	3.25	2,953	2.72	5,016
S-132	0.72	Rare	3.71	3,526	1.00	15,560
D-30	0.76	Rare	2.82	2,656	1.06	6,942
D-65	1.22	Abundant	3.12	2,949	1.04	1,335
D-132	9.17	Abundant	3.00	2,869	1.01	2,589
M. hypogaeus	NA	NA	3.70	3,623	NA	NA

Table 2. Name, estimated abundance, estimated genome size (mega basepairs, Mbp), estimated number of protein coding genes, strain level metagenomic estimation of growth rate (SMEG), and number of single nucleotide variants detected for metagenome assembled genomes (MAGs) recovered from discrete intervals in well BA1A. MAG names are labeled as belonging to the Type I surface phylogenetic clade (S) or the Type II deep phylogenetic clade (D) followed by the maximum depth interval (m) from where they were recovered (30, 65, 132). MAGs were also empirically defined as abundant versus rare based on read mapping (see section “Materials and methods”). Genome characteristics of the most closely related *Meiothermus* cultivar that is not from Oman, *Meiothermus hypogaeus*, are provided for context. A SMEG score of 1.00 indicates a non-replicating MAG. NA, not available.

Meiothermus Genome Sizes

To further examine differences in the BA1A MAGs that correspond to the Type I (S) and Type II (D) clades, their genome sizes were inferred. The inferred genome size was similar for MAGs from each respective clade (Table 2). MAGs belonging to the Type I (S) clade had an estimated genome size of 3.5 Mega-basepairs (mbp), which is comparable to the observed genome sizes of most other described *Meiothermus* species, including the closest non- Oman relative, *M. hypogaeus*, which has an observed genome size of 3.7 mbp (Mori et al., 2012). However, the Type II (D) clade MAGs had much smaller inferred genomes, with an average size of 3.0 mbp.

Protein Clustering of BA1A *Meiothermus* MAGs and their Inferred Metabolism

The larger genome sizes of MAGs that cluster within Type I (S) versus Type II (D) *Meiothermus* clades suggested differences in the functional potential of the associated organisms. To determine the differences in encoded protein content between the six *Meiothermus* MAGs, the inferred proteins encoded in MAGs were subjected to protein clustering and comparative analysis (Supplementary Table 5). The 17,477 inferred protein sequences [encoded by genes or open reading frames (ORFs)] among the six BA1A *Meiothermus* MAGs partitioned into 4,114 clusters, 2,169 of which included protein homologs encoded in both the Type I (S) and the Type II (D) clade MAGs. Among the proteins that were shared between the Type I (S) and the Type II (D) clade MAGs are homologs of proteins for complete glycolytic and tricarboxylic acid cycles as well as homologs of bidirectional NAD⁺/NADP⁺-reducing [NiFe]-hydrogenases. Both clades also encoded a homolog of group 1 [NiFe]-hydrogenases, however, they are likely involved in different processes based on phylogenetic relationships of the large subunit with characterized

homologs and the presence/absence of motifs indicative of translocation across the membrane.

The Type I (S) clade group 1 [NiFe]-hydrogenase homolog is predicted to be located in the periplasm based on the presence of a twin-arginine translocation (tat) motif and to be a high affinity H₂ oxidation enzyme (group 1h) that functions in aerobic respiration (Søndergaard et al., 2016). In contrast, the Type II (D) clade group 1 [NiFe]-hydrogenase homolog is predicted to be cytoplasmic or inner membrane-associated (lack of tat motif) and is predicted to be involved in H₂ oxidation coupled to anaerobic respiration of sulfate, fumarate, nitrate, or metals (Søndergaard et al., 2016). Interestingly, homologs allowing for respiration of such oxidants were not identified among the Type II (D) clade MAGs.

In addition, of the 4,114 protein clusters, 1,196 were unique to the Type I (S) clade MAGs, and 749 were unique to the Type II (D) clade MAGs. Importantly, because these MAGs are incomplete (Supplementary Table 2) it cannot be known confidently that a protein sequence is necessarily unique to a clade. Nonetheless, that they were identified in three Type I (S) MAGs and no Type II (D) MAG and vice versa gives additional confidence to their uniqueness to a given clade. A far greater proportion of the proteins unique to the Type I (S) clade of BA1A *Meiothermus* MAGs were annotated in the KEGG categories of ABC transporters (26× as likely) and amino sugar and nucleotide sugar metabolism (10× as likely). Proteins unique to the Type I (S) clade of *Meiothermus* MAGs were also more likely annotated as two component regulatory systems (10× as likely) or as involved in quorum sensing (4× as likely). A number of protein encoding genes associated with sulfur metabolism were also identified in Type I (S) clade MAGs, including those encoding SoxABCDX that would potentially allow for oxidation of thiosulfate/elemental sulfur. Intriguingly, while these genes are encoded by most

non-Oman *Meiothermus* species, including those that are sister to the Oman *Meiothermus* clade (Figure 2), they are absent from the Type II (D) Oman clade. This suggests loss of these genes and the functionalities in the Type II (D) clade.

Proteins unique to the Type II (D) clade MAGs tended to be annotated in the KEGG categories of thiamine biosynthesis ($3.5\times$ as likely) and folate biosynthesis ($5\times$ as likely). Moreover, a complete thiamine biosynthesis pathway (ThiEMDLGOS and its regulator TenA) was identified exclusively in the Type II (D) clade MAGs. Based on close sequence homology of these subunits to Acetothermia and their absence in other non-Oman *Meiothermus* genomes, it is likely that the thiamine biosynthesis pathway was acquired via a horizontal gene transfer event from Acetothermia that tend to be enriched in Type II waters (Colman et al., 2022). Additionally, the Type II (D) clade encodes a copy of pyruvate ferredoxin oxidoreductase (PFOR; porABC), which, based on sequence homology, may have been obtained via HGT from cohabitating and anaerobic *Thermodesulfovibrio* that often predominates in Type II waters (Templeton et al., 2021). In addition, all six MAGs, regardless of whether they correspond to the shallow or the Type II (D) clade, encode pyruvate dehydrogenase (PDH) that typically facilitates oxidative conversion of pyruvate to acetyl CoA and CO₂ (unidirectional) in aerobes (de Kok et al., 1998).

SNV Profiles of Individual BA1A *Meiothermus* MAGs

To assess the extent of gene exchange among spatially or ecologically differentiated *Meiothermus* populations, a population biology study of MAGs via analyses of SNVs was undertaken. A total of 35,019 curated (see section “Materials and methods”) SNVs were identified among the six BA1A *Meiothermus* MAGs, 22,476 of which were in ORFs (Table 3). SNVs were not distributed evenly across the MAGs. MAGs classified as rare [i.e., are the less

abundant *Meiothermus* population in their respective community such as S-132 (Table 2)] had a significantly higher number of ORFs containing SNVs than their abundant counterparts ($P = 0.00047$; two tailed T-test). Further, they had a significantly higher total number of SNVs in ORFs ($P = 0.012$) and a significantly higher mean number of SNV's per ORF ($P = 0.0013$). There was no significant difference between the mean number of SNVs of MAGs that belonged to different clades [i.e., Type I (S) versus Type II (D)] or that were recovered from the different depths. However, the clade the MAG belonged to [i.e., Type I (S) versus Type II (D)] was a slightly stronger, albeit not significant, predictor of the mean number of SNVs across the MAGs (phylogenetic clade $P = 0.075$; abundance $P = 0.20$).

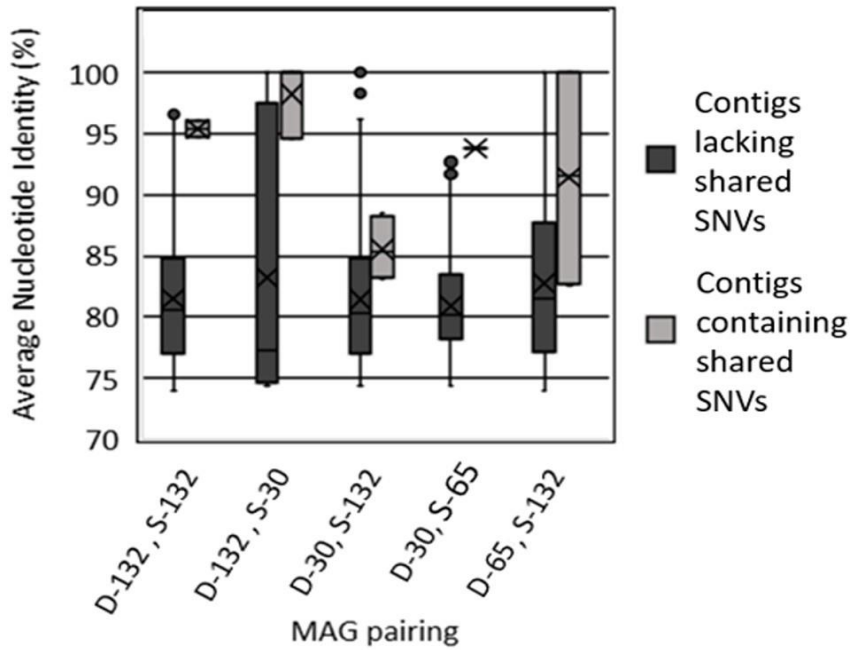


Figure 3. Average nucleotide identity of contigs between select pairs of metagenome assembled genomes (MAGs) that contain shared single nucleotide variants (SNVs) versus contigs from those pairs of MAGs that lack shared SNVs. MAGs are labeled as belonging to the Type I surface phylogenetic clade (S) or the Type II deep phylogenetic clade (D) followed by the maximum depth interval (m) from where they were recovered (30, 65, 132). Values for all MAG pairings are reported in Table 3.

SNVs Shared Between MAGs

Of the 4,484 total protein clusters identified among the six *Meiothermus* MAGs, 369 protein clusters had no SNVs within their associated ORF, while 1,322 clusters had at least one SNV in an associated ORF for each *Meiothermus* MAG. The 369 invariable ORFs encode proteins that are highly conserved and/or that are essential (e.g., housekeeping proteins) (data not shown). There was no significant difference in the number of shared SNVs by MAGs in the same clade or shared by MAGs with similar abundances (Table 2). Pairs of MAGs where both members came from the same depth interval, such as S-30/D-30, S-65/D-65, and S-132/D-132, have significantly fewer shared SNVs at the same site within a protein encoding gene than other

pairs. The pairing of S-65/D-65 had no SNVs at the same site whereas the pairing of S-30/D-30 had two SNVs at the same site but with different base substitutions. Finally, the pairing of S-132/D-132 had three SNVs at the same site, two of which had the same base substitution. Surprisingly, two pairs of MAGs (S-65/D-132 and S-30/S-132) have a significantly higher number of shared SNVs in protein encoding ORFs than other pairs of MAGs, with 219 and 148 SNVs at the same site with the same base substitution, respectively (Table 3). This was despite the fact that the organisms with these MAGs have been evolutionarily/spatially more isolated from each other than those organisms from the same aquifer. The limited evidence for shared SNVs in MAGs recovered from the same depth is attributed at least in part to limitations associated with the approach utilized herein, where a given contig is only allowed to be present in a single bin due to the non-redundant nature of the contig binning process. As such, the results presented herein for co-inhabiting populations should be considered a conservative estimate of gene flow/recombination across phylogenetic lineages.

The only triplet sets of BA1A MAGs that had SNVs with the same substitution at the same site within a protein encoding gene occurred among MAGs that belonged to the same phylogenetic clade (Table 3). MAGs comprising the Type II (D) clade had ten SNVs at the same site of a protein encoding gene, two of which had the same base substitution. MAGs comprising the Type I (S) clade had five SNVs at the same site of a protein encoding gene, of which four had the same base substitution.

Correlation Between Shared SNVs and Pairwise MAG Genome Similarity

The pairwise ANI between each contig in pairs of BA1A *Meiothermus* MAGs was calculated to determine if genomic regions containing shared SNVs at the same position were

more similar than regions lacking shared SNVs. Importantly, it is also possible to calculate the ANI of individual protein coding genes (i.e., genes encoding proteins among protein clusters) that are shared between MAGs to identify those that likely arose from recent recombination (high ANI compared to average ANI). However, such an approach could confound an assessment of the influence of recombination due to potential differences in selective pressures on encoded proteins (i.e., false-positives among housekeeping genes under purifying selection, false-negatives among protein genes undergoing drift). A comparison of ANIs between contigs with shared SNVs among MAG pairs thus is a more conservative approach to identifying regions of MAGs that have been influenced by recombination and is the approach utilized herein.

Of the 10 MAG pairings that shared SNVs (Table 3), all exhibited higher genomic similarity in the regions containing those SNVs than in the other regions of the genome (Figure 3 and Table 3).

Metagenome assembled genome (MAG) pair or triplet	Shared SNVs at same site	Shared SNVs at same site with same substitution	Expected number of convergent SNVs	Average ANI between contigs containing shared SNVs	Average ANI between contigs lacking shared SNVs
S-30, D-30	2	0	<1	NA	81.0
S-65, D-65	0	0	0	NA	81.8
S-132, D-132	3	2	<1	95.4	81.5
S-65, S-132	259	219	12.14	100	83.5
D-65, D-132	62	44	2.91	100	83.9
D-30, D-65	25	11	1.17	98.1	82.5
D-30, S-132	73	9	3.42	85.6	81.4

S-30, S-65	25	22	1.17	100	86.6
S-30, D-132	8	0	0.38	NA	81.4
S-30, D-65	37	0	1.73	NA	81.5
S-30, S-132	173	148	8.11	100	84.1
D-65, S-132	34	20	1.59	91.4	82.7
D-30, S-65	15	3	<1	93.8	80.9
D-30, D-132	27	15	1.27	98.4	83.5
S-65, D-132	1	0	<1	NA	81.4
D-30, D-65, D-132 (deep					
clade MAGs)	10	2	<1	NA	NA
S-30, S-65, S-132 (shallow clade MAGs)	5	4	<1	NA	NA
S-30, D-65, D-132 (common MAGs)	0	0	0	NA	NA
D-30, S-65, S-132					
(rare MAGs)	1	0	<1	NA	NA

Table 3. Single nucleotide variants (SNVs) identified between the specified pair or triplet of metagenome assembled genomes (MAGs). MAGs are labeled as belonging to the Type I surface phylogenetic clade (S) or the Type II deep phylogenetic clade (D) followed by the maximum depth interval (m) from where they were recovered (30, 60, 132). The number of shared SNVs at the same site and those with the same substitution are reported, as is the expected number of SNVs that could arise through evolutionary convergence (see section “Materials and methods”). The average nucleotide identity (ANI) of contigs that share versus those that lack shared SNVs is provided. Bolded pairings or triplets denote MAGs that exhibit evidence for gene flow/recombination. A) The number SNVs that would theoretically be present in MAG pairs or triplets due to convergent evolution in the absence of gene flow/recombination as calculated based on substitution frequencies (see section “Materials and methods”) multiplied by the number of SNVs shared by a pair or triplet of MAGs, as identified by InStrain (first column above). Values were rounded to the nearest integer. B) ANIs are not reported for pairs of MAGs with no shared SNVs. C) ANI calculations were not performed for triplets of MAGs. overall sequence quality ($P = 0.12$; two tailed T-test). Further, while not every contig pair could be evaluated, as ANI is not reliable for sequences with less than 80% sequence similarity, the sample size of those contigs that were evaluated indicates that while the actual percentage similarities may not be accurate, the overall trend is accurate. In MAG pairs belonging to the same clade, much of this genomic similarity is likely a consequence of shared ancestry. Since the majority of contigs had shared SNVs, their high ANI reflects the overall high ANI between the MAG pairs. However, for the five pairs belonging to different clades but still exhibiting shared SNVs, this increase in genomic similarity near shared SNVs is likely to be a consequence of recombination in at least part of that contig.

Likelihood of SNVs Arising Convergently

A simple site substitution model was developed to evaluate the likelihood that SNVs in MAG pairs or MAG triplets arose convergently as opposed to having been either a property of their ancestor or result from gene flow/recombination (see section “Materials and methods” for model description). Ultimately, the model predicts that 12 out of 256 of all SNVs occurring at the same site between two MAGs and that 12 out of 4,096 of all SNVs occurring at the same site between three MAGs should have the same base substitution. In other words, the expected number of shared SNVs with the same substitution in a pair of MAGs in the absence of gene flow/recombination was calculated as: $12/256 \times \text{number of shared SNVs}$. As an example, the expected number of convergent SNVs between MAG pair D-65 and S-132 that have the same

base substitution is ~ 2 ($12/256 \times 34 = 1.59$ or ~ 2). However, 20 out of 34 shared SNVs at the same site in these two MAGs were found to have the same base substitution (Table 3). This suggests that convergent evolution alone (accounting for only ~ 2 of the 20 observed shared SNVs with the same base substitution) is unlikely to explain the prevalence of shared SNVs with the same base substitution at the same site. This is the case for other MAG pairs from different clades (S versus D) (Table 3). This model is an oversimplification of mutation dynamics in natural result of gene flow/recombination or common ancestry rather than convergent evolution. This same model was applied to each other MAG pair and MAG triplet to determine the likelihood that a SNV arose convergently, rather than being a result of gene flow/recombination or common ancestry. The number of expected convergent SNVs varied from near 0 (S-65/D-65 pairing) to 12 (S-65/D-132 pairing; Table 3). Ultimately, nine out of the total 15 possible pairs of *Meiothermus* MAGs exhibited probabilistic evidence for gene flow/recombination or shared ancestry as evidenced by observed shared SNVs with the same substitution exceeding expected values based on convergence alone. Seven of these nine pairs were from different depths and, more importantly, from different clades.

Identification of Viral and Non-Viral Integrated Mobile Genetic Elements (IMGEs) in BA1A Communities

Metagenome assembled genomes were evaluated for virus signatures and IMGEs. Within the BA1A30, BA1A65, and BA1A132 communities, 4,610, 7,969, and 4,482 viral sequences, respectively, were identified. However, none of the retrieved viral sequences were found on *Meiothermus* contigs. Nonetheless, 19 total IMGEs were identified on *Meiothermus* contigs, 13 of which were classified as insertion sequences (i.e., a transposase gene flanked by two inverted repeats) while the remaining six were classified as composite transposons or a composite mobile

genetic element resulting from a transposase acting on the inverted repeat of a related mobile genetic element (MGE) that transposes both the original and new element together (Clark et al., 2019). *Meiothermus* MAGs from BA1A that comprised the Type II (D) clade consistently had more IMGES than those that comprised the Type I (S) clade. Of the 19 IMGES, four were present in more than one MAG from the Type II (D) clade. The first shared IMGE, ISGlo6, is present in D-132, D-65, and S-132. The sequence includes the transposase InsK, which may be able to act on additional insertion sequences besides ISGlo6 (Fernández De Henestrosa et al., 2000). The D-132 and D-65 sequences exhibited 100% identity (Table 4), but the S-132 MAG sequence is truncated and comprised only 131 residues compared to the full 214 residues in the other MAGs. Further, the S-132 MAG ISGlo6 sequence shared only 61% sequence identities with that of D-132 and D-65 MAGs. In the D-132 and D-65 MAGs, the insertion sequence is preceded by upwards of 10 kilo basepairs (kbp) that exhibited 100% sequence identity, which may constitute part of a larger transferred sequence.

The second shared IMGE, ISMyca1, was present in D-30, D-65, D-132, and S-132 MAGs (Table 4). The D-132 and D-65 MAG ISMyca1 sequences are identical. Moreover, the shorter D-132 MAG contig containing the ISMyca1 exhibited 100% sequence similarity to a portion of the larger D-65 MAG contig. The D-30 MAG ISMyca1 sequence, however, shared only 83% sequence identity with the D-132 and D-65 MAG ISMyca1 sequences. The S-132 MAG ISMyca1 sequence exhibited less than 50% similarity to the other three ISMyca1 sequences and lacked an identifiable transposase. The third potentially shared IMGE is ISPlu21 (Table 4), which is present in the D-30 MAG and that was duplicated in the D-132 MAG, one on

the sense strand and one on the antisense strand. All three ISPlu21 sequences exhibited 100% sequence identity and are surrounded by upwards of 10 kbp also bearing 100% sequence identity.

	IS640	ISGlo6	ISMyca1	ISPa85	ISPlu21	ISUNCu3
D-132	NA	1.00	1.00	NA	1.00	NA
D-65	NA	1.00	1.00	NA	NA	NA
D-30	1.00	NA	0.83	1.00	1.00	NA
S-132	NA	0.61	0.42	NA	NA	1.00
S-65	NA	NA	NA	NA	NA	NA
S-30	NA	NA	NA	NA	NA	NA

Table 4. Percent shared nucleic acid sequence identities between pairs of mobile genetic elements (MGE) retrieved from BA1A *Meiothermus* metagenome assembled genomes (MAGs). MAGs are labeled as belonging to the Type I surface phylogenetic clade (S) or the Type II deep phylogenetic clade (D) followed by the maximum depth interval (m) from where they were recovered (30, 65, 132). The MAGs encoding MGEs are in rows and the individual MGEs are in columns. The percent nucleic acid identity of the specified MGE in each MAG is given in reference to the MGE from MAG D-132, or from whichever MAG encoded the MGE when it was absent from D-132. NA, not available.

The fourth sequence is Tn125, a common composite mobile genetic element with a wide phylogenetic distribution (Acman et al., 2022). One copy of Tn125 was detected in each of the six BA1A *Meiothermus* MAGs. Intriguingly, the degree of similarity between these Tn125 sequences does not recapitulate the presumed evolutionary history of the six BA1A *Meiothermus* MAGs, based on housekeeping genes or RpoB proteins (Table 5 and Figure 2). The Tn125 sequences of the D-132 and D-65 MAGs are identical but exhibited only 43% sequence identity to the Tn125 sequence present in the D-30 MAG. Further, the D-132 and D-65 MAG Tn125 sequences exhibited 87% identity to the S-65 MAG Tn125 sequence. Similarly, S-132 and S-30

MAG Tn125 sequences exhibited 100% identity to each other, 44% identity to the S-65 MAG Tn125 sequence, and 87% identity to the D-30 MAG Tn125 sequence.

Discussion

Spatially segregated microbial communities from three depth intervals spanning 132 m in BA1A were shown to contain abundant populations of *Meiothermus*, a genus that thus far comprises only obligately aerobic cultivars (Mori et al., 2012). This was true even in the highly reduced Type II waters encountered at depth in BA1A (>132 m). While perhaps surprising, these observations align with previous 16S rRNA gene (Rempfert et al., 2017; Nothaft et al., 2021) and metagenomic sequencing studies (Fones et al., 2019, 2021; Kraus et al., 2021) that also revealed the presence of *Meiothermus* in anoxic Type II waters in the Samail Ophiolite. For example, sequences affiliated with *Meiothermus* were detected in highly reduced fluids from samples collected from depths of 85 m in well NSHQ14 where Type II waters are encountered (Rempfert et al., 2017; Fones et al., 2021). Like the Samail Ophiolite, investigations of highly reduced fluids from the Cedars, Zambales Ophiolite, and Lobios hot springs, all of which are influenced by serpentinization, also identified *Meiothermus* sequences (López-López et al., 2015; Woycheese et al., 2015; Suzuki et al., 2017).

Several non-mutually exclusive explanations were put forth to account for the presence of *Meiothermus* in highly reduced waters including (1) *Meiothermus* was present in Type II waters due to inadvertent mixing with Type I waters during their collection and thereby represent inactive cells in said waters, (2) distinct populations of *Meiothermus* exist in Type I and Type II waters in the Samail Ophiolite that are indistinguishable via the sequence of the conservative 16S rRNA marker gene, (3) *Meiothermus* in Type II waters are aerobes capable of generating

endogenous O₂ to fuel their energy metabolism, (4) *Meiothermus* in Type II waters are facultative anaerobes, and/or (5) that there is more connectivity among the 0-30 m, 41-65 m, and 108-132 m aquifers than expected based on geochemical and hydrological data (Lods et al., 2020; Nothaft et al., 2021). More connectivity of aquifers (explanation 5) would potentially allow for movement/dispersal of *Meiothermus* cells, viruses, or DNA between surface and subsurface aquifers that could impact the extent of gene flow/recombination and shape patterns of speciation.

The use of packers to collect samples from discrete intervals in BA1A, combined with SMEG scores indicating that abundant Type II (D) MAGs are in an active state of genome replication at depth (albeit at low levels), discounts explanation 1 as a source of abundant *Meiothermus* in the 108-132 m depth interval community in BA1A (Nothaft et al., 2021). Importantly, the detection of Type I (S) *Meiothermus* (i.e., S-132), which was not in an active state of genome replication in the 108-132 m depth interval at the time of sample collection, could be due to residual water from mixing during drilling, small ambient downflow after drilling, or sources of vertical flow (other than from the borehole) from shallower aquifer levels during pumping with packers. This possibility is suggested based on a geochemical model that indicated that ~7.0% of the 108-132 m fluids are of the Type I, Mg-HCO₃ end member type (Nothaft et al., 2021). Regardless, such possibilities cannot explain the increased prevalence of Type II (D) *Meiothermus* with depth in BA1A.

MAGs from other Oman wells, and those in the NCBI NR and SRA database were unsuccessful due to the highly fragmented nature of those genes in metagenome assemblies (data not shown). This indicates that distinct *Meiothermus* phylotypes exist in the Samail Ophiolite

and that these were not distinguished in previous studies via analyses of the more conservative 16S rRNA marker gene. Together, these observations suggest that an ancestral *Meiothermus* population colonized the Samail Ophiolite or its geologic predecessor (Figure 4) and likely underwent speciation that led to two sublineages that have only been identified in the Samail Ophiolite to date. Additional metagenomic sequencing of other globally distributed ophiolites will be required to determine if these two sublineages are truly endemic to the Samail Ophiolite. These two sublineages of MAGs were termed Type I (S) and Type II (D) clades to reflect the water type (Type I versus Type II) and depth [Shallow (S) versus deep (D)] where they were most abundant (Figure 4).

	D-132	D-65	D-30	S-132	S-65	S-30
D-132	1.00	1.00	0.42	0.44	0.86	0.43
D-65	—	1.00	0.41	0.43	0.87	0.42
D-30	—	—	1.00	0.88	0.44	0.87
S-132	—	—	—	1.00	0.43	1.00
S-65	—	—	—	—	1.00	0.44
S-30	—	—	—	—	—	1.00

Table 5. Percent nucleic acid identities between pairs of transposon 125 (Tn125) sequence homologs retrieved from BA1A *Meiothermus* metagenome assembled genomes (MAGs). MAGs are labeled as belonging to the Type I surface phylogenetic clade (S) or the Type II deep phylogenetic clade (D) followed by the maximum depth interval (m) from where they were

To begin to investigate explanations (3 and 4) above, MAGs were compared and subjected to metabolic reconstruction, with the specific aim of determining if MAGs were capable of aerobic or anaerobic metabolism and/or generating their own endogenous O₂. Type II

(D) MAGs are ~14% smaller than the Type I (S) MAGs, a finding that is potentially consistent with genome streamlining to reduce the nutrient and energetic costs associated with genome replication. Genome streamlining would presumably be advantageous in the carbon and oxidant limited Type II waters, such as has been suggested previously for populations inhabiting such waters in the Samail Ophiolite (Fones et al., 2019, 2021; Colman et al., 2022). Despite differences in genome sizes, MAGs from both *Meiothermus* clades encode Cox homologs, indicating an ability to respire O₂. Intriguingly, members of both clades also encode homologs of chlorite dismutase that, in other organisms, have been shown to function in the generation of endogenous O₂ via dismutation of chlorite (Bardiya and Bae, 2011; Mlynek et al., 2011; Hofbauer et al., 2014; Celis et al., 2015). Such a mechanism could conceivably be used to power aerobic metabolism in the presence of sufficient chlorine oxyanions. However, attempts to measure perchlorate and chlorite in Oman fracture waters have been unsuccessful to date (Templeton, unpublished data), suggesting they are either maintained at undetectable levels or that these compounds are not present in the system. Further, attempts to cultivate *Meiothermus ruber* (ATCC strain 35948), which also encodes Cld, and to enrich *Meiothermus* from Oman fracture fluids under anaerobic conditions with a variety of electron donors/carbon sources and chlorite, chlorate, and perchlorate have been unsuccessful to date (Munro-Ehrlich, unpublished data). Together, these observations point to endogenous production of O₂ by *Meiothermus* (explanation 3) as being an unlikely reason for the abundance of *Meiothermus* sequence in BA1A132 communities.

The inferred proteomes indicate that the Type I (S) and Type II (D) *Meiothermus* are capable of aerobic, heterotrophic growth. However, the Type II (D) *Meiothermus* MAGs

intriguingly encode several proteins suggestive of an ability to potentially grow anaerobically and/or autotrophically. This includes a group 1 [NiFe]-hydrogenase homolog that is predicted to provide reducing equivalents for the anaerobic respiration of sulfate, fumarate, nitrate, or metals (Søndergaard et al., 2016). Further, unlike Type I (S) MAGs, Type II (D) MAGs encode PFOR that allows for the anaerobic conversion of pyruvate to acetyl CoA and CO₂ (Furdui and Ragsdale, 2000; Fuchs, 2011; Mall et al., 2018; Witt et al., 2019). PFOR is a thiamine-dependent iron-sulfur ([Fe-S]) cluster- containing enzyme (Pieulle et al., 1995) which may, at least in part, explain why Type II (D) MAGs also encode a complete thiamine biosynthesis pathway, unlike Type I (S) MAGs.

Pyruvate ferredoxin oxidoreductase is reversible, allowing for the conversion of acetyl CoA and CO₂ to pyruvate (Furdui and Ragsdale, 2000; Fuchs, 2011; Mall et al., 2018; Witt et al., 2019). When operating in this direction, it is often termed pyruvate synthase and this enzyme/activity is required for three of the six known autotrophic pathways of CO₂ fixation including (i) the reductive acetyl CoA pathway [Wood-Ljungdahl pathway (Furdui and Ragsdale, 2000; Fuchs, 2011)], (ii) the reverse tricarboxylic acid (rTCA) cycle (Fuchs, 2011) and the more recently discovered functional reversal of the oxidative TCA cycle, termed the reverse oxidative TCA cycle or roTCA (Mall et al., 2018; Nunoura et al., 2018), and (iii) the dicarboxylate/hydroxybutyrate (DC/HB) cycle (Fuchs, 2011). In the anaerobe *Desulfurella acetivorans* when growing via the roTCA cycle, the directionality of PFOR is dictated by the availability of organic carbon. When acetate is present, heterotrophic growth (pyruvate oxidation) is favored whereas autotrophic growth is favored in its absence (Mall et al., 2018; Nunoura et al., 2018). The distribution and abundance of Type II (D) *Meiothermus* MAGs in

anoxic waters in BA1A and in other waters impacted by serpentinization (discussed above), when combined with these MAGs uniquely encoding PFOR, oxidative [NiFe]-hydrogenase enzyme homologs that are typically involved in anaerobic respiration, and full TCA cycles that could potentially be reversed (e.g., roTCA) points to the intriguing possibility that these organisms are capable of anaerobic growth and/or autotrophy. Additional physiological and/or cultivation experiments are needed to further evaluate the plausibility that Type II (D) *Meiothermus* MAGs are facultative anaerobes (explanation 4) and possibly capable of autotrophic growth.

The recovery of Type II (D) MAGs in near-surface, less reacted aquifer waters (0–30 m, 41–65 m) and Type I (S) MAGs in deep, more reacted aquifer waters (108–132 m) suggested the possibility of dispersal between aquifers (explanation 5) and prompted a population-level analysis to examine the extent of gene flow/recombination among MAGs. Surprisingly, few shared SNVs with the same base substitution were identified among pairs of MAGs from different clades that co-occurred in the same depth. Interestingly, shared SNVs with the same base substitution were more prevalent among MAGs from different clades from differing depths. These shared SNVs could not be accounted for based on convergent (random) evolutionary processes alone indicating a potential role for limited gene flow/recombination among depth stratified populations despite chemical and physical barriers to dispersal (Figure 4). The role of gene flow/recombination, as opposed to shared ancestry, is potentially bolstered by the observation that contigs from MAG pairs where shared SNVs are located had, on average, significantly higher ANI than those contigs that did not share SNVs. While signatures of viruses were not evident in BA1A MAGs, several classes of IMGEs including the insertion sequences

ISplu6 and the composite transposon Tn125 were detected thereby providing a potential mechanistic explanation for the observed but limited gene flow/recombination between *Meiothermus* populations comprising different clades. Intriguingly, the pattern of sequence identities among IMGs, including that of Tn125, do not recapitulate the evolutionary history of the *Meiothermus* MAGs from which they derive. This provides further evidence for relatively recent gene flow/recombination between depth-stratified *Meiothermus* populations belonging to the same clade and, to a lesser extent, members of the different clades among depth intervals. The time scales over which gene flow/recombination occurred among spatially segregated *Meiothermus* is expected to be much shorter than the estimated residence times of Type II waters in the Samail Ophiolite [$>20,000$ years (Paukert Vankeuren et al., 2019)] based on the amount of time since they were in contact with the atmosphere. Such differences are likely a consequence of geochemical measurements made on bulk fluids as opposed to the resolution provided by the sensitive genomic techniques utilized herein.

Taken together, these findings suggest chemical and physical gradients in the subsurface of the Samail Ophiolite (or a geologically similar precursor) facilitated parapatric speciation of *Meiothermus* (Figure 4), a process where co-inhabiting populations speciate to inhabit specific ecological niches, with limited gene exchange between them (Butlin et al., 2008). In the case of *Meiothermus* in Oman, it is suggested that chemical variation generated by serpentinization creates opportunities for spatial variation between populations to develop (a form of geographic isolation), and this spatial variation of populations represents the starting point for further genetic divergence and speciation. As serpentinization of host rock progresses, the chemistry of deep fluids and near surface fluids further diverge, allowing for and/or promoting additional genetic

divergence. Concomitantly, the porosity of host rocks (e.g., dunite, harzburgite), which likely become even less porous as serpentinization reactions progress (Klein and Le Roux, 2020), likely limits further dispersal and gene exchange. In this scenario, spatially and ecologically fragmented *Meiothermus* populations adapted to local environmental conditions and thereby independently accumulated distinct mutations, single nucleotide variants, and as illustrated above, differences in encoded functionalities. Parapatric speciation is thus likely to be a prominent mode of evolution in the Samail Ophiolite and other active subsurface geologic systems that have limited spatial connectivity and whose physical and chemical features are in actively evolving or changing.

Data Availability Statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ Supplementary material.

Author Contributions

DN and EB collected the samples. AT, JM, and EB designed the experiment. EF extracted DNA. MM-E conducted all bioinformatics analyses. EB and MM-E wrote the manuscript with help from other co-authors. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by grants from NASA (NNA15BB02A and 80NSSC21K0489) to AT and EB. Drilling and research in the Oman Drilling Project were supported by the Alfred P. Sloan

Foundation (in association with the Deep Carbon Observatory, DCO), the International Continental scientific Drilling Program (ICDP), US National Science Foundation (NSF) Research Grant NSF-EAR-1516300, the Japanese Marine Science and Technology Center (JAMSTEC), grant number 16H06347 from the Japanese Society for the Promotion of Science (JSPS), the US National Aeronautics and Space Administration (Rock Powered Life NASA Astrobiology Institute NNA15BB02A), the European Science Foundation, the German Science Foundation, the Swiss Science Foundation, and the International Ocean Discovery Program (IODP), with contributions from the Sultanate of Oman Ministry of Regional Municipalities and Water Resources, the Oman Public Authority of Mining, Sultan Qaboos University, CNRS, University of Montpellier, Columbia University of New York, and the University of Southampton.

Acknowledgments

We are grateful to the Sultanate of Oman Ministry of Regional Municipalities and Water Resources for allowing sampling and export of well waters. We are grateful to Philippe Gouze, Delphine Roubinet, Richard Leprovost, and Gérard Lods of CNRS, University of Montpellier for operating the packer system in 2018 and Amelia Paukert Vankeuren and Martin Stute for help with sample collection in 2019. We are grateful for two independent reviewers whose comments greatly improved this manuscript.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

CHAPTER 3 CONCLUSIONS AND FUTURE STUDIES:

Conclusions:

This dissertation aims to investigate how microbial dispersal and gene flow impact the evolution of microbial populations. Microbial populations are capable of undergoing gene flow even when dispersal is impossible. When dispersal is impossible or limited but gene flow can still occur, an understudied ecological and evolutionary scenario occurs that has implications for community structure, selection, and the surrounding environment. While few microbial systems have been studied to the extent where gene flow and dispersal are well characterized, it seems that most microbial metapopulations have some degree of disconnect between dispersal and gene flow. This dissertation explores the impacts of dispersal and gene flow by investigating three microbial communities from different depths of well BA1A in the subsurface of the Samail Ophiolite. Community composition differences between the three different depths of BA1A imply substantial barriers to dispersal, which were initially expected to limit gene flow as well. Yet, we were able to identify substantial gene flow between six populations of *Meiothermus*. While the possibility remains that cryptic dispersal events are contributing to this gene flow, the replete viral and mobile genetic elements present in the system are likely responsible for the bulk of the gene flow. The gene flow patterns seem to be contributing to disruptive selection leading to the continuing speciation of the two *Meiothermus* clades into a deep and shallow niche. Without this gene flow, it seems possible that one or more of these populations would introgress, i.e., be absorbed into the larger population. While it is unknown how other members of the community are affected, this pattern of gene flow across a barrier to dispersal appears to be a contributing factor to the speciation of *Meiothermus* in Oman.

We first hypothesized that the physical and chemical barriers separating the three microbial communities would constitute a barrier to dispersal, and that this barrier would be reflected in heterogeneous community composition. Previous studies in the region characterized the microbial communities of multiple different well sites, encompassing the extent of the different fluid chemistries and geologies in the Samail ophiolite subsurface. Communities inhabiting similar environments, e.g., similar pH and fluid chemistries, often exhibited different community structures. If there were minimal barriers to dispersal between these sites, then the similar environmental characteristics should select upon the wide range of immigrating populations to converge on homogenous community structures. Based on this prior evidence, there appeared to be barriers to dispersal across wells. We thusly expected to also see these barriers to apply within wells.

Reconstruction of each microbial community revealed three distinct microbial communities, while characterizations of the hydrology revealed two different geochemical regimes. The community occupying the 108 – 132 meter depth was substantially different from the two shallower communities. This was unsurprising since there are substantial environmental differences between the 108 – 132 m depth and the two shallower depths. However, the 0 – 30 m and 31 – 45 m communities are also quite different from one another, despite these sites having very similar fluid chemistry. Assuming these two communities are not experiencing differential selection pressures due to some uncaptured environmental variable, then the community level differences evidence a barrier to dispersal. If substantial dispersal was possible, we would expect to see homogenization of communities inhabiting environmentally similar sites (Vieira et al. 2022), as microbial populations disperse between sites and the similar local selection pressures

of different sites constrains a wide immigrating pool into a homogenized, highly adapted community archetype.

We secondly hypothesized that gene flow could still occur between these separated populations despite barriers to dispersal since the range of horizontal gene transfer mechanisms should allow DNA to cross temporal or physical barriers that the microbial populations themselves cannot cross. We evaluated if gene flow was occurring by identifying and comparing single nucleotide variants (SNVs) shared between each population of *Meiothermus*. Phylogenetic parsimony rules and probabilistic models were used to determine which SNV pairings were due to gene flow, instead of random convergence or inheritance from a common ancestor. In total, we identified a substantial number of shared SNVs across spatially segregated populations, including across the two clades of *Meiothermus*, that could be reasonably attributed to gene flow.

Multiple mechanisms which may have facilitated this gene flow were identified. Many viral sequences were identified throughout the metagenomes, both associated with or outside of the reconstructed microbial genomes. These viral sequences came from a wide taxonomic range, and many are known to facilitate transduction in other settings. With sequence data alone, it is difficult to link hosts and viruses, but given the bulk and diversity of viral sequences recovered it seems reasonable to assume that transduction is an active force shaping the evolution of many of these microbial populations. We also identified mobile genetic elements integrated into the *Meiothermus* chromosomes. Many of these elements appear intact and functional and have likely mediated recombination events in the recent evolutionary past. Paired with a mechanism of horizontal gene transfer, like a transformosome or virus, these elements would be able to integrate the exogenous DNA directly into the *Meiothermus* chromosome, preserving the

horizontal gene transfer events for generations. Most of the large mobile genetic elements identified contained at least one shared SNV, confirming our suspicions that these elements are likely involved in the gene flow we explored. Critically, one of these elements, Tn125 exhibited a unique sequence similarity pattern indicating two transfer events of over 10,000 basepairs of genetic material. The phylogenetic structure of the Tn125 sequences as well as proximal genes implies a substantial recombination and gene flow event occurring after the split of the founder and original colonizing *Meiothermus* population into the two extant clades.

We finally hypothesized that this unique dispersal structure would impact the evolution of the two *Meiothermus* clades. These clades have distinct genetic and genomic adaptations that aid in their specific niche. For example, the deep clade, which is more abundant at depth, has a streamlined genome. A smaller genome is less costly to maintain and replicate, which is preferable in the more nutrient poor fluids this population inhabits. In contrast, the shallow clade has a much faster genome replication time, preferable for its comparatively nutrient replete habitat. Additionally, each clade has many unique proteins constituting local adaptations, like the thiamine biosynthesis pathway unique to the deep clade, which would allow for the synthesis of thiamine when none can be found in the local environment.

The SNV profiles of the *Meiothermus* populations provide strong evidence of disruptive selection between these two clades. We identified gene flow between nearly every *Meiothermus* population except for those that share a habitat. This was interpreted as a strong disruptive selection pressure eliminating hybrid genotypes, accelerating the speciation of these two clades, and driving each further into its respective niche. Given the environmental structure, and how it impacts gene flow and dispersal, this constitutes an example of parapatric speciation. The degree

to which the unique environmental structure, i.e., one that limits dispersal but allows gene flow, would've impacted this speciation is not exactly known, but it is assumed to have contributed to or accelerated it. In the two shallow sample sites, the assumed population size of each clade is roughly similar, but in the deep site, the deep clade is over an order of magnitude more abundant. In the absence of gene flow, the shallow clade likely would've not been able to genetically distinguish itself from the deep clade, and possibly would've introgressed, i.e., genetically merged, into the larger population.

The dispersal structure of microbial communities also impacts their surrounding environment. For example, these findings imply a greater degree of fluid connectivity in the Samail ophiolite than previously thought. Previous estimations of fluid residence times based on isotope fractions (Lods, Roubinet, Matter, Leprovost, and Gouze 2020) imply long residence times and limited fluid connectivity, but this work provides a genomic indication that fluids bearing DNA, viruses, or possibly even cells, are exchanging between microbial communities separated by meters of rock. The findings of this work are evidence that genomic approaches can inform hydrological questions as well.

While the scope of this study is limited to the Samail Ophiolite, the ecological and evolutionary paradigms observed here are likely applicable to other environments in which inhabitant microbial communities face barriers to dispersal. However, owing to the recent development of many of these methods, this type of investigation has not been performed on any other system. While metagenomics has been a standard approach for years, 16s based approaches are still used in many cases. While 16s based approaches are advantageous in that they require less DNA and are easier to evaluate, they preclude the ability to investigate gene flow. Given the

findings of this thesis, studies using 16s based approaches to evaluate the connectivity of microbial systems should be aware that they can only determine the degree of microbial dispersal, not the degree of microbial gene flow. Additionally, the ever improving quality of metagenomic sequences has improved our ability to reliably identify single nucleotide variants. This process is highly contingent on having high quality sequence data, and as such, becomes more accurate with each improvement to the entire metagenomics pipeline. To the best of our knowledge, the probability based model presented in Munro-Ehrlich et al is the first application of a statistical model using the observed number of SNVs versus the expected number of SNVs to assess gene flow. In sum, the methods used to characterize the Samail Ophiolite microbial communities are comparatively novel, and there is a significant opportunity in retroactively applying these methods to other well characterized microbial systems. These findings encourage a re-evaluation of other systems where community structure alone was used to draw conclusions regarding the connectivity of separated microbial communities.

Future Directions

This work investigated the unique scenario of gene flow across dispersal barriers between populations of *Meiothermus* inhabiting subsurface fluids in the Samail Ophiolite. This gene flow is presumed to be intensifying speciation and local adaptation, amongst other impacts. To

determine if these consequences are unique to *Meiothermus*, the methods utilized in this work should be applied firstly to other populations in the Samail ophiolite, and then to populations in other systems. For example, Fones et al 21 (Fones et al. 2021), identified two clades of Methanobacteria in the Samail ophiolite that are evolutionary and phenotypically distinct. Importantly, the clades exhibit local adaptation, including a unique formate dehydrogenase which should reduce the carbon requirement of that archaeon (Fones et al. 2021). An SNV based approach could provide insight into the potential origin of this adaptation, as well as provide a clearer insight into the degree of disruptive selection presumed to be driving these two clades apart.

The appendices of this thesis investigate three additional microbial systems: the human gut microbiome, Yellowstone hotsprings, and the Great Salt Lake. Given the quality metagenomic sequence data available from each of these systems, it should be possible to investigate these systems with the same gene flow analysis applied to Samail ophiolite microbial communities. Investigating gene flow dynamics in hotsprings or in the Great salt lake would help contextualize the findings presented in this thesis. Those two systems are also environmental systems, but are different in that there are different barriers to dispersal. For example, communities distributed around the Great Salt Lake may have few barriers to dispersal. Moreover, these environments host different microbial taxa than those in Oman. In applying the same gene flow approach to these systems, we could further our understanding of how the environment impacts gene flow and evolution, and the degree to which those findings are consistent across taxa or habitat.

Studying community dynamics in the human gut would be uniquely valuable given the transience and volatility of the microbial communities in that system. The human gut microbiome experiences substantial directional immigration and emigration and would provide an interesting foil to the Samail Ophiolite communities which are likely very dispersal limited. Similarly, the gut is a nutrient replete system. As such, communities there face very different environmental pressures than those in the Samail ophiolite. Arguably, the gut microbiome is clear foil to the Samail ophiolite communities. An investigation into the gene flow within human gut microbiome communities would be very effective in contextualizing the findings discussed here, and would have the bonus of being pertinent to human health.

Moreover, the findings of this work imply that gene flow is possible through the rock separating fluid hosted microbial communities, but the question remains as to how this DNA is able to move through the rock. Previous work in the Samail ophiolite has identified cells in rock interiors (Templeton et al. 2021). In one sample site biologically mediated transformation of carbon compounds was observed, presumably via the activities of methanogens inhabiting rock interiors (Fones et al. 2022). In the case of Fones et al 22, the rock was incubated in media and organic acids with radiolabeled carbon. In the biological condition, the carbon compounds were transformed, presumably due to the metabolism of endogenous methanogens. However, while rock hosted cells appear metabolically active under certain conditions, it is yet unknown whether those endolithic cells would be able to facilitate gene flow between fluid hosted communities. It is also unknown if they are dispersing members of those communities or if they are trapped. Ongoing efforts have been undertaken in the Boyd lab to further investigate the possibility of active endolithic microbial communities. These efforts entail incubating 16 rock cores from

different depths in three different wells. The cores were selected to recapitulate the hydrological and geological diversity of the Samail Ophiolite. Using a radio-isotope based approach analogous to Fones et al 22 (Fones et al. 2021), these rocks will be incubated in media containing tritiated water, and the quantity of tritium uptake will be measured under biotic and abiotic conditions to estimate biological activity. Tritiated water is metabolism agnostic; most cells incorporate hydrogen from water during lipid synthesis, and all cells can incorporate tritium from isotopic exchange with hydrogen (Zhang, Gillespie Al Fau - Sessions, and Sessions). Thus, this approach is expected to give an indication of which sites host the most active communities. This should also identify conditions influencing the habitability of the sample sites, like porosity, mineral composition, or the pH of the surrounding fluids. The results of this work will inform further investigations; We will attempt to extract DNA from the most active communities to infer community composition. We can then compare the community composition of the rock interiors to the surrounding fluids to gain insight into dispersal dynamics in the system; If microbes are able to freely disperse through rock interiors, then the rock interior community should resemble the community of the surrounding fluids. Further work will entail attempts to identify the possibility for gene flow without dispersal in this system, including attempts to determine the permeability of these rocks to DNA, both exogenous and in viral particles.

REFERENCES CITED

- Alderliesten, Jesse B., Sarah J. N. Duxbury, Mark P. Zwart, J. Arjan G. M. de Visser, Arjan Stegeman, and Egil A. J. Fischer. 2020. 'Effect of donor-recipient relatedness on the plasmid conjugation frequency: a meta-analysis', *BMC Microbiology*, 20: 135.
- Barnes, M. A., Christopher L. Turner Cr Fau - Jerde, Mark A. Jerde Cl Fau - Renshaw, W. Lindsay Renshaw Ma Fau - Chadderton, David M. Chadderton Wl Fau - Lodge, and D. M. Lodge. 'Environmental conditions influence eDNA persistence in aquatic systems'.
- Burns, Adam R., Elizabeth Miller, Meghna Agarwal, Annah S. Rolig, Kathryn Milligan-Myhre, Steve Seredick, Karen Guillemin, and Brendan J. M. Bohannon. 2017. 'Interhost dispersal alters microbiome assembly and can overwhelm host innate immunity in an experimental zebrafish model', *Proceedings of the National Academy of Sciences*, 114: 11181-86.
- Custer, Gordon F., Luana Bresciani, and Francisco Dini-Andreote. 2022. 'Ecological and Evolutionary Implications of Microbial Dispersal', *Frontiers in microbiology*, 13.
- Diop, Awa, Ellis L. Torrance, Caroline M. Stott, and Louis-Marie Bobay. 2022. 'Gene flow and introgression are pervasive forces shaping the evolution of bacterial species', *Genome Biology*, 23: 239.
- Dodsworth, Jeremy A., Lei Li, Shiping Wei, Brian P. Hedlund, John A. Leigh, and Paul de Figueiredo. 2010. 'Interdomain Conjugal Transfer of DNA from Bacteria to Archaea', *Applied and Environmental Microbiology*, 76: 5644-47.
- Fones, Elizabeth M., Daniel R. Colman, Emily A. Kraus, Ramunas Stepanauskas, Alexis S. Templeton, John R. Spear, and Eric S. Boyd. 2021. 'Diversification of methanogens into hyperalkaline serpentinizing environments through adaptations to minimize oxidant limitation', *The ISME Journal*, 15: 1121-35.
- Fones, Elizabeth M., Alexis S. Templeton, David W. Mogk, and Eric S. Boyd. 2022. 'Transformation of low-molecular-weight organic acids by microbial endoliths in subsurface mafic and ultramafic igneous rock', *Environmental Microbiology*, 24: 4137-52.
- Fournier, Gregory P., and J. Peter Gogarten. 2008. 'Evolution of Acetoclastic Methanogenesis in Methanosarcina via Horizontal Gene Transfer from Cellulolytic Clostridia', *Journal of Bacteriology*, 190: 1124.
- Guo, Li, Kim Brügger, Chao Liu, A. Shah Shiraz, Huajun Zheng, Yongqiang Zhu, Shengyue Wang, K. Lillestøl Reidun, Lanming Chen, Jeremy Frank, David Prangishvili, Lars Paulin, Qunxin She, Li Huang, and A. Garrett Roger. 2011. 'Genome Analyses of

- Icelandic Strains of *Sulfolobus islandicus*, Model Organisms for Genetic and Virus-Host Interaction Studies', *Journal of Bacteriology*, 193: 1672-80.
- Hall, Rebecca J., Fiona J. Whelan, James O. McInerney, Yaqing Ou, and Maria Rosa Domingo-Sananes. 2020. 'Horizontal Gene Transfer as a Source of Conflict and Cooperation in Prokaryotes' <https://www.frontiersin.org/articles/10.3389/fmicb.2020.01569>, *Frontiers in microbiology*, 11: 2020-July-17 Perspective HGT in prokaryotes.
- Hoetzing, Matthias, Alexandra Pitt, Andrea Huemer, and Martin W. Hahn. 2021. 'Continental-Scale Gene Flow Prevents Allopatric Divergence of Pelagic Freshwater Bacteria', *Genome Biology and Evolution*, 13: evab019.
- Hu, Yongfei, George F. Gao, and Baoli Zhu. 2017. 'The antibiotic resistome: gene flow in environments, animals and human beings', *Frontiers of Medicine*, 11: 161-68.
- Koonin, Eugene, and Yuri Wolf. 2012. 'Evolution of microbes and viruses: a paradigm shift in evolutionary biology?', *Frontiers in Cellular and Infection Microbiology*, 2.
- Legendre, M., A. Lartigue, L. Bertaux, S. Jeudy, J. Bartoli, M. Lescot, J. M. Alempic, C. Ramus, C. Bruley, K. Labadie, L. Shmakova, E. Rivkina, Y. Couté, C. Auid-Orcid Abergel, and J. M. Claverie. 'In-depth study of Mollivirus sibericum, a new 30,000-y-old giant virus infecting *Acanthamoeba*'.
- Lennon, Jt Auid-Orcid, M. E. Muscarella, S. A. Placella, and B. K. Lehmkuhl. 'How, When, and Where Relic DNA Affects Microbial Diversity. LID - 10.1128/mBio.00637-18 [doi] LID - e00637-18'.
- Lods, Gérard, Delphine Roubinet, Jürg M Matter, Richard Leprovost, Philippe Gouze, and Oman Drilling Project Science Team. 2020. 'Groundwater flow characterization of an ophiolitic hard-rock aquifer from cross-borehole multi-level hydraulic experiments', *Journal of Hydrology*, 589: 125152.
- Lods, Gérard, Delphine Roubinet, Jürg M. Matter, Richard Leprovost, and Philippe Gouze. 2020. 'Groundwater flow characterization of an ophiolitic hard-rock aquifer from cross-borehole multi-level hydraulic experiments', *Journal of Hydrology*, 589: 125152.
- Maddamsetti, R. 'Gene flow in microbial communities could explain unexpected patterns of synonymous variation in the *Escherichia coli* core genome'.
- Mallet, J. 2008. 'Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363: 2971-86.

- Merino, Nancy, Tracie R. Jackson, James H. Campbell, Annie B. Kersting, Joshua Sackett, Jenny C. Fisher, James C. Bruckner, Mavrik Zavarin, Scott D. Hamilton-Brehm, and Duane P. Moser. 2022. 'Subsurface microbial communities as a tool for characterizing regional-scale groundwater flow', *Science of The Total Environment*, 842: 156768.
- Munro-Ehrlich, M., D. B. Nothaft, E. M. Fones, J. M. Matter, A. S. Templeton, and E. S. Boyd. 'Parapatric speciation of *Meiothermus* in serpentinite-hosted aquifers in Oman'.
- Muthukumar, M. 'Mechanism of DNA transport through pores'.
- Neu, Josef. 2021. 'Chapter 1 - Evolutionary microbiology.' in Josef Neu (ed.), *The Microbiome in Prenatal and Neonatal Life* (Academic Press).
- O'Malley, Maureen A. 2008. "Everything is everywhere: but the environment selects': ubiquitous distribution and ecological determinism in microbial biogeography', *Studies in history and philosophy of biological and biomedical sciences*, 39: 314-25.
- Pan, Donald, Jason Nolan, Kenneth H. Williams, Mark J. Robbins, and Karrie A. Weber. 2017. 'Abundance and Distribution of Microbial Cells and Viruses in an Alluvial Aquifer', *Frontiers in microbiology*, 8.
- Paukert Vankeuren, Amelia N., Jürg M. Matter, Martin Stute, and Peter B. Kelemen. 2019. 'Multitracer determination of apparent groundwater ages in peridotite aquifers within the Samail ophiolite, Sultanate of Oman', *Earth and Planetary Science Letters*, 516: 37-48.
- Rempfert, Kaitlin R., Hannah M. Miller, Nicolas Bompard, Daniel Nothaft, Juerg M. Matter, Peter Kelemen, Noah Fierer, and Alexis S. Templeton. 2017. 'Geological and Geochemical Controls on Subsurface Microbial Life in the Samail Ophiolite, Oman', *Frontiers in Microbiology*, 8: 56.
- Russell, M. J., A. J. Hall, and W. Martin. 2010a. 'Serpentinization as a source of energy at the origin of life', *Geobiology*, 8: 355-71.
- Russell, MJ, AJ Hall, and W Martin. 2010b. 'Serpentinization as a source of energy at the origin of life', *Geobiology*, 8: 355-71.
- Soler, N. Auid-Orcid, and P. Forterre. 'Vesiduction: the fourth way of HGT'.
- Stanojković, Aleksandar, Svatopluk Skoupý, Pavel Škaloud, and Petr Dvořák. 2022. 'High genomic differentiation and limited gene flow indicate recent cryptic speciation within the genus *Laspinema* (cyanobacteria)', *Frontiers in microbiology*, 13.
- Templeton, Alexis S., Eric T. Ellison, Clemens Glombitza, Yuki Morono, Kaitlin R. Rempfert, Tori M. Hoehler, Spencer D. Zeigler, Emily A. Kraus, John R. Spear, Daniel B. Nothaft,

- Elizabeth M. Fones, Eric S. Boyd, Mason Munro-Ehrlich, Lisa E. Mayhew, Dawn Cardace, Juerg M. Matter, Peter B. Kelemen, and Party the Oman Drilling Project Science. 2021. 'Accessing the Subsurface Biosphere Within Rocks Undergoing Active Low-Temperature Serpentinization in the Samail Ophiolite (Oman Drilling Project)', *Journal of Geophysical Research: Biogeosciences*, 126: e2021JG006315.
- Vieira, Helena Henriques, Inessa Lacativa Bagatini, Guilherme Pavan de Moraes, Roberta Mafra Freitas, Hugo Sarmento, Stefan Bertilsson, and Armando Augusto Henriques Vieira. 2022. 'Regional factors as major drivers for microbial community turnover in tropical cascading reservoirs', *Frontiers in microbiology*, 13.
- Virolle C Fau - Goldlust, Kelly, Sarah Goldlust K Fau - Djermoun, Sarah Djermoun S Fau - Bigot, S. Bigot, and C. Lesterlin. 'Plasmid Transfer by Conjugation in Gram-Negative Bacteria: From the Cellular to the Community Level. LID - 10.3390/genes11111239 [doi] LID - 1239'.
- Volz, Erik M., Katia Koelle, and Trevor Bedford. 2013. 'Viral Phylodynamics', *PLOS Computational Biology*, 9: e1002947.
- Warwick-Dugdale, Joanna, Natalie Solonenko, Karen Moore, Lauren Chittick, Ann C. Gregory, Michael J. Allen, Matthew B. Sullivan, and Ben Temperton. 2019. 'Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands', *PeerJ*, 7: e6800-e00.
- Weitz, Joshua S., and Steven W. Wilhelm. 2012. 'Ocean viruses and their effects on microbial communities and biogeochemical cycles', *FI000 biology reports*, 4: 17-17.
- Williams, L. E., and J. J. Wernegreen. 'Sequence context of indel mutations and their effect on protein evolution in a bacterial endosymbiont'.
- Zamani Dahaj, Seyed Alireza. 2015. "revised-thesis-zamani."
- Zhang, X., Alex L. Gillespie Al Fau - Sessions, and A. L. Sessions. 'Large D/H variations in bacterial lipids reflect central metabolic pathways'.