Intra-species Genomic Variations in Hydrogen-oxidising Bacteria is Delineated by Horizontal Gene Flow

A thesis

submitted in fulfilment

of the requirements for the degree

of

Doctor of Philosophy in Biological Sciences

at

The University of Waikato

by

GAMALIEL LYSANDER (G.L.) BENIG CABRIA

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2024

Statement of Authorship

I hereby declare that this is a submission of my work, and to the best of my knowledge does not contain material that has been previously by another individual, nor has it been used for awarding of another degree.

Acknowledgements

First, I would like to give thanks to my supervisors, Charles Lee, Craig Cary, Matt Stott and Huw Richards, who kept me in line and did not stop supporting me even through the troubling times of COVID-19 and being locked out of the country. Thank you, Charlie, for the opportunity to learn and better myself as a scientist and a person. I will always be grateful for all the wisdom you've shared, be it about science or otherwise. Thank you, Matt, for sharing your knowledge on science, the maori people and their culture. I am always fascinated by how you bridge science and the people you work with through a deep understanding of their culture. Huw, thank you for your company these past five years. Your calming presence and wisdom help me be grounded and composed whenever I am troubled, scientifically and emotionally. Thanks for being my pseudo-therapist. Craig, you will always be an inspiration to me as I pursue science. Your unending enthusiasm to learn and explore inspires me to go over the hill and learn more. Thanks for all the ideas for research you always share. You will be deeply missed.

This thesis manuscript would not be possible without my dear colleagues and friends in the Thermophile Research Unit, who provided technical and logistical support. I am always grateful for all the assistance Charlotte, John and Lynn gave. More hands make work a little easier. And I owe a huge thanks to Roanna Richards-Babbage, who is always there day or night when I am in trouble in the lab.

To all my friends who always support me in dire need, who laugh when I am in pain, who provide distraction to calm my worries, '*Dakulaong pasasalamat*' to Maria, Anya, Shir, Nauv, Shaun, Nav, Emily, Jacinda, Georgia and Georgia. Thank you too to my buddy who did not made me feel alone through the struggles of this past few months,خیلی لطف دارید Jaber! Thank you too to all the people I've met that make this journey through TRU more colourful: Mafalda, Alexis, Matt, Jen! Thank you too to all the Filipinos who I've met in my journey here in New Zealand whom fed me and made me feel at home. Thank you Ate Karen, Kuya Toti and Nina!

Lastly I dedicate this thesis to my mum and dad, my brother and sister and nephew Aslan Bo whom without their support and sacrifice I won't be here. You were always there when I needed you most.

To all who are asking why your name isn't here. I apologize but I have limited time. If you grumbled that your name isn't here while reading this, then you know I am also thankful for your support.

Padagos!-

Abstract

Species are the fundamental unit of ecology and evolutionary biology. However, there is no universal definition of a species that spans all domains of life. Eukaryotic species are traditionally defined through reproductive isolation, where reproductive barriers result in adaptive traits being shared only within the isolated population. Therefore, discontinuities in gene flow in eukaryotes generally result in the formation of an ecologically and genotypically cohesive unit. In contrast, delineating asexual bacterial species is often hampered by extensive horizontal gene flow and is further confounded by the difficulty of identifying a bacteria's ecological niche. Therefore, the working definition of a prokaryotic species is based on arbitrary standards of relatedness and often lacks a cohesive ecological niche. To address this problem, the ecological population theory, which defines bacterial 'species' as ecologically distinct populations delineated through discontinuities in total gene flow, was proposed. It postulates that, rather than a hindrance, the prevalence of horizontal gene flow, primarily through homologous recombination, plays a significant role in bacterial speciation. Horizontal gene flow allows gene-specific sweeps in bacterial populations, where adaptive genes are shared preferentially between recombining bacteria independent of their genomic background. Thus, discontinuities in horizontal gene flow can be utilised to demarcate a genotypically and ecologically cohesive bacterial species.

Previous studies have demonstrated that ecological populations exist in heterotrophic bacteria. However, examining the generalisability of the ecological population theory requires understanding the role of horizontal gene flow in delineating populations from extreme bacterial communities, especially bacteria with limited metabolic redundancy due to their higher impact of deleterious consequences (e.g., pathway dysregulation by recombination of a maladaptive gene) during horizontal gene flow. Therefore, in this work, I utilised the chemolithoautotrophic hydrogen-oxidising *Aquificota* bacteria from natural populations as a model system to test the generalisability of the ecological population theory. Utilising a novel high-throughput isolation method, I isolated co-habiting *Aquificota* bacteria with minimal enrichment from geothermal springs. Sequencing and comparative analysis of the *Aquificota* genomes reveal intra-species genomic variations driven by varying degrees of horizontal gene flow and a population structure composed of co-occurring clonal lineages and recombinogenic bacteria. To examine whether these *Aquificota* populations exhibit ecological and evolutionary patterns akin to eukaryotic species, I investigated the changes in their population structure across four additional time points (1 year). Analysis of recent horizontal gene flow and mutations in *Aquificota* populations show that clonal lineages remain clonal across time, with minimal changes in their genomic background diversity. In contrast, recombining populations experience variation in intra-population diversity and a shift in the dominant genomic background across time—implying an

intra-population turnover. Lastly, I also investigated the role of both non-homologous and homologous recombination in maintaining ecological populations in *Aquificota* by analysing the presence and function of mobile genetic elements (MGEs) in its different populations. The lack of spatial and functional overlap between MGEs and horizontally swept regions confirms that homologous recombination drives horizontal gene flow in Aquificota populations.

This work demonstrated that ecological populations in *Aquificota* can be delineated through discontinuities in horizontal gene flow driven primarily by homologous recombination. Thus, *ecological population* theory can be a generalisable model for speciation and evolution in bacteria.

Table of Contents

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

1.1. Introduction

Species are the fundamental unit of ecology and evolutionary biology. Defining species is relatively simpler for sexual organisms, such as plants and animals, since barriers to reproduction are relatively easily observed, and gene flow almost exclusively occurs from parents to offspring (Balakrishnan, 2005; Hull, 1977; Mayr, 1999; Queiroz, 2005a; Shapiro et al., 2009; Staley, 2006). However, applying the same criteria to asexual prokaryotes is problematic due to the difficulty in consistently observing genotypic and phenotypic traits and extensive horizontal gene flows between even distantly related bacterial 'species' (Doolittle, 2012; Lan and Reeves, 2000; Queiroz, 2005b). This has led to arbitrary definitions of bacterial 'species' or taxonomic units, which often do not reflect the taxa's ecological and evolutionary history (Stackebrandt and Goebel, 1994). The lack of an evolutionarily and ecologically coherent definition of prokaryotic species has also hindered the application of many ecological and evolutionary theories developed for metazoans on bacterial communities (e.g., habitat fragmentation, migration, and dispersal theories) (Oakley et al., 2010).

Despite a clear need for an evolutionarily and ecologically cohesive definition of bacterial species, numerous challenges exist, and a universally accepted solution has yet to be identified. In sexual organisms, the most widely used definition demarcates species based on reproductive barriers (i.e., the absence of gene flow) between organisms (Mayr, 1999; Queiroz, 2005a); a similar metric should be used to delineate bacterial species to ensure consistency. Rather than being a problem, horizontal gene flow could provide the solution. Demarcating bacterial populations based on disruption in total gene flow, including horizontal gene transfers, allows bacterial populations to be split into groups that preferentially share genetic information (Cordero and Polz, 2014; Shapiro and Polz, 2015, 2014). These genetically isolated groups are akin to reproductive isolation, resulting in ecologically and evolutionarily cohesive species in metazoan eukaryotes. Thus, ecological and evolutionary theories can be utilised for these populations.

Here, I will explore further the fundamentals of the most widely used species definition, the Biological Species Concept, and why aligning asexual prokaryotes with this definition is problematic. I will provide an overview of historical and current popular methods for demarcating a 'species' in asexual prokaryotes and their problems. I will then discuss the alternative theories proposed to resolve this issue and the role of horizontal gene flow as a solution to demarcate an ecologically and evolutionarily cohesive species in bacteria. Afterwards, I will explore the recent advancements in sequencing and computational genomics that allow the evaluation of alternatives in defining a bacterial species. Lastly, I will discuss how we can

examine the general applicability of these alternative theories in demarcating an ecologically and evolutionarily cohesive species from natural bacterial communities.

1.2. The Biological Species Concept: Species as a Biological Unit

Ernst Mayr established the most widely accepted definition of species as "groups of interbreeding natural populations which are reproductively isolated from other similar groups" (Mayr, 1999; Queiroz, 2005a; Rieseberg et al., 2004). Mayr's concept, named the Biological Species Concept (BSC), highlights reproductive isolation as the defining trait of a species - implying that advantageous mutations can spread within a species without affecting other similar coexisting species (Queiroz, 2005a; Shapiro and Polz, 2015). Therefore, species in the BSC are characterised based on gene flow and the barriers to gene flow that demarcate species boundaries (Mayr, 1954; Rieseberg et al., 2004). The limitation of gene flow to only reproductively isolated groups allows adaptive genes to be preferentially shared within that group, resulting in the coupling of the genotype and ecological niche of all the members within a species (Mayr, 1999, 1954). Linking an organism's ecological niche to its genotypic characteristics roots the BSC on a biologically meaningful definition—unlike the preceding definitions equating species only as arbitrary rankings based on their overall similar properties (Queiroz, 2005b). An ecological and genotypically coherent species allows it to become the fundamental unit in ecology, evolution, and other fields of biology.

Understanding gene flow is essential in understanding species in the BSC (Mayr, 1954; Rieseberg et al., 2004). In the BSC, species are metapopulations or metapopulation lineages, meaning that species are composed of several distinct spatially separated populations linked through migration and interbreeding (i.e. gene flow) (Mayr, 1999, 1975; Queiroz, 2005b). With metapopulations connected through gene flow, a genetic change within a member population can still affect other distant populations due to their connection to the species' shared gene pool (Queiroz, 2005a). However, differences in the gene flow strength directly affect each population's and species' evolutionary fate. A strong gene flow can maintain cohesion between two populations, minimising the divergence brought by genetic drift or local adaptation of individual populations by having a common gene pool (Barker and Wilson, 2010; Mayr, 2014; Slatkin, 1987). Meanwhile, a weak or decreased gene flow between populations can lead to the genes or alleles sweeping to fixation only within an individual population and independent of the other populations (Mayr, 1999, 1975, 1954; Rieseberg et al., 2004; Slatkin, 1987). The preferential sharing of genes within only a population increases divergence between distinct populations, which can further disrupt the gene flow and lead to speciation. Therefore, disruption in gene flow between

closely related populations is indicative of speciation, meaning identification of these disruptions can aid in delineating species boundaries (Petit and Excoffier, 2009; Sites and Marshall, 2003).

In metazoan species, identification of species boundaries through gene flow disruptions and their accompanying reproductive barriers has been the norm due to the ease of observing their reproductive behaviour, biogeography, and dispersal. In contrast, a similar convention of delineating species through gene flow and reproductive barriers is hard to apply in microscopic organisms (i.e. protists, bacteria and archaea) and asexual organisms (i.e. parthenogenetic, binary fissioned) (Achtman and Wagner, 2008; Cohan, 2002; Cohan and Koeppel, 2008; Shapiro et al., 2016). This difficulty is due to their small size, which makes studying their phenotype, ecological niche, and dispersal difficult. This problem is also further exacerbated by the unusual reproductive behaviour that is common among them (Achtman and Wagner, 2008; Fraser et al., 2007; Hanage et al., 2005; Polz et al., 2013; Wiedenbeck and Cohan, 2011). Prokaryotes can share genes not only vertically but also horizontally (e.g., homologous recombination and conjugation), even from distantly related organisms (Doolittle and Papke, 2006). Thus, identifying ecologically meaningful species boundaries in microscopic organisms based on gene flow and aligned to the biological species concept still presents a challenge to microbiologists (Cohan, 2002; Cohan and Perry, 2007a; Shapiro et al., 2016). As metazoan systematics moved on from phenotypic clustering to classification with a sound species concept, bacterial systematics, unfortunately, has not. Instead, it remains the same as it was prior to the introduction of the Biological Species Concept and relies mostly on similarity clustering. Innovations in the field can only be characterised as improvements in techniques for cluster demarcations (Brenner et al., 2005; Cohan, 2002).

1.3. The Operational Definitions of a Bacterial Species

Current schemes for identifying bacterial species are reliant on a polyphasic approach of comparing properties between bacterial type strains. Generally, a polyphasic approach in identifying bacteria utilises both the phenotypic and genetic properties of bacterial strains being compared (Brenner et al., 2005; Stackebrandt et al., 2002; Vandamme et al., 1996; Wayne et al., 1987). The phenotypic analysis uses bacterial physiology, morphology, and biochemical characteristics to provide a piece of general descriptive information that aids in categorising the taxa (i.e. genus, species and subspecies) to which a particular organism belongs (Raina et al., 2019; Vandamme et al., 1996). However, many of these phenotypic characteristics are irrelevant in measuring genetic relatedness between organisms as they often exhibit alternate characteristics depending on the environmental condition (Beier et al., 2015). Furthermore, the methods in classifying phenotypic characteristics can often change between users and over time as the protocols and guidelines set by the International Journal of Systematics and Evolutionary Microbiology are constantly being

improved (Stackebrandt et al., 2002; Wayne et al., 1987) and often results in discrepancy in characterisation. Thus, comparison using genetic data was seen as a more stable alternative than phenotypic and chemotaxonomic classifications as it can classify strains into clusters of discrete DNA sequence similarity directly, and it also directly involves the unit of heredity: the genes (Krieg, 1988; Rosselló-Mora, 2006).

Analysis of genetic data provides a distance-based method for measuring genetic relatedness and delineating species independent of the bacteria's current phenotypic presentation. DNA-DNA hybridisation (DDH), Mean Sequence Similarity index and conserved gene sequence analysis are the most widely used among these methods. DNA-DNA hybridisation is the gold standard in demarcating genomic similarity between pairs of candidate bacterial species (Raina et al., 2019; Stackebrandt et al., 2002; Wayne et al., 1987). DNA-DNA hybridisation measures genetic relatedness by exploiting DNA's intrinsic property of reassociation between paired nucleotides after thermal denaturation (Johnson, 1985; Rosselló-Móra et al., 2011). Bacterial genomes that hybridised over 70% are observed to belong to the same bacterial species identified through traditional phenotypic comparison (Johnson, 1980; Raina et al., 2019) and has been the threshold for delineating bacterial species (Wayne et al., 1987). However, DNA-DNA hybridisation is cumbersome as it must be performed repeatedly for each pair of bacteria being compared (Rosselló-Móra et al., 2011; Stackebrandt et al., 2002). An alternate distance-based method for measuring genetic relatedness is the mean sequence similarity (nucleotide and protein sequences) score from paired genome sequences. Average Nucleotide Identity (ANI) and Average Amino Acid Identity (AAI) delineated genomes into the same species using a threshold of 94-95% nucleotide sequence and 85-90% sequence similarity, respectively. In addition, it was shown to correlate with observed DDH values (Figure 1.1) (Jain et al., 2018; Konstantinidis and Tiedje, 2007; Rodriguez-R and Konstantinidis, 2014). Meanwhile, conserved gene sequence analysis is a cheaper and more widely used method since it uses short nucleotide sequences of ubiquitous bacterial genes with essential functions as phylogenetic markers (e.g., *16S rrn, recA, gyrB, rpoB*, etc) (Raina et al., 2019). As these genes occupy an essential role in routine bacterial function, they are often under constant purifying selection to remove any deleterious mutations. These genes are thought to accumulate nucleotide changes more gradually than the background mutation rate and, thus, are ideal for measuring evolutionary time (Das et al., 2014). Similar to mean sequence similarity indices, similarity values of some conserved genes, such as 97% similarity of *16S rRNA* gene (*16S rrn*), have been correlated with DDH values (Stackebrandt and Goebel, 1994). Unfortunately, the correlation between these techniques and DDH means that they also have the same shortcomings as DNA-DNA hybridisation and possess several problems of their own.

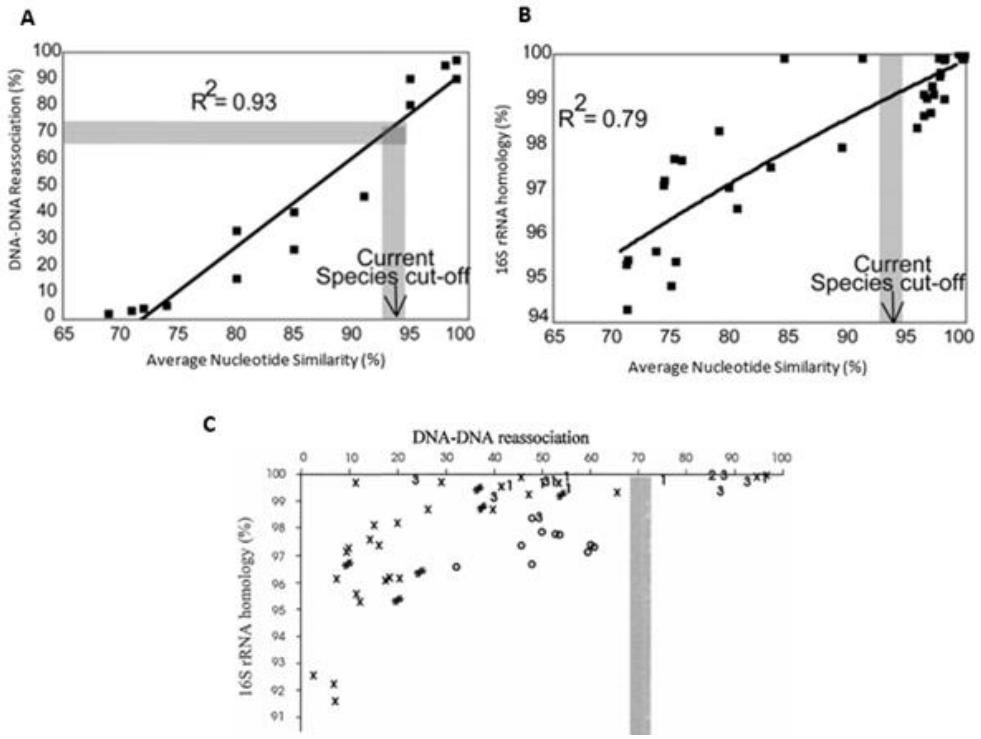

Figure 1.1 Relationships between DNA-DNA Hybridization, Average Nucleotide Identity and *16S rRNA* gene sequence homology. (A) A comparison of Average Nucleotide Identity with DNA-DNA Hybridization values of paired organisms shows a linear correlation between the two metrics, with 70% homology in DDH corresponding to around the 94-95% ANI similarity threshold value. (B) Meanwhile, the relationship of ANI values with *16S rRNA* gene sequence similarity is less linear. It also shows that ANI species threshold values can not easily correlate with the *16S rRNA gene* identity, as several organisms with >99% 16S rRNA gene identity fall below the ANI threshold value. (C) Similarly, a *16S rRNA gene* sequence homology comparison with DDH shows that most organisms with>70% DDH similarity have 99% *16S rRNA gene* identity. However, not all organisms with >99% *16S rRNA gene* identity pass the DDH cutoff. Figures from Konstantinidis and Tiedje 2005 (A,B), and Stackerbrandt and Goebel 1994 (C).

One of the main problems in using sequence-based techniques in delineating species boundaries is the correlation and interpretation of the often arbitrary threshold values with actual natural biological boundaries (Johnson, 1985). For example, the *16S rRNA gene* sequence threshold of <97% similarity indicates a putatively novel species, but the opposite is not necessarily true (Figure 1.1C). The meaning of similarity scores >97% is not clear as many distinct bacterial species with distinct biochemical properties and low genomic DNA homology (e.g., several *Bacillus, Stenotrophomonas, Edwardsiella, Actinomyces* and *Enterobacter* species) exhibit *16S rRNA gene* sequence similarity of >99.5% (Janda and Abbott, 2007; Stackebrandt and Goebel, 1994). A similar problem was also observed in mean similarity indices (i.e. ANI and AAI). Despite utilising newer sequencing technologies and a larger pool of genomes to delimit species boundaries (Carroll et al., 2021; Jain et al., 2018; Rodriguez-R and Konstantinidis, 2014), several genomes have been identified to belong to the same species even with nucleotide similarity of less than 90% or identified as distinct species even with greater than 98% similarity (Figure

1.2). These ANI threshold values that could be used to delimit species boundaries across bacterial taxa (Jain et al., 2018) have been attributed to be likely a product of sampling bias and an overabundance of genome duplicates from the same species (Murray et al., 2021). These problems in delimiting species boundaries are rooted in DNA-DNA hybridisation shortcomings: its high experimental errors (Rosselló-Mora, 2006) and arbitrary cutoff in homology delineation (Johnson, 1985, 1980). Furthermore, even though these methods utilise genetic properties to classify bacteria into taxonomic groups, they still rely on taxonomic references that are often delineated through phenotypic characterisation and other chemotaxonomic properties (Carroll et al., 2021). These references may contain several polyphyletic bacteria within a single identified 'species', or a monophyletic bacteria split into several 'species', which hinders proper delineation.

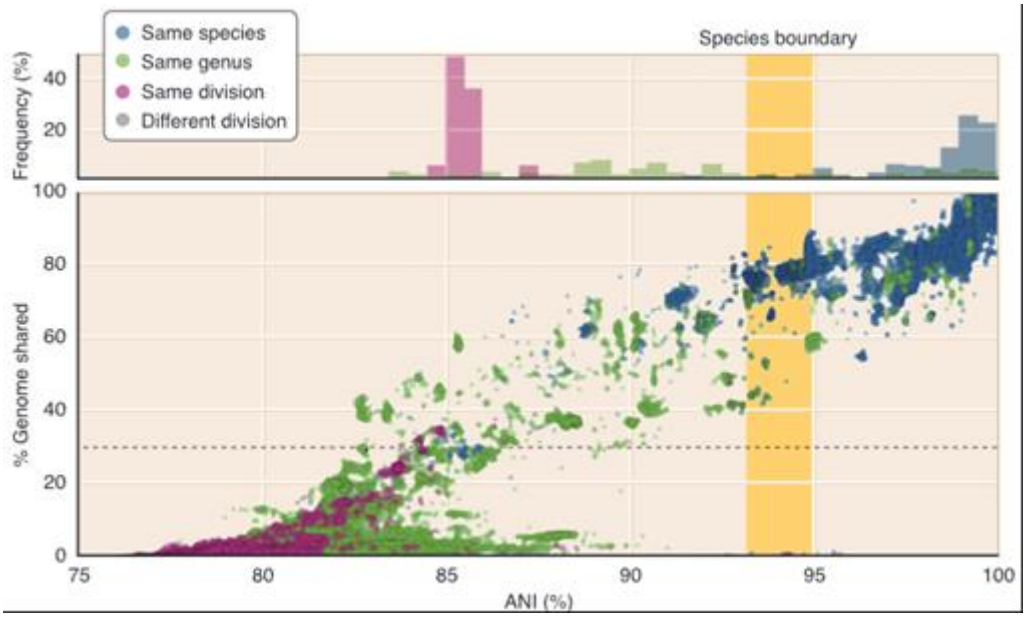

Figure 1.2 Distribution of ANI values of completed genomes across taxa boundaries. A much larger comparison of all available genomes with DDH has shown fuzzy boundaries between named species and their DDH and ANI values. Traditionally, 95% or higher average nucleotide identity score delineates genomes to belong to the same species. However, several genomes within this range (greens) have been categorised as distinct species belonging to the same genus and, thus, indicating the technique's imprecision or the need for reclassification of those bacteria. Figure taken from Rodriguez-R & Konstantinidis, 2014.

Unfortunately, all these methods for delineating bacterial species do not directly reflect a species' biological meaning as a group of reproductively isolated organisms. As such, sequence similarity-based techniques do not directly reflect the disruption in gene flow between diverging bacterial organisms to produce a distinct species. These techniques assume that the mutation rate among bacteria greatly overshadows lateral or horizontal gene transfer rates (Jain et al., 1999; Miyazaki and Tomariguchi, 2019). Therefore, these techniques assume that divergence between sequences can only occur due to the gradual accumulation of point mutations, which can later affect environmentally advantageous genes and speciation. Of course, this assumption

forsakes the prominent role of homologous recombination as a cohesive force in preventing divergence in bacterial species and disregards the observed high prevalence of lateral gene flow between bacteria (Hanage et al., 2005; Polz et al., 2013; Shapiro et al., 2016). Therefore, sequence similarity techniques can classify bacteria neatly into discrete sequence clusters, known as operational taxonomic units, but they do not clearly explain or correlate how speciation occurs. These techniques also often fail to resolve genetic boundaries at species and lower taxa levels (Hanage et al., 2005). Thus, a new method of delineating species that aligns with the gene flow and speciation is needed.

1.4. Proposed Models for Biological Species Concept in Bacteria

Delineating bacterial species based on a biologically meaningful definition necessitates identifying ecologically and genetically cohesive clusters of organisms. These ecologically and genetically cohesive populations of bacteria should bear similar fundamental dynamic properties of species as seen in metazoans, such as the ability to maintain ecological and genetic cohesion or to diverge into a separate population (Cohan, 2002; VanInsberghe et al., 2020). The difference in the models' understanding of the formation of an ecologically and genetically cohesive bacterial population stems from the prevalence of different gene flow works in bacteria (vertical descent vs homologous recombination).

1.4.1. Ecotype and Periodic Selection

The ecotype model is one of the first models to align bacterial speciation with a biological meaning (Konstantinidis et al., 2006). In this model, an ecotype is described as a set of ecologically homogenous bacteria brought by an adaptive mutant outcompeting all other mutants that occupy the same ecological role but not mutants from other ecotypes (Cohan, 2002, 2001). The ecotype model adheres to the dominance of vertical gene flow in bacterial reproduction (Cohan, 2002; Palys et al., 1997), which implies adaptive genes reach fixation within the population through genome-wide sweeps and clonal expansion (Figure 1.3). A genome-wide sweep occurs when adaptive alleles or genes spread throughout the population along with its genomic background (neutral genes) (Bendall et al., 2016; Burke, 2012; Cohan, 2001). Afterwards, a selection event purges non-adaptive mutants, allowing only mutants with the advantageous trait to expand and take over the population, thus forming a genotypic and ecologically cohesive cluster of bacteria (Cohan, 2001, 2002; Cohan and Perry, 2007a; Koeppel et al., 2008; Kopac et al., 2014). As a genetically and ecologically cohesive unit, the ecotype model

proposes that ecotypes exhibit similar fundamental properties of species (i.e. cohesion and divergence) observed in sexual eukaryotes (Cohan, 2019, 2002).

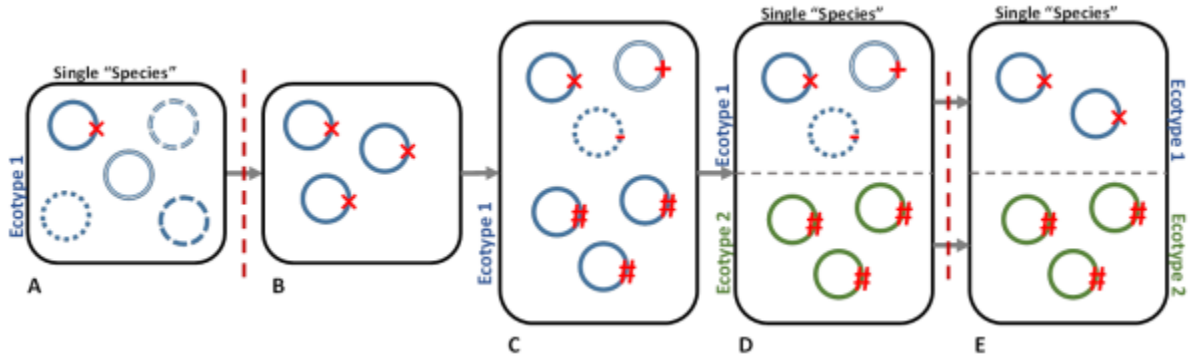

Figure 1.3 Periodic Selection purifies diversity in ecotypes. (A) Bacteria that acquired advantageous mutations (red cross) can survive a periodic selection (red dash line), (B) which can then repopulate and expand within the population. (C) These populations acquire new genotypes (red plus, minus and hash), resulting in increased diversity within the population and/or (D) expanding into a new ecological role to form a new ecotype. (E) Periodic selection, however, can purify an ecotype's diversity, preventing further divergence without affecting other ecotypes. Circle line patterns indicate a unique genomic background. Different colour of circles indicates different ecotypes. Figure adapted from Cohan 2001, 2002, 2019.

The ecotype model proposes that intermittent bouts of natural selection (periodic selection) maintain genetic and ecological cohesion in clonally bacterial populations by periodically removing non-adaptive diverging genomes (Cohan, 2002, 2001; Koeppel et al., 2008). Periodic selection resets the diversity of acquired mutations within the population as only adaptive mutants remain in the population (Figure 1.3A-B). The selective advantage brought by the adaptive mutation causes the adaptive mutant to take over the population through clonal expansion (genome-wide sweep), thus becoming a genetically and ecologically homogenous group (Figure 1.3B). In between periodic selection, the incipient population of adaptive mutants may expand, occupy new habitats and acquire novel mutations (Figure 1.3C). This expansion may allow the population to acquire new adaptive mutations, which allows it to occupy a new ecological role and diverge into a new ecologically distinct population, a new ecotype (Figure 1.3D). Afterwards, the cycle continues in which periodic selection purifies diversity within the ecotypes and minimises divergence within them. Periodic selection within the ecotypes occurs independently of each other, thus retaining diversity between ecotypes and their distinct evolutionary history (Figure 1.3D-E) (Cohan, 2019, 2016, 2002; Koeppel et al., 2008; Kopac et al., 2014).

Unlike eukaryotes that can maintain species cohesion through the frequent exchange of genetic material, the assumed rarity of horizontal gene flow in bacteria meant that horizontal gene flow and recombination does not aid in maintaining the genotypic and ecological cohesion

of a bacterial species (Cohan, 2001, 2002; Cohan and Koeppel, 2008; Vos and Didelot, 2009a; Wiedenbeck and Cohan, 2011). In the ecotype model, maintenance of cohesion in bacterial populations relies mainly on selection as genome-wide sweeps can only occur if the rate of mutation and effect of selection is greater than the rate of horizontal gene flow and recombination ($r/s << 1$) (Cohan, 2002). A constant high rate of recombination in the absence of selection is problematic as it leads to the formation of a single homogenous genotypic cluster, thus preventing genetic divergence and the formation of an ecologically distinct population. However, when mutation rates become more frequent than recombination but without selection, clonal clusters can form through random mutation but rapidly drift to extinction as deleterious mutations also quickly accumulate, which suggests that neutral cluster formation is unlikely in bacteria (Fraser et al., 2007; Shapiro and Polz, 2015). Thus, a positive natural selection for an adaptive trait is required to produce stable genotypic clusters (Cohan, 2019, 2016; Cohan and Koeppel, 2008; Fraser et al., 2007; Shapiro and Polz, 2015; Wiedenbeck and Cohan, 2011).

The periodic purge of divergence means that given enough time, ecotypes would be identifiable as discrete sequence clusters in which sequence divergence between ecotypes is more prominent than within each ecotype (Cohan, 2002; Palys et al., 1997). As each ecotype propagates clonally, phylogenetic studies of each ecotype would show that each discrete sequence cluster represents a single ecotype lineage that is monophyletic and derived from a single ancestor. Inversely, an ecotype cannot be separated into multiple sequence clusters as adaptive mutants should be able to drive other sequence clusters of the same ecotype to extinction during every periodic selection. However, a caveat is that two or more sequence clusters in different geographical habitats belonging to a single ecotype can occur as there is no opportunity for them to interact and outcompete the others during each periodic selection (Cohan, 2019, 2002). Therefore, the ecotype model suggests that the currently described bacterial 'species' is a metapopulation composed of multiple ecotype lineages, which can be observed as intra-species genomic variations (Cohan, 2016; Majewski and Cohan, 1999).

1.4.2. Ecological Populations and Horizontal Gene Flow

The ecological population model is another model that tries to align bacterial speciation with the BSC. In this model, there is a greater emphasis on the effect of horizontal gene flow (homologous recombination in particular) in the formation of an ecologically and genetically distinct population of bacteria (Shapiro and Polz, 2014). Unlike in the ecotype model, in which the homogenising property of recombination hinders genome-wide sweeps, the ecological population model recognises that rather than a problem, horizontal gene flow through

homologous recombination is a mechanism that aids in the formation of ecologically cohesive populations (Fraser et al., 2009, 2007; Hanage et al., 2005; Polz et al., 2013; Shapiro and Polz, 2015). The prevalence of horizontal gene flow and, specifically, homologous recombination suggests that gene-level sweeps are as likely to occur as genome-wide sweeps in the formation of an ecologically cohesive population (Bendall et al., 2016; Doolittle, 2012; Fraser et al., 2009). In gene-level sweeps, adaptive genes spread to fixation within the population independent of the genomic background (Figure 1.4). A decrease in gene flow between bacteria can result in the formation of gene flow boundaries that limit the sharing of adaptive genes between bacteria connected by horizontal gene flow (Arevalo et al., 2019a; Cordero and Polz, 2014; Fraser et al., 2007; Hanage et al., 2005; Shapiro and Polz, 2015). Taking into account the total gene flow in bacteria, the ecological population model postulates that ecologically cohesive populations of bacteria should be delineated exclusively by gene flow boundaries between bacteria, regardless of whether vertical or horizontal gene flow dominates these populations (Cordero and Polz, 2014; Fraser et al., 2009; Shapiro and Polz, 2014).

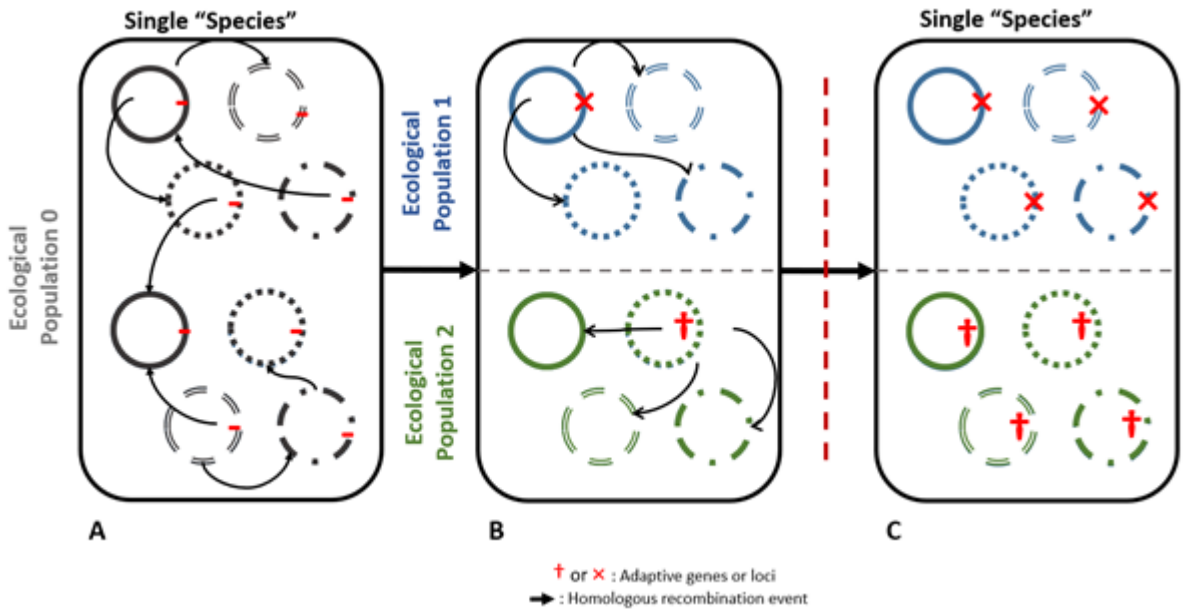

Figure 1.4 Formation of ecological population brought by homologous recombination and selection. (A) In the absence of selection, homologous recombination maintains the ecological and genotypic cohesion of a population. (B) Decline or sudden loss of recombination between organisms (grey dash line) results in preferential sharing of adaptive genes within the cluster rather than between clusters. (C) This results in the formation of an ecologically cohesive cluster brought by a gene-level sweep of an adaptive gene upon selection (red dash line). Solid arrow lines indicate homologous recombination; Each distinct circular pattern indicates a distinct genomic background; Red minus, cross and daggers indicate adaptive alleles.

Horizontal gene flow in the ecological population model is thought to be primarily driven by homologous recombination, which allows genetic exchange between closely related bacteria and forms ecological populations. It requires homology of flanking regions of ~20bp identical

sequences and proximity between bacteria to initiate recombination (Arevalo et al., 2019a; Polz et al., 2013; Shapiro and Polz, 2014). A drop in recombination frequency can be observed as sequence divergence between bacteria increases, and thus, homologous recombination can be observed more frequently between closely related bacteria. Unlike other mechanisms of HGT, which utilise mobile elements or bacteriophages to incorporate non-homologous DNA, homologous recombination is thought to be facilitated by the uptake of DNA (Arevalo et al., 2019a, 2019b; Fraser et al., 2009; Shapiro et al., 2012). The susceptibility of environmental DNA to environmental degradation requires the protection of DNA or close proximity of bacteria for uptake of naked DNA and homologous recombination to occur (Nagler et al., 2022). Therefore, ecological populations are formed more easily along closely related bacteria occupying similar microhabitats.

Table 1.1 Comparison of bacterial speciation in varying degrees of selection and recombination

	Clonal Population ($r/s << 1$)	**Recombinogenic Population** ($r/s >> 1$)
Stage 1	New niche-specifying variants(s) acquired by mutation, homologous recombination, or HGT	
Stage 2	Ecological separation: new variant spreads in new niche by clonal expansion	Ecological separation: new variant spreads in new niche by recombination
Stage 3	Genetic separation driven by periodic selection and drift	Genetic separation driven by genome-wide depression in recombination between new and ancestral niches
Stage 4	Genetic separation maintained by further periodic selection and drift events; lineages are permanently separated	Genetic separation maintained by genetic barriers to recombination; otherwise lineages may merge back together
Stage 5	Lineages remain ecologically and genetically distinct (at both adaptive and neutral loci) until extinction	

Note: Adapted from Shapiro and Polz (2014). Copyright 2014 by Cell Press.

However, in the ecological population model, the formation of ecologically and genotypically cohesive populations of bacteria does not solely rely on recombination but depends on the interplay between the effect of recombination and selection (r/s) (Table 1.1) (Fraser et al.,

2007; Polz et al., 2013; Shapiro and Polz, 2015). In the ecological population model, both clonal and recombinogenic bacteria can form ecologically distinct populations, albeit in different manners. Asexual and highly clonal bacteria form ecologically distinct populations similar to the ecotype model (Table 1.1), in which selection plays a greater role than recombination ($r/s<<1$), resulting in the formation of distinct genotypic clusters through the emergence of random mutations (Fraser et al., 2007; Shapiro and Polz, 2015). Meanwhile, in a recombinogenic or 'sexual' bacterial population (Table 1.1), in which the recombination rate is much higher than selection ($r/s>>1$), the formation of an ecologically distinct population is a much more complex process in which barriers to recombination plays a major role in both genetic divergence and maintenance of genetic cohesion (Shapiro and Polz, 2014).

In recombinogenic populations, gene-adaptive genes are easily shared between population members independent of the neutral genomic background (Figure 1.4A) (Fraser et al., 2009; Shapiro and Polz, 2015, 2014). However, when recombination declines or is hindered between bacteria, distinct genotypic clusters of bacteria are formed in which adaptive genes are preferentially shared within a specific cluster than between other clusters (Figure 1.4B). This preferential sharing of genes leads to the formation of a distinct ecologically significant cluster upon selection (Figure 1.4C) (Shapiro et al., 2012; Shapiro and Polz, 2015, 2014). The rapid decline in the recombination rate brought by genetic barriers (i.e. sequence divergence, epistasis or dispersal) counteracts the homogenising property of recombination (Fraser et al., 2007; Shapiro and Polz, 2015) and can be analogous to how reproductive barriers prevent interbreeding in a eukaryotic species. The persistence of such genetic barriers allows the maintenance of ecological and genotypic clusters as a distinct unit which would otherwise merge back as a homogenous cluster (Fraser et al., 2007; Polz et al., 2013; Shapiro and Polz, 2015, 2014).

Even with a higher effect of recombination, selection still plays a significant role in the formation of a stable ecological population (Shapiro and Polz, 2015). Similar to the ecotype model, the absence of selection in clonal populations results in genetic drift and extinction. Meanwhile, selection's absence in recombinogenic populations necessitates an absurdly high rate of decline in recombination for an ecological population to form. The formation of an ecologically distinct cluster through neutral processes in the ecological population model is unlikely (Fraser et al., 2007; Shapiro and Polz, 2015). However, even with selection, divergence is not guaranteed if nascent populations are not accompanied or followed by habitat partitioning (Figure 1.4B) (Chase et al., 2019; Shapiro et al., 2016; Shapiro and Polz, 2014). Lack of habitat partitioning would allow the homogenising properties of horizontal gene flow to occur between the nascent ecological populations, thereby preventing the formation of an ecologically distinct population (Shapiro and Polz, 2014; Wiedenbeck and Cohan, 2011).

Taken together, horizontal gene flow and homologous recombination in the ecological population model play the leading role in diverging and maintaining cohesion in the bacterial population. The presence of horizontal gene flow allows the homogenisation of diversity within an ecological population, while its absence leads to the formation of distinct genotypic clusters in both clonal and recombinogenic ecological populations (Fraser et al., 2009; Shapiro and Polz, 2015). Therefore, we could compare horizontal gene flow as akin to reproduction in sexual eukaryotes and gene flow boundaries forming ecological populations as analogous to reproductive barriers leading to the formation of reproductively isolated populations or species. This analogy also implies that ecological population is the fundamental unit of ecology and evolution in bacteria, similar to how species are treated in BSC and should exhibit similar ecological and evolutionary behaviours observed in metazoan species. Furthermore, the intra-species genomic variations observed in natural bacterial communities are the diversity brought by the varied neutral mutation in the genomic background.

1.4.2.1. Mode of horizontal gene flow and maintenance of ecological population boundaries

Recombination or horizontal gene flow is often facilitated by three modes of dispersal and acquisition of extrageneous genetic material: conjugation, transduction, and transformation. Conjugation is the direct cell-to-cell transfer of genetic material between bacteria through pili or adhesins (Thomas and Nielsen, 2005). It is often facilitated by transferring small mobile genetic elements such as plasmids and transposons (Zarei-Baygi and Smith, 2021). Conjugation is often thought to be the primary driver of the dispersal of antibiotic resistance genes (ARGs) and, thus, of significant environmental and epidemiological importance (von Wintersdorff et al., 2016). Meanwhile, transduction, the process of transferring genetic material through viral intermediaries, is observed to occur at a much lesser frequency in transferring ARGs (Volkova et al., 2014; von Wintersdorff et al., 2016), even though phages harbour specific families of ARGs in the environment (Zarei-Baygi and Smith, 2021). Unlike conjugation, which is a directed transfer, the transfer of bacterial DNA through transduction is assumed to be a secondary effect of erroneous bacteriophage replication or gene replication (Norman et al., 2009). However, One advantage of transduction is that bacteriophages have a wide range of microbial hosts and thus can spread genetic material across taxonomically distant bacteria (von Wintersdorff et al., 2016). The third canonical mode of HGT is natural transformation, which occurs through direct uptake and integration of environmental DNA (Thomas and Nielsen, 2005). Although natural transformation is limited by the requirement of natural competence of bacterial cells (i.e. presence of *ComEC*, *secretin* and *rec* genes) (Mell and Redfield, 2014) and the persistence of extracellular DNA in the environment (Thomas and

Nielsen, 2005; von Wintersdorff et al., 2016), natural transformation could acquire ARGs at a similar rate to conjugation (Zarei-Baygi and Smith, 2021).

Other than these three modes of transfer, we can also categorise horizontal gene flow and recombination through the results of the integration of genes/alleles into the genome: additive gene transfer, paralogous addition, and xenologous recombination (Novick and Doolittle, 2019). Gene acquisition can be done through the acquisition of novel gene families that were not previously present in the genome (additive gene transfer or more known by misnomer as non-homologous recombination), thereby acquiring novel functions or through paralogous addition (Schaller et al., 2021; Soucy et al., 2013). Paralogous addition results in the acquisition of copies of genes from duplication events (Olendzenski et al., 2006; Soucy et al., 2013). Unlike the acquisition of novel genes or gene copies in these additive gene transfers, replacement gene transfer or xenologous recombination is observed to be more frequent (Choi et al., 2012; Khayi et al., 2015). The frequent xenologous recombination could be due to the need for only at least 20bp flanking sequences to initiate recombination, which could occur more frequently between closely related bacteria (Hua et al., 1997; Novick and Doolittle, 2019; Olendzenski et al., 2006; Shapiro and Polz, 2015). Additionally, the lack of novel function that can impact the host's fitness cost may facilitate faster integration of replaced genes (Novick and Doolittle, 2019).

In ecological population theory, the formation of ecologically and genotypically coherent populations through gene-level sweep is attributed to frequent horizontal gene flow, whether homologous or non-homologous recombination (Cordero and Polz, 2014; Shapiro et al., 2016; Takeuchi et al., 2015). However, it is primarily assumed, based on a few studies, that homologous recombination through xenologous exchange of alleles drives the maintenance of ecologically and genotypically coherent populations (Chase et al., 2019; Hülter and Wackernagel, 2008; Mell and Redfield, 2014). Replacement gene transfers often rely on flanking sequences, which are inefficient in introducing novel genes but could be efficient in maintaining populations as they can occur more frequently and do not have the fitness cost attached to acquiring new genes or functions (Novick and Doolittle, 2019). Meanwhile, homology-independent gene acquisition (i.e. additive gene transfer), often facilitated through mobile genetic elements and viral sequences, could acquire novel genes faster and thus are more likely to facilitate the novel function and adaptation to new ecological niches by the bacterial population (Mell and Redfield, 2014). Therefore, novel niches are acquired through geographical expansion and acquisition of novel functions through non-homologous gene transfer in a single or few bacteria, but these genes get fixed throughout the population primarily through homologous recombination or genome-wide expansion, resulting in a novel ecological population.

1.5. Ecologically Significant Units in Natural Environments

The ecotype model's reliance on the widely accepted assumption that selection and mutation greatly outperform recombination allows it to be one of the most widely used models to describe ecologically significant units in bacteria (Bendall et al., 2016; Cordero and Polz, 2014). The common notion of horizontal gene flow's rarity in a bacterial population helps perpetuate the idea of genome-wide sweeps' dominant role in bacterial speciation and evolution. Studies in *Bacillus*, *Synechococcus* and *Chlorobium* show the formation of ecologically distinct populations under positive selection and genome-wide sweeps, suggesting agreement with the ecotype model (Becraft et al., 2015; Bendall et al., 2016; Kopac et al., 2014). In the *Bacillus* and *Synechococcus* studies, phylogenetic analysis of genes undergoing positive selection (genes with elevated SNPs over time) allows the grouping of isolates into distinct genotypic clusters. Its ecological coherence is subsequently verified by similarity in environmental distributions and response to environmental perturbations (Becraft et al., 2015; Kopac et al., 2014). Meanwhile, Bendall's (2016) time-series study of *Chlorobium* showed the first direct evidence of a natural bacterial population undergoing genome-wide selective sweeps. Throughout the eight-year study, a significant loss of neutral SNPs and the increase in the abundance of a few select genes indicate that the population is undergoing purifying selection, thus preventing ecological divergence (Bendall et al., 2016; Cohan, 2016).

However, observations from among the other 29 natural bacterial populations exhibited different gene flow patterns in which genome-wide sweeps do not strictly follow the predictions of the ecotype model. In the same Bendall's (2016) study of *Chlorobium*, the other populations undergoing genome-wide selective sweeps are not completely clonal even after eight years, indicating incompleteness of the sweep or a possible soft sweep. In contrast to a hard sweep which completely purges genomic diversity, selection under a soft sweep favours several persistent genotypes that could have acquired advantageous genes independently or through intra-population recombination (Messer and Petrov, 2013). Soft sweeps are more likely to occur when there is a long time between selection events, which allows for a greater chance of acquiring advantageous genes. Lack of a complete purge of intra-population genomic diversity in *Salinibacter ruber* and *Haloquadratum walsybi* populations that occupy an ecologically distinct niche was also observed, suggesting soft sweeps of ecologically distinct populations (Viver et al., 2021).

Aside from soft sweeps, formation of ecologically and genotypically distinct populations that deviates from ecotype model have been observed to be frequently linked with high intra-population horizontal gene flow. Example of the deviations in ecotype theory can be found in Synechococcus and Chlorobium populations. The *Synechococcus* and *Chlorobium* ecological

populations were previously observed to go under genome-wide sweeps (Becraft et al., 2015; Bendall et al., 2016) but also exhibit high intra-population diversity brought by large genome regions sweeping independently through some populations (Bendall et al., 2016; Rosen et al., 2015). At least in *Synechococcus*, extensive recombination allows selective sweeps of specific loci and genes, resulting in the formation of an ecologically adapted population and thus retaining the diversity of its unlinked genomic regions (Miller and Carvey, 2019). The occurrence of both genome-wide sweeps and gene-specific sweeps in similar bacterial taxa suggests the larger role of recombination in maintaining ecologically distinct populations.

An accompaniment to the ecotype model, the "Adapt Globally, Act Local" model suggests that recombination events are still compatible with genome-wide sweeps and periodic selection in the formation of ecologically distinct populations (Cohan, 2016; Wiedenbeck and Cohan, 2011). This model suggests that rarely, adaptive mutations can confer selective advantage across multiple ecotypes. This happens when a generally adaptive mutation (adaptive for many ecotypes) arises randomly within a single ecotype (Figure 1.5A). This mutation initially swept within its initial ecotype before homologous recombination allows the transfer of small loci or genes with the adaptive mutation into other ecotypes (Figure 1.5B), thus transferring selective advantage to other ecotypes. Afterwards, global periodic selection targeting this adaptive mutation results in a loss of divergence both between and within ecologically distinct populations or ecotypes (Figure 1.5C) (Cohan, 2016). Thus, rare recombination events can still maintain cohesion and prevent further divergence between different ecotypes (Cohan, 2016; Majewski and Cohan, 1999; Melendrez et al., 2016). This results in the formation of a metapopulation of ecologically heterogenous ecotypes, which can be observed as intra-species genomic heterogeneity (Cohan, 2016). Although recombination in this model plays a role in maintaining distinct ecological populations, the model insists that recombination is still rare and that the effect of periodic selection drives the ecological cohesion of the population (Cohan, 2011; Melendrez et al., 2016; Wiedenbeck and Cohan, 2011). Additionally, in the "Adapt Globally, Act Local" model, gene-level sweeps are suggested to occur due to rapid diversification, like those in generalist heterotrophic populations occupying new energy sources. In contrast, genome-wide sweeps are more likely to be found in ecologically homogenous clusters such as energy-constrained photoautotrophic populations (Cohan, 2019, 2016).

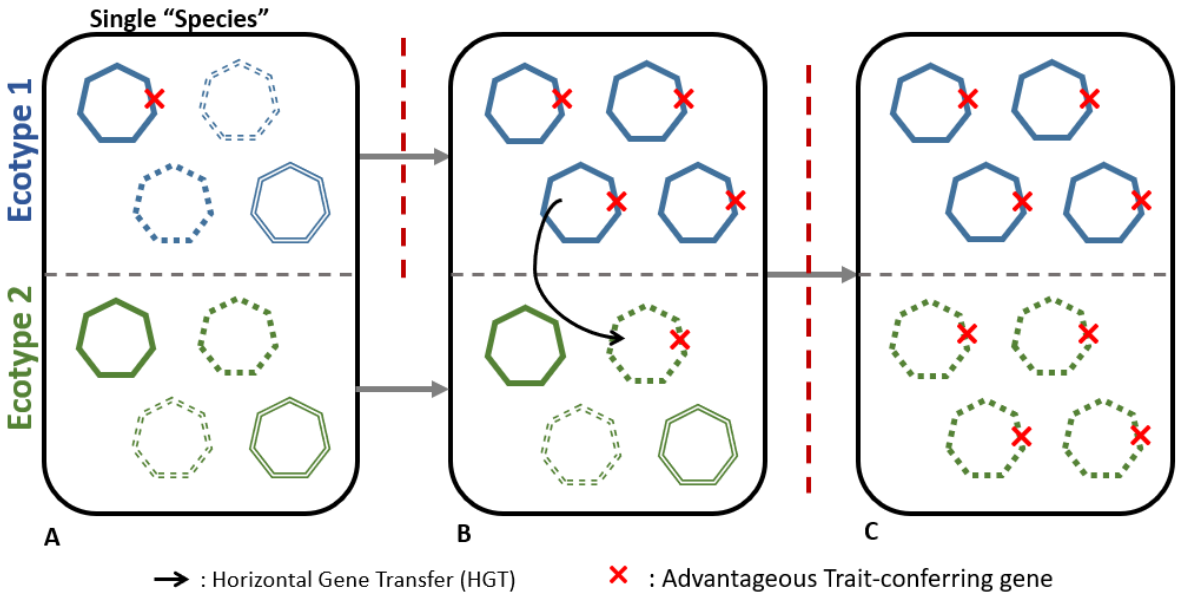

Figure 1.5 Rare recombination events between ecotypes preserve intra-species genomic variation. (A) Advantageous mutations often appear in a single ecotype and sweep through its own distinct population before being recombined into another ecotype, (B) which allows it to survive a similar periodic selection. (C) The bacteria that acquired the adaptive mutation can then expand within its own population. The heptagon colour indicates ecotypes; the heptagon outline indicates a distinct genomic background; the red dashes line indicates a selection event; the red cross indicates an adaptive mutation. Figures adapted from Cohan 2016.

The "Adapt Globally, Act Locally" model suggests that a single recombination event between ecotypes allows a much simpler explanation than the recurrent recombination events proposed by the ecological population model and agrees with the supposed rarity of bacterial recombination (Cohan, 2019, 2016; Melendrez et al., 2016). However, such a model provides an additional layer of complexity in which inter-population recombination would be frequent to allow horizontal gene flow but rare enough at intra-population levels to allow genome-wide sweeps to form ecologically distinct populations (Bendall et al., 2016; Cohan, 2016). This model and the overarching ecotype model insist that genome-wide sweeps and vertical descent are the primary drivers of ecological divergence and speciation in bacteria. Nevertheless, is it the only one? Given the increasing observation of horizontal gene flow and gene-level sweeps in natural populations, it would be better to describe bacterial speciation as a spectrum of genome-wide and gene-level sweeps depending on their rate of horizontal gene-flow. A highly clonal bacterial population would be more inclined to undergo genome-wide sweeps, while gene-level sweeps occur more likely in recombinogenic populations (Bendall et al., 2016; Fraser et al., 2009; Shapiro and Polz, 2015).

Recent studies on homoplasies in sequenced genomes belonging to the same species in the NCBI database have shown almost half of named species undergo continuous recombination across different strains, while a quarter of named species exhibit sharp discontinuities in gene

flow between their members (Bobay and Ochman, 2017). These observed variations in gene flow in named species support the fact that bacterial speciation is on a spectrum. Therefore, it is more apt to describe bacterial speciation in terms of models and methods that explicitly consider and mathematically describe the role of horizontal gene flow in bacterial populations, like the ecological population model. As previously discussed, the ecological population model utilises a method in which species are delineated solely based on discontinuities in total gene flow (vertical and horizontal gene flow) without the need to identify the ecological niche or habitat of the populations beforehand (Arevalo et al., 2019b, 2019a; Shapiro et al., 2009; Shapiro and Polz, 2014). With genetic isolation as its primary criterion, the ecological population model can delineate both recombinogenic and clonal populations without prior knowledge of their recombinogenicity or the lack thereof (Fraser et al., 2009; Shapiro and Polz, 2014).

1.5.1. Detecting gene flow and Delineating gene flow boundaries

Since variation in the rate of horizontal gene flow between bacteria results in the formation of a distinct genotypic cluster, identification of discontinuities (i.e. breaks or decline) in gene flow between co-habiting bacteria allows for the demarcation of ecological populations (Arevalo et al., 2019a; Shapiro and Polz, 2014). The lack of any horizontal gene flow between members of a population is indicative of the formation of highly clonal ecological populations, with each ecological population possessing a highly identical genomic background (Figure 1.6 - Red). However, for a recombinogenic population, the analysis of the change of gene flow between population members needs to be more scrutinised as there could be just a decrease in horizontal gene flow and not its total absence (Figure 1.6 - Blue/Green). Demarcating ecologically distinct clusters among recombinogenic bacteria necessitates the identification of the intensity of recent horizontal gene flow (Arevalo et al., 2019a; Shapiro et al., 2012).

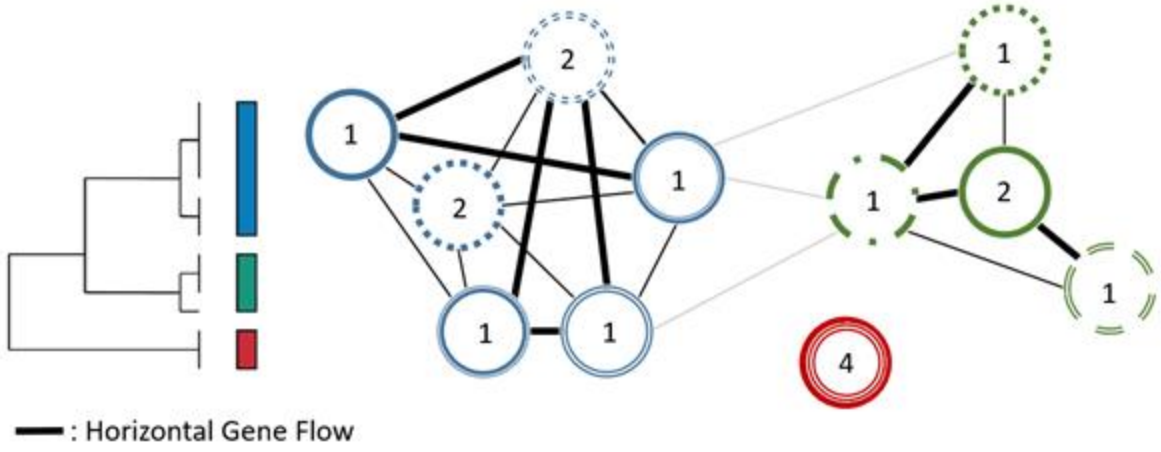

Figure 1.6 Visualisation of the role of horizontal gene flow and the formation of ecological populations. Strong horizontal gene flow between organisms allows the preferential sharing of adaptive alleles within an ecological population compared to other such clusters (Blue and Green). Meanwhile, a

highly clonal ecological population would result in genomes being observed as almost completely identical and lacking horizontal gene flow occurring between organisms and other clusters (red). Phylogenetic tree reconstruction of these organisms using the recombined regions or genes would result in a genotypically coherent topology. Circles denote bacteria with distinct genomic backgrounds, with numbers denoting the number of bacteria sharing that distinct genomic background; Colours indicate distinct ecological populations; the Black line indicates recent horizontal gene flow, and thickness indicates its prevalence; the Grey line indicates historical horizontal gene flow. Figure adapted from Arevalo 2019, Chase 2019, VanInsberghe 2020.

Identifying horizontal gene transfer, particularly those driven by homologous recombination, between closely related bacteria is more complicated than identifying genes transferred from distantly related organisms. Horizontal gene flow is traditionally identified through parametric methods (i.e. nucleotide composition, oligonucleotide frequencies, structural features) and phylogenetic studies that identify genes or loci that significantly differ from surrounding regions or genomic backgrounds in terms of their sequence composition and evolutionary history (Becq et al., 2010; Ravenhall et al., 2015). With these methods, horizontal gene transfers can be identified between distant bacteria, with up to around 25% genome similarity (Ravenhall et al., 2015). Unlike HGT between distantly related bacteria, homologous recombination occurs between closely related bacteria that often occupy similar habitats to allow efficient uptake of transferred DNA (Fraser et al., 2009; Shapiro et al., 2012; Shapiro and Polz, 2014). Closely related bacteria that undergo homologous recombination could have genome similarity of >99% and would have similar or identical genomic signatures (i.e. GC content, oligonucleotide patterns) (Meziti et al., 2019; Shapiro et al., 2012). Identification of recent homologous recombination, however, could be difficult using shared sequence signatures as it may not be possible to disentangle it from historical recombination (Arevalo et al., 2019a). This is unfortunate as analysis of recent recombination is needed to accurately delineate ecologically cohesive populations under selective pressure. Therefore, a method of identifying potentially recombining regions of DNA is through investigation of the distribution of random point mutation throughout the genome.

1.5.1.1. The PopCOGenT Method

PopCOGenT (Populations as Clusters of Gene Transfer) is an ecological population prediction method that can measure and differentiate recent horizontal gene flow among closely related organisms without prior knowledge of the core or flexible genome (Arevalo et al., 2019b). Differentiating recent gene flow from the overall gene flow allows for the precise demarcation of ecological populations, as historical gene flow can blur population boundaries(Arevalo et al., 2019a; Shapiro et al., 2012). The PopCOGenT method utilises a pairwise comparison of entire genomes to identify all recombining regions irrespective of whether all population members share them. This method takes into account that preferentially shared genetic materials of the whole population as population-differentiating genes have been observed in both the core and

flexible genomes (Arevalo et al., 2019a; Shapiro et al., 2012; VanInsberghe et al., 2020). Measurement of recent horizontal gene flow is done by identifying and measuring the distribution of identical shared regions, termed locally collinear blocks (LCBs), throughout the paired genomes.

An LCB is a long stretch of identical genomic regions that do not recently experience rearrangements and are absent of any SNPs which arise randomly throughout the genome due to mutation and vertical descent or recent horizontal transfer (Darling et al., 2004; Minkin and Medvedev, 2020). As SNPs randomly interrupt a stretch of DNA, long LCBs between two genomes would be less observed in non-recombining organisms (VanInsberghe et al., 2020). In contrast, recently recombined homologous DNA should have long LCBs between compared genomes (Figure 1.7A). Long LCBs imply recent recombination of that region as mutations or rearrangements have not had time to alter the DNA sequences of the region to form SNPs. As more time passes without recombination, the stretch of identical DNA can be observed to become shorter in size due to the appearance of SNPs brought through random mutation or genomic rearrangements. Genomes of highly recombinogenic bacteria are, therefore, predicted to have a higher fraction of long LCBs in a genome alignment than non-recombining bacterium genomes (Figure 1.7B and Figure 1.8) (Arevalo et al., 2019a; VanInsberghe et al., 2020).

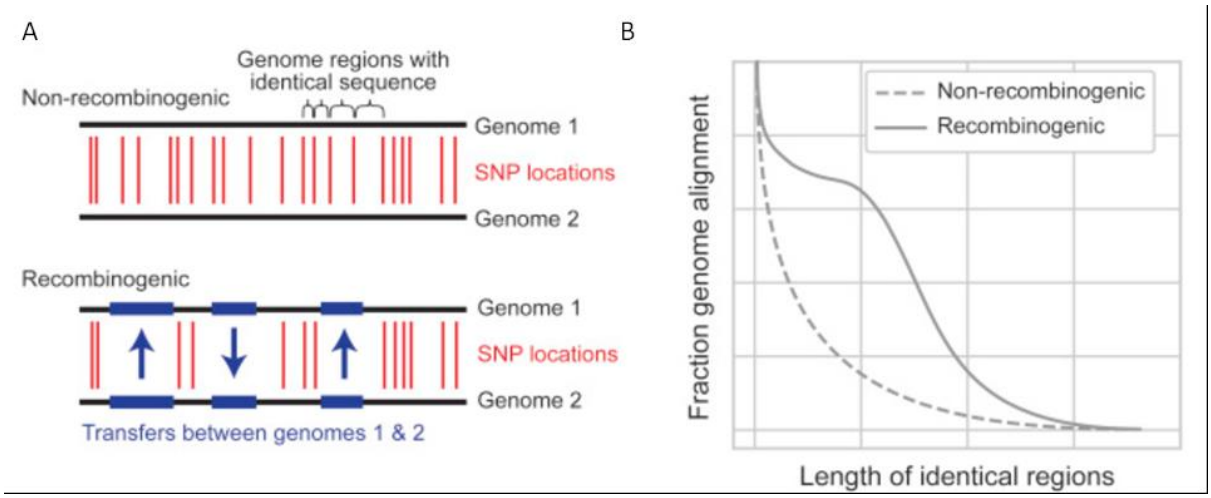

Figure 1.7 Locally Collinear Blocks between recombining and non-recombining genomes. (A) Non-recombining genomes should exhibit fewer and shorter LCBs due to random mutations punctuating any homologous region. In recombining genomes, constant and recent homologous recombination allows the maintenance of homology between regions being shared. (B) Therefore, genomes of recombining strains (B) should exhibit a higher frequency of longer homologous regions than non-recombining strains. Figure taken from Arevalo 2019

The fraction and length of all LCBs (length bias) in the genome allow the estimation of the amount of homologous recombination between bacteria and categorise whether compared organisms are from recombinogenic or clonal populations (Arevalo et al., 2019a; VanInsberghe

et al., 2020). Comparing genomes from different species with varying degrees of recombination shows that the length bias of recombinogenic populations (e.g., *Vibrio cyclitrophicus*, *Sulfolobus islandicus,* and recombinogenic *Salmonella enterica*) differs from non-recombinogenic populations, which follows the null model for a highly clonal bacteria (Figure 1.8A) (Arevalo et al., 2019a). The increase in length bias for non-recombinogenic bacteria as compared to the null model could be based on either low levels of previously undetected recombination in the model organisms used or by a higher number of genomic regions with decreased divergence due to the effect of purifying selection, which scales with both genome and population size (Arevalo et al., 2019a; Bobay and Ochman, 2018). To differentiate between undetected recombination or effect of purifying selection, the PopCOGenT method utilises a genome size-adjusted model to set the lower bound for detection of recombination based on the trend of model non-recombinogenic populations (Figure 1.8B) (Arevalo et al., 2019b).

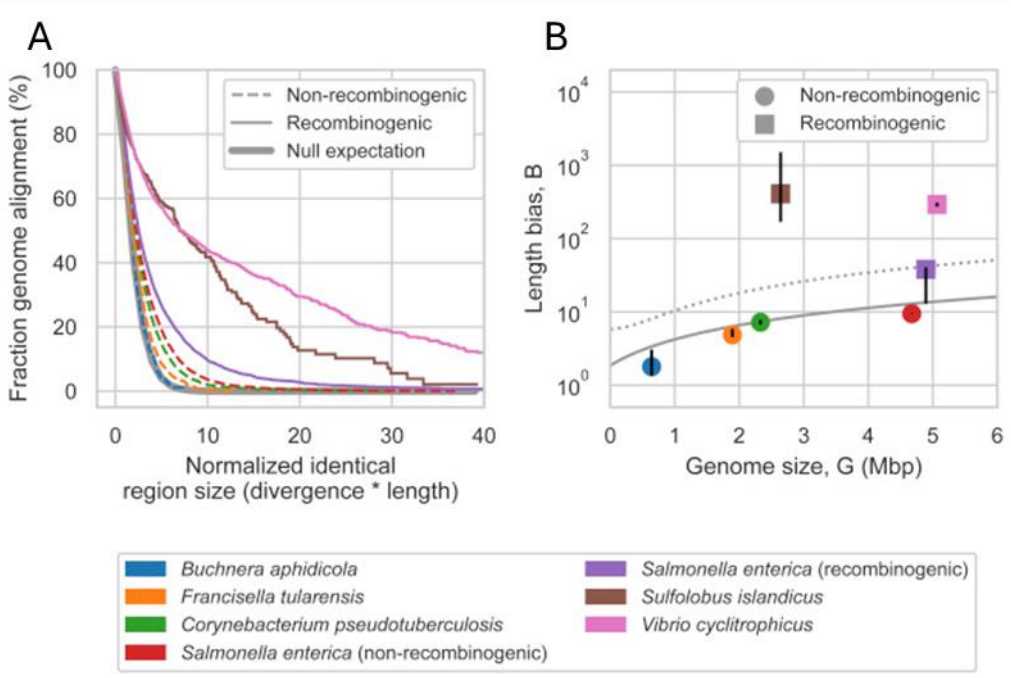

Figure 1.8 Length Bias distribution in known recombinogenic and non-recombinogenic populations. (A) Known non-recombinogenic populations (dash line) of *Buchnera aphidicola, Francisella tularensis, Corynebacterium pseudotuberculosis* and *Salmonella enterica* show similar distribution of length bias as the null model of mutational accumulation, while recombinogenic populations of *Salmonella enterica*, *Sulfolobus islandicus* and *Vibrio cyclitrophicus* shows considerable deviation from the null model (solid line). (B) Non-recombinogenic populations (circles) also follow a linear trend in which the length bias of identical genome region distributions increases proportionally with genome size. Meanwhile, the recombinogenic populations also deviate from this trend, as indicated by the median length bias measured for each indicated population (square). Error bars represent an interquartile range of length bias. The solid grey line represents the best-fit linear regression of length bias against genome size for non-recombinogenic strains, while the dotted grey line is the upper bound of the 90% prediction interval of the linear regression. Figure taken from Arevalo 2019.

The measured length bias between paired genomes is used to construct population clusters, which are networks of gene transfer events with genomes acting as nodes and the weight of measured length bias as edges between paired genomes (Figure 1.9). Highly identical genomes, genomes with less than 0.0355% divergence score as determined using known non-recombining model genomes, are collapsed into a single node as clonal complexes. Meanwhile, edges of clonal complexes with other genomes are estimated as the mean measured length bias between each member of the clonal complex and the other genomes (Figure 1.9) (Arevalo et al., 2019a). These clusters of gene-flow network were found to be highly congruent with previously established genetic and ecological units as indicated by the similarity with the clusters formed and taxonomic trees of the populations (Figure 1.9) (Arevalo et al., 2019a; VanInsberghe et al., 2020).

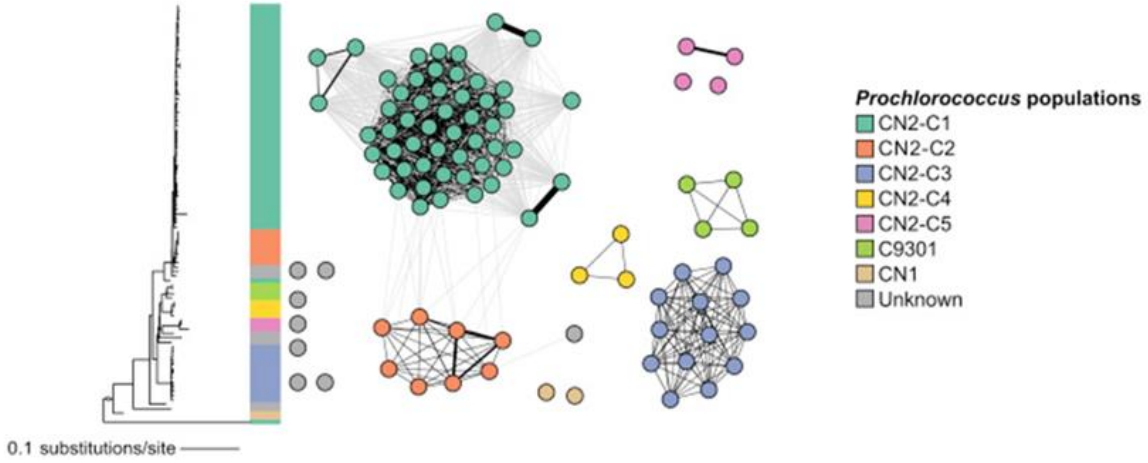

Figure 1.9 Horizontal gene flow network of Prochlorococcus bacteria. Bacterial genomes (nodes) that are connected to each other through gene flow (edges/lines) form a cluster that is categorised as an ecological population (same colour). These clusters of ecological population aligned similarly with the ribosomal reference tree (left panel) with deviations in some populations such as CN2-C1, which is observed to be spread out in the ribosomal tree. Black and grey coloured edge lines denote gene flow within and between populations, with their thickness corresponding to the amount of gene flow occurring between nodes. The node size indicates the number of genomes included within clonal complexes with <0.0355% divergence. Figure from Arevalo 2019.

Unlike methods that rely on global sequence alignments that can only detect and measure gene transfer in core genomes, PopCOGenT can estimate and detect gene transfers independent of global alignments, which allows identification of DNA regions transferred both in core and flexible genomes (Arevalo et al., 2019a). Recent horizontal gene transfer in the core genome alone is presumed to have slower evolutionary turnover as it occurs mainly through homologous recombination (Novick and Doolittle, 2019). Meanwhile, gene transfers in the flexible genomes are thought to occur more rapidly through mobile genetic elements or other transfer mechanisms but are harder to identify in global alignments, as they can be completely rearranged or involve gene loss or addition (Arevalo et al., 2019a; Cordero and Polz, 2014).

However, recent studies have shown that core genome could have a higher degree of homologous recombination to maintain population cohesion, while flexible genomes are more likely to experience gene gain/loss for a rapid adaptation to environmental cues (Novick and Doolittle, 2019; Preska Steinberg et al., 2022). The identification of gene-level sweeps in the flexible genome is done through a simple presence/absence-based method. Flexible genes found within a population cluster but not present in other clusters are considered candidates for gene-level sweeps in the population. For gene-level sweeps in core genomes, windows of 100 SNPs were used to scan the whole genome alignments of a population cluster for low intra-population nucleotide diversity regions. These regions were used to construct phylogenetic trees to search for genes or regions that would create monophyletic clades that match the distribution of genomes in all population clusters (Arevalo et al., 2019a). Thus, alleles that are found only within a monophyletic population but not in others are considered population-differentiating alleles.

The genes from regions that swept both the core and flexible components of the genome can now be used for a reverse ecology analysis (Arevalo et al., 2019a, 2019a; VanInsberghe et al., 2020). Population-differentiating SNPs, SNPs found in swept regions of a population but not in other distinct populations, can be used to study the changes in gene structure and function that allow the bacterial population to diverge ecologically. The SNPs may directly change the peptide and protein structure of the genes and thus affect their function, but they can also be subtle, such as changes in codon bias or transcription factor activation. A thorough study of the SNPs' effect (e.g., protein-ligand binding, RNA structure, etc.) could also provide a glimpse of the changing environment of the bacteria and its effect on its molecular functions. Otherwise, these SNPs can also be helpful as population-differentiating markers for ecological and environmental studies (Arevalo et al., 2019b, 2019a).

1.5.2. Modelling ecological populations

Exploring the biological properties that affect ecological populations' formation is best done by studying whole genomes from co-habiting closely related bacterial populations (Kashtan et al., 2014). These bacterial populations are ideal for examining bacterial speciation driven by discontinuities in gene flow as cohabitation, and the close genetic relatedness affects efficient homologous recombination (Arevalo et al., 2019a; Shapiro et al., 2012; Shapiro and Polz, 2015; VanInsberghe et al., 2020; Wiedenbeck and Cohan, 2011). Habitats with discrete spatial variations in their environmental conditions or drastic variations across time are also ideal for examining bacterial speciation, as varying degrees of environmental shifts are correlated with selective pressure that leads to bacterial speciation (Kassen and Rainey, 2004; Rainey and Travisano, 1998; Weltzer and Miller, 2013). However, such isolates are currently lacking

primarily due to difficulties in obtaining intra-specific heterogeneity representative of the environment.

1.5.2.1. Acquiring non-enriched genomes

Isolating co-habiting closely related bacterial genomes for studying ecological populations is significantly more complex than traditional isolation of type strains or species. The genomes and bacteria isolated should be axenic to minimise contamination, which could add false-positive SNPs and hinder the identification of locally collinear blocks. At the same time, isolated genomes must capture and reflect the intra-species diversity of the co-habiting natural bacterial communities. Traditional enrichment techniques used to isolate type strains and species can obscure the study of intra-population diversity and ecological populations in bacteria as an enrichment step can inadvertently select traits and genes favouring one or two genotypes, which can potentially mask intra-species diversity (Boynton and Greig, 2014; Steensels et al., 2019; Vaz-Moreira et al., 2011).

Other non-culturing methods (metagenome-assembled genomes and single-cell amplified genomes) that reflect the environment's bacterial diversity can also have other potential downsides. Metagenome-assembled genomes (MAGs) can provide higher throughput and a more accurate picture of community diversity than traditional culturing techniques (Alneberg et al., 2018). However, MAGs assembled using short-read sequences can often form chimeric genomes, which can mask the SNPs and the intra-species genomic diversity of the genomes (Sangwan et al., 2016). Single-cell amplified genomes (SAGs) can better reflect the intra-species genomic diversity of the natural bacterial populations but can also have its own disadvantages (Alneberg et al., 2018). SAGs allow high-throughput isolation of genomes but are expensive to perform and require high-quality genomic DNA to yield a good-quality genome (Stepanauskas, 2012). However, it can also result in highly discontiguous genomes, hindering the efficient identification of LCBs and ecological populations (Troell et al., 2016). A downside of SAGs, and even MAGs, is that live bacterial isolates cannot be isolated and cannot be used to verify the ecological properties hypothesised by the reverse ecology approach.

A method of isolating axenic bacteria without enrichment is through a direct serial dilution to extinction approach. The initial enrichment step in the traditional culturing method is disregarded, with environmental samples directly and serially diluted into the growth media. The serial dilution is done to ensure that the last well with growth comes from only a single cell and maintains an axenic bacterium (Henson et al., 2020; Yang et al., 2016). For this method to reflect the bacterial population diversity, high amounts of bacteria need to be isolated, and their genomes are sequenced. The axenicity of isolated bacteria can be confirmed through analysis of the sequenced genome. Significantly different genomes, such as those belonging to a different

genus, can be identified and separated using binning methods that search for discrepancies in genomic signatures (i.e. tetranucleotide composition or GC content).

1.5.2.2. The Taupō Volcanic Zone

Understanding biological systems and their processes is often done by examining the extremes or edge-cases that highlight deviations from the norm, thus providing insights on how the normal processes or systems work (Pérez-Gracia et al., 2010). Extreme biomes, such as those in geothermal springs, represent one apt edge case for understanding bacterial speciation. Geothermal springs often have extreme ranges of geochemical and geophysical conditions (i.e., pH, salinity, temperature) that vary between geothermal features and through time (Oliverio et al., 2018; Power et al., 2018; Sharp et al., 2014). The different level of environmental variability in geothermal springs acts as a selective pressure that shapes the structure and dynamics of its bacterial communities, which is ideal in testing models of population divergence and cohesion (Kassen and Rainey, 2004; Rainey and Travisano, 1998; Sharp et al., 2014). Unlike terrestrial environments, geothermal springs are often discrete habitats with distinct but similar geochemical profiles suitable for identifying changes in community structures and diversity trends (Power et al., 2018). The physical separation of each geothermal spring also allows for the effect of geographic dispersal to be distinguished from other ecological factors affecting speciation (Whitaker et al., 2005). Lastly, horizontal gene flow has been documented to occur frequently in bacterial communities for geothermal springs (Fuchsman et al., 2017; Klatt et al., 2011; Ward et al., 2018).

The Taupō Volcanic Zone (TVZ) is a geologically active region in the heart of the North Island of New Zealand and rich in geothermal features ideal for studying bacterial speciation. The subduction in the Pacific-Australian plate tectonic boundary causes intense magmatism, resulting in deep groundwater convection and the formation of the geophysically and geochemically heterogeneous geothermal features in the TVZ (Power et al., 2018). Several of these geothermal features experience temporal variability, resulting in changes in bacterial composition (Scott, 1994; Ward et al., 2017). Furthermore, a wide range of environmental variability can be observed across the geographically dispersed features (Power et al., 2018). Temporal variations in various environmental parameters are ideal for understanding bacterial communities' diversification and speciation (Kashtan et al., 2014). Therefore, the TVZ geothermal sites would also allow observation of the ecological populations' response to environmental variability and thus be used to determine if they are coherent with metazoans' ecological and evolutionary patterns as described in BSC.

1.5.2.3. Phylum Aquificota

Among the bacteria found in TVZ, members of *Aquificota* present themselves as a suitable model for the observations of gene-level and genome-wide sweeps and the study of the formation of ecological populations. The phylum *Aquificota* and members of order *Aquificales* and *Hydrogenothermales* dominate a large number of the geothermal springs in TVZ, with an average of 31% relative abundance across the springs sampled (Power et al., 2018). Out of the 19 described genera of *Aquificota*, six are found to be prevalent in the Taupō Volcanic Zone: *Hydrogenobacter*, *Thermocrinis*, *Sulfurihydrogenibium*, and among them are three of the most abundant genera in the TVZ, namely, *Venenivibrio*, *Hydrogenobaculum*, and *Aquifex*. Most *Aquificota* species are obligate chemolithoautotrophs that solely utilise oxidation of hydrogen gas (H2) and sulphur species as electron donors for energy production (Gupta and Lali, 2013; Huber and Eder, 2006; Hügler et al., 2007; Reysenbach, 2015) with a couple showing potential limited heterotrophic capabilities (e.g., *Thermocrinis ruber* and *Hydrogenobacter subterraneus*) (Huber and Eder, 2006; Takai et al., 2001). *Aquificota* thrives primarily in thermophilic or hyperthermophilic conditions, from 40°C and up to 95°C for some species (Gupta, 2014; Huber and Eder, 2006; Reysenbach, 2015). Additionally, the genome size of its members is very compact and ranges from around 1.4-2 million bases (Kawasumi et al., 1984; Wäber and Hartmann, 2019). Their relatively high abundance and highly limited growth parameters would allow precisely targeted isolation without enrichment, which was previously discussed as needed to generate high-quality genomes for PopCOGenT analysis (Arevalo et al., 2019a, 2019b).

Members of the phylum *Aquificota* are also edge-cases by themselves in terms of their evolutionary history and ecology. Phylogenetic studies of *Aquificota* have shown that its position is highly variable within the universal tree of life (Brown, 2004; Gupta, 2014; Oshima et al., 2012). Tree reconstruction using rRNA and protein synthesis genes has put the phyla among the earliest branches of bacteria near *Thermotoga*. In contrast, conserved sequence indels and GC content of *dnaK*, *rpoC* and other genes indicate the phylum to be late branching and closely grouped to *Epsiloproteobacteria* and *Bacteroidetes* (Boussau et al., 2008). These phylogenetic discrepancies have been attributed to the ancient horizontal gene transfer of many essential genes, such as the *rrn* operon, ribosomal proteins, *EF-Tu* and *EF-G* genes which are needed for adaptation in thermophilic environments (Gupta, 2014). Recent studies also show a high degree of linkage-disequilibrium and prevalence of mobile genetic elements in the genomes of the genus *Hydrogenobacter sp.* from natural bacterial communities (Bowers et al., 2022). However, horizontal gene transfer in highly specialised organisms such as *Aquficota* should have a significant fitness benefit to counteract the considerable risk of HGT, such as recombination of maladaptive or disadvantageous genes and alleles is detrimental to very compact genome and energy production mechanisms (Aris-Brosou, 2005; Jain et al., 1999; Novick and Doolittle, 2019).

1.5.3. Generalisability of Ecological population theory

Previous studies on delineating bacterial populations through horizontal gene flow or gene-specific sweeps have been primarily done on heterotrophic bacteria (Arevalo et al., 2019a; den Bakker et al., 2008; Szabo et al., 2013; Wang et al., 2022) or those with metabolic redundancies (Whitaker et al., 2005) and minimal energy stress (Kashtan et al., 2014; Ward et al., 2018). The ecotype theory asserts that horizontal gene flow and gene-specific sweeps are more likely to occur in ecologically heterogeneous populations (i.e., heterotrophs) brought by their ability to switch metabolic strategies or ecological niches (Cohan, 2016). Meanwhile, ecologically homogeneous populations with very narrow living conditions (i.e. specialised chemolithoautotrophs) are less likely to undergo gene flow and reach fixation within the population solely through genome-wide sweeps (Cohan, 2016). It can be argued that lack of metabolic redundancies meant multiple processes rely on a few or single genes (i.e. energy production pathway), thus making these genes highly entrenched and connected. The complexity hypothesis posits that these entrenched features are less likely to undergo modification (Aris-Brosou, 2005; Jain et al., 1999). Modifying highly connected features can lower fitness costs brought about by maladaptive mutations or disadvantageous recombination (Novick and Doolittle, 2019).

The ecotype theory's assertions still raise the question of how prevalent populations adhering to ecological population theory are in nature. Two methodologies can be employed to address this question. One method is to perform an extensive examination similar to Bobay and Ochman's study (2017) on genomes in NCBI databases but on natural bacterial populations across varied environments, ecological niches and diverse taxa. However, this method can be laborious and highly expensive. Alternatively, challenging the assertion of genome-wide sweep restrictions in ecologically homogeneous populations would suggest that gene-specific sweeps can occur in broader categories than previously thought. The logic is that if gene-specific sweeps occur in unlikely populations such as chemolithoautotrophs, then horizontal gene flow and ecological population theory's assertion of delineating species through discontinuities in horizontal gene flow applies not only to ecologically heterogeneous populations but also to other taxa, possibly across other natural bacterial populations.

1.6. Research Questions

Considering the above literature review, I propose that even with edge-case organisms like *Aquificota*, ecological populations can form in natural communities through horizontal gene flow. Therefore, this research aims to examine whether *Aquificota's* ecological populations from

natural bacterial communities can be delineated through horizontal gene flow and investigate whether these populations exhibit ecological and evolutionary patterns that are coherent with the BSC.

1.6.1. Chapter 2: Gene flow boundaries delineate natural bacterial populations of *Aquificaceae*

This chapter aims to answer whether intra-species genomic variations in chemolithoautotrophic *Aquificota* bacteria in natural communities are connected through horizontal gene flow. Using a high-throughput dilution-to-extinction method with minimal enrichment, I isolated closely related co-habiting populations of *Aquificota* from TVZ geothermal pools to test the null hypothesis that *Aquificota* populations do not undergo horizontal gene flow (or gene-level sweeps) due to their metabolic inflexibility. Non-enriched axenic isolates of *Aquificota* bacteria were sequenced and used for comparative genomic analysis to investigate intra-species genomic variations in natural *Aquificota* populations. The presence of horizontal gene flow through homologous recombination and analysis of the population structure of *Aquificota* bacteria were determined using the PopCOGenT pipeline.

1.6.2. Chapter 3: Adaptive mechanisms of metabolically limited bacteria in response to ecological perturbations

This chapter aims to examine the adaptive mechanisms utilised by *Aquificota* bacteria in response to ecological shifts. Through a year-long longitudinal study, I tested the hypothesis postulated by the ecotype theory in which metabolically constrained organisms such as *Aquificota* bacteria respond to ecological shifts solely through genome-wide sweeps (Cohan, 2016; Majewski and Cohan, 1999). To test this null hypothesis, I isolated and sequenced co-habiting *Aquificota* bacteria from TVZ geothermal pools experiencing shifts in their environmental conditions and determined changes in their intra-population diversity and population structure using comparative genomic analysis (i.e., ANI and PopCOGenT). Furthermore, I utilised metagenomic analysis to examine the changes in the relative abundance and intra-population diversity of *Aquificota* populations in natural bacterial communities.

1.6.3. Chapter 4: Homologous recombination drives horizontal gene flow in *Aquificota* populations

Chapter 4 aims to understand the role of non-homologous recombination in the bacterial speciation of *Aquificota* bacteria, motivated by the previously observed abundance of mobile

genetic elements in *Aquificota* bacteria (Bowers et al., 2022). With the assumption that homologous recombination is the primary mechanism in the formation of ecologically and genotypically distinct populations delineated by horizontal gene flow (Arevalo et al., 2019b; Shapiro et al., 2012), this chapter aims to test the hypothesis that non-homologous recombination mechanisms also play a significant role in horizontal gene flow-driven bacterial speciation in *Aquificota* bacteria. Thus, I compared the abundance of mobile genetic elements between recombining and clonal populations of *Aquificota* to determine correlations between MGE abundance and a recombining population structure. Additionally, I identified horizontally swept genes assumed to be driven by homologous recombination and compared their spatial and functional overlap with mobile genetic elements and bacterial host defence systems (i.e., CRISPR-Cas) that regulate against mobile genetic elements.

1.7. References

Achtman, M., Wagner, M., 2008. Microbial diversity and the genetic nature of microbial species. Nature Reviews Microbiology 6, 431–440. https://doi.org/10.1038/nrmicro1872

Alneberg, J., Karlsson, C.M.G., Divne, A.-M., Bergin, C., Homa, F., Lindh, M.V., Hugerth, L.W., Ettema, T.J.G., Bertilsson, S., Andersson, A.F., Pinhassi, J., 2018. Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. Microbiome 6, 173. https://doi.org/10.1186/s40168-018-0550-0

Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J., Polz, M.F., 2019a. A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. Cell 178, 820-834.e14. https://doi.org/10.1016/j.cell.2019.06.033

Arevalo, P., VanInsberghe, D., Polz, M.F., 2019b. A Reverse Ecology Framework for Bacteria and Archaea, in: Polz, M.F., Rajora, O.P. (Eds.), Population Genomics: Microorganisms, Population Genomics. Springer International Publishing, Cham, pp. 77–96. https://doi.org/10.1007/13836_2018_46

Aris-Brosou, S., 2005. Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis. Mol. Biol. Evol. 22, 200–209. https://doi.org/10.1093/molbev/msi006

Balakrishnan, R., 2005. Species Concepts, Species Boundaries and Species Identification: A View from the Tropics. Syst Biol 54, 689–693. https://doi.org/10.1080/10635150590950308

Barker, M.J., Wilson, R.A., 2010. COHESION, GENE FLOW, AND THE NATURE OF SPECIES. The Journal of Philosophy 107, 61–79.

Becq, J., Churlaud, C., Deschavanne, P., 2010. A Benchmark of Parametric Methods for Horizontal Transfers Detection. PLOS ONE 5, e9989. https://doi.org/10.1371/journal.pone.0009989

Becraft, E.D., Wood, J.M., Rusch, D.B., Kühl, M., Jensen, S.I., Bryant, D.A., Roberts, D.W., Frederick M. Cohan, Ward, D.M., 2015. The molecular dimension of microbial species: 1. Ecological distinctions among, and homogeneity within, putative ecotypes of Synechococcus inhabiting the cyanobacterial mat of Mushroom Spring, Yellowstone National Park. Front. Microbiol. 6. https://doi.org/10.3389/fmicb.2015.00590

Beier, S., Rivers, A.R., Moran, M.A., Obernosterer, I., 2015. Phenotypic plasticity in heterotrophic marine microbial communities in continuous cultures. The ISME Journal 9, 1141–1151. https://doi.org/10.1038/ismej.2014.206

Bendall, M.L., Stevens, S.L., Chan, L.-K., Malfatti, S., Schwientek, P., Tremblay, J., Schackwitz, W., Martin, J., Pati, A., Bushnell, B., Froula, J., Kang, D., Tringe, S.G., Bertilsson, S., Moran, M.A., Shade, A., Newton, R.J., McMahon, K.D., Malmstrom, R.R., 2016. Genome-

wide selective sweeps and gene-specific sweeps in natural bacterial populations. The ISME Journal 10, 1589–1601. https://doi.org/10.1038/ismej.2015.241

Bobay, L.-M., Ochman, H., 2018. Factors driving effective population size and pan-genome evolution in bacteria. BMC Evolutionary Biology 18, 153. https://doi.org/10.1186/s12862-018-1272-4

Bobay, L.-M., Ochman, H., 2017. Biological species are universal across Life's domains. Genome Biol Evol 9, 491–501. https://doi.org/10.1093/gbe/evx026

Boussau, B., Guéguen, L., Gouy, M., 2008. Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria. BMC Evol Biol 8, 272. https://doi.org/10.1186/1471-2148-8-272

Bowers, R.M., Nayfach, S., Schulz, F., Jungbluth, S.P., Ruhl, I.A., Sheremet, A., Lee, J., Goudeau, D., Eloe-Fadrosh, E.A., Stepanauskas, R., Malmstrom, R.R., Kyrpides, N.C., Dunfield, P.F., Woyke, T., 2022. Dissecting the dominant hot spring microbial populations based on community-wide sampling at single-cell genomic resolution. ISME J 16, 1337–1347. https://doi.org/10.1038/s41396-021-01178-4

Boynton, P.J., Greig, D., 2014. The ecology and evolution of non-domesticated Saccharomyces species. Yeast 31, 449–462. https://doi.org/10.1002/yea.3040

Brenner, D.J., Staley, J.T., Krieg, N.R., 2005. Classification of Procaryotic Organisms and the Concept of Bacterial Speciation, in: Brenner, D.J., Krieg, N.R., Staley, J.T., Garrity, G.M. (Eds.), Bergey's Manual® of Systematic Bacteriology: Volume Two: The Proteobacteria, Part A Introductory Essays. Springer US, Boston, MA, pp. 27–32. https://doi.org/10.1007/0-387-28021-9_4

Brown, J.R., 2004. Reconstructing the Universal Tree of Life, in: Ribas de Pouplana, L. (Ed.), The Genetic Code and the Origin of Life. Springer US, Boston, MA, pp. 15–33. https://doi.org/10.1007/0-387-26887-1_2

Burke, M.K., 2012. How does adaptation sweep through the genome? Insights from long-term selection experiments. Proceedings of the Royal Society B: Biological Sciences 279, 5029–5038. https://doi.org/10.1098/rspb.2012.0799

Carroll, L.M., Wiedmann, M., Kovac, J., 2021. Proposal of a Taxonomic Nomenclature for the Bacillus cereus Group Which Reconciles Genomic Definitions of Bacterial Species with Clinical and Industrial Phenotypes. mBio 11, e00034-20. https://doi.org/10.1128/mBio.00034-20

Chase, A.B., Arevalo, P., Brodie, E.L., Polz, M.F., Karaoz, U., Martiny, J.B.H., 2019. Maintenance of Sympatric and Allopatric Populations in Free-Living Terrestrial Bacteria. mBio 10. https://doi.org/10.1128/mBio.02361-19

Choi, S.C., Rasmussen, M.D., Hubisz, M.J., Gronau, I., Stanhope, M.J., Siepel, A., 2012. Replacing and Additive Horizontal Gene Transfer in Streptococcus. Mol Biol Evol 29, 3309–3320. https://doi.org/10.1093/molbev/mss138

Cohan, F.M., 2019. Transmission in the Origins of Bacterial Diversity, From Ecotypes to Phyla, in: Microbial Transmission. John Wiley & Sons, Ltd, pp. 311–343. https://doi.org/10.1128/9781555819743.ch18

Cohan, F.M., 2016. Bacterial Speciation: Genetic Sweeps in Bacterial Species. Current Biology 26, R112–R115. https://doi.org/10.1016/j.cub.2015.10.022

Cohan, F.M., 2011. Are Species Cohesive?—A View from Bacteriology, in: Population Genetics of Bacteria. John Wiley & Sons, Ltd, pp. 43–65. https://doi.org/10.1128/9781555817114.ch5

Cohan, F.M., 2002. What are Bacterial Species? Annual Review of Microbiology 56, 457–487. https://doi.org/10.1146/annurev.micro.56.012302.160634

Cohan, F.M., 2001. Bacterial Species and Speciation. Syst Biol 50, 513–524. https://doi.org/10.1080/10635150118398

Cohan, F.M., Koeppel, A., 2008. The Origins of Ecological Diversity in Prokaryotes. Current Biology 18, R1024–R1034. https://doi.org/10.1016/j.cub.2008.09.014

Cohan, F.M., Perry, E.B., 2007. A Systematics for Discovering the Fundamental Units of Bacterial Diversity. Current Biology 17, R373–R386. https://doi.org/10.1016/j.cub.2007.03.032

Cordero, O.X., Polz, M.F., 2014. Explaining microbial genomic diversity in light of evolutionary ecology. Nature Reviews Microbiology 12, 263–273. https://doi.org/10.1038/nrmicro3218

Darling, A.C.E., Mau, B., Blattner, F.R., Perna, N.T., 2004. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Res. 14, 1394–1403. https://doi.org/10.1101/gr.2289704

Das, S., Dash, H.R., Mangwani, N., Chakraborty, J., Kumari, S., 2014. Understanding molecular identification and polyphasic taxonomic approaches for genetic relatedness and phylogenetic relationships of microorganisms. Journal of Microbiological Methods 103, 80–100. https://doi.org/10.1016/j.mimet.2014.05.013

den Bakker, H.C., Didelot, X., Fortes, E.D., Nightingale, K.K., Wiedmann, M., 2008. Lineage specific recombination rates and microevolution in Listeria monocytogenes. BMC Evolutionary Biology 8, 277. https://doi.org/10.1186/1471-2148-8-277

Doolittle, W.F., 2012. Population Genomics: How Bacterial Species Form and Why They Don't Exist. Current Biology 22, R451–R453. https://doi.org/10.1016/j.cub.2012.04.034

Doolittle, W.F., Papke, R.T., 2006. Genomics and the bacterial species problem. Genome Biology 7, 116. https://doi.org/10.1186/gb-2006-7-9-116

Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G., Hanage, W.P., 2009. The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity. Science 323, 741–746. https://doi.org/10.1126/science.1159388

Fraser, C., Hanage, W.P., Spratt, B.G., 2007. Recombination and the Nature of Bacterial Speciation. Science 315, 476–480. https://doi.org/10.1126/science.1127573

Fuchsman, C.A., Collins, R.E., Rocap, G., Brazelton, W.J., 2017. Effect of the environment on horizontal gene transfer between bacteria and archaea. PeerJ 5, e3865. https://doi.org/10.7717/peerj.3865

Gupta, R.S., 2014. The Phylum Aquificae, in: Rosenberg, E., DeLong, E.F., Lory, S., Stackebrandt, E., Thompson, F. (Eds.), The Prokaryotes: Other Major Lineages of Bacteria and The Archaea. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 417–445. https://doi.org/10.1007/978-3-642-38954-2_119

Gupta, R.S., Lali, R., 2013. Molecular signatures for the phylum Aquificae and its different clades: proposal for division of the phylum Aquificae into the emended order Aquificales, containing the families Aquificaceae and Hydrogenothermaceae, and a new order Desulfurobacteriales ord. nov., containing the family Desulfurobacteriaceae. Antonie van Leeuwenhoek 104, 349–368. https://doi.org/10.1007/s10482-013-9957-6

Hanage, W.P., Fraser, C., Spratt, B.G., 2005. Fuzzy species among recombinogenic bacteria. BMC Biology 3, 6. https://doi.org/10.1186/1741-7007-3-6

Henson, M.W., Lanclos, V.C., Pitre, D.M., Weckhorst, J.L., Lucchesi, A.M., Cheng, C., Temperton, B., Thrash, J.C., 2020. Expanding the Diversity of Bacterioplankton Isolates and Modeling Isolation Efficacy with Large-Scale Dilution-to-Extinction Cultivation. Appl Environ Microbiol 86, e00943-20. https://doi.org/10.1128/AEM.00943-20

Hetzer, A., McDonald, I.R., Morgan, H.W., 2008. Venenivibrio stagnispumantis gen. nov., sp. nov., a thermophilic hydrogen-oxidizing bacterium isolated from Champagne Pool, Waiotapu, New Zealand. International Journal of Systematic and Evolutionary Microbiology 58, 398–403. https://doi.org/10.1099/ijs.0.64842-0

Hua, S., Qiu, M., Chan, E., Zhu, L., Luo, Y., 1997. Minimum Length of Sequence Homology Required for*in Vivo*Cloning by Homologous Recombination in Yeast. Plasmid 38, 91–96. https://doi.org/10.1006/plas.1997.1305

Huber, R., Eder, W., 2006. Aquificales, in: Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K.-H., Stackebrandt, E. (Eds.), The Prokaryotes: Volume 7: Proteobacteria: Delta, Epsilon Subclass. Springer New York, New York, NY, pp. 925–938. https://doi.org/10.1007/0-387-30747-8_39

Hügler, M., Huber, H., Molyneaux, S.J., Vetriani, C., Sievert, S.M., 2007. Autotrophic CO2 fixation via the reductive tricarboxylic acid cycle in different lineages within the phylum Aquificae: evidence for two ways of citrate cleavage. Environmental Microbiology 9, 81–92. https://doi.org/10.1111/j.1462-2920.2006.01118.x

Hull, D.L., 1977. The Ontological Status of Species as Evolutionary Units, in: Butts, R.E., Hintikka, J. (Eds.), Foundational Problems in the Special Sciences: Part Two of the Proceedings of the Fifth International Congress of Logic, Methodology and Philosophy of Science, London, Ontario, Canada-1975, The University of Western Ontario Series in Philosophy of Science. Springer Netherlands, Dordrecht, pp. 91–102. https://doi.org/10.1007/978-94-010-1141-9_6

Hülter, N., Wackernagel, W., 2008. Double illegitimate recombination events integrate DNA segments through two different mechanisms during natural transformation of Acinetobacter baylyi. Mol Microbiol 67, 984–995. https://doi.org/10.1111/j.1365-2958.2007.06096.x

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., Aluru, S., 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9, 5114. https://doi.org/10.1038/s41467-018-07641-9

Jain, R., Rivera, M.C., Lake, J.A., 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. U.S.A. 96, 3801–3806.

Janda, J.M., Abbott, S.L., 2007. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. Journal of Clinical Microbiology 45, 2761–2764. https://doi.org/10.1128/JCM.01228-07

Johnson, J.L., 1985. 2 DNA Reassociation and RNA Hybridisation of Bacterial Nucleic Acids, in: Methods in Microbiology. Elsevier, pp. 33–74. https://doi.org/10.1016/S0580-9517(08)70471-9

Johnson, J.L., 1980. The use of DNA homology in bacterial taxonomy and identification. Clinical Microbiology Newsletter 2, 1–3. https://doi.org/10.1016/S0196-4399(80)80006-7

Kashtan, N., Roggensack, S.E., Rodrigue, S., Thompson, J.W., Biller, S.J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R.R., Stocker, R., Follows, M.J., Stepanauskas, R., Chisholm, S.W., 2014. Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus. Science 344, 416–420. https://doi.org/10.1126/science.1248575

Kassen, R., Rainey, P.B., 2004. The Ecology and Genetics of Microbial Diversity. Annu. Rev. Microbiol. 58, 207–231. https://doi.org/10.1146/annurev.micro.58.030603.123654

Kawasumi, T., Igarashi, Y., Kodama, T., Minoda, Y., 1984. Hydrogenobacter thermophilus gen. nov., sp. nov., an Extremely Thermophilic, Aerobic, Hydrogen-Oxidizing Bacterium. International Journal of Systematic and Evolutionary Microbiology 34, 5–10. https://doi.org/10.1099/00207713-34-1-5

Khayi, S., Blin, P., Pédron, J., Chong, T.-M., Chan, K.-G., Moumni, M., Hélias, V., Van Gijsegem, F., Faure, D., 2015. Population genomics reveals additive and replacing horizontal gene transfers in the emerging pathogen Dickeya solani. BMC Genomics 16, 788. https://doi.org/10.1186/s12864-015-1997-z

Klatt, C.G., Wood, J.M., Rusch, D.B., Bateson, M.M., Hamamura, N., Heidelberg, J.F., Grossman, A.R., Bhaya, D., Frederick M. Cohan, Kühl, M., Bryant, D.A., Ward, D.M., 2011. Community ecology of hot spring cyanobacterial mats: predominant populations and their functional potential. ISME J 5, 1262–1278. https://doi.org/10.1038/ismej.2011.73

Koeppel, A.F., Perry, E.B., Sikorski, J., Krizanc, D., Warner, A., Ward, D.M., Rooney, A.P., Brambilla, E., Connor, N., Ratcliff, R.M., Nevo, E., Frederick M. Cohan, 2008. Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics. PNAS 105, 2504–2509. https://doi.org/10.1073/pnas.0712205105

Konstantinidis, K.T., Ramette, A., Tiedje, J.M., 2006. The Bacterial Species Definition in the Genomic Era. Philosophical Transactions: Biological Sciences 361, 1929–1940.

Konstantinidis, K.T., Tiedje, J.M., 2007. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. Current Opinion in Microbiology, Antimicrobials/Genomics 10, 504–509. https://doi.org/10.1016/j.mib.2007.08.006

Kopac, S., Wang, Z., Wiedenbeck, J., Sherry, J., Wu, M., Frederick M. Cohan, 2014. Genomic Heterogeneity and Ecological Speciation within One Subspecies of Bacillus subtilis. Appl. Environ. Microbiol. 80, 4842–4853. https://doi.org/10.1128/AEM.00576-14

Krieg, N.R., 1988. Bacterial classification: an overview. Canadian Journal of Microbiology 34, 536–540. https://doi.org/10.1139/m88-091

Lan, R., Reeves, P.R., 2000. Intraspecies variation in bacterial genomes: the need for a species genome concept. Trends in Microbiology 8, 396–401. https://doi.org/10.1016/S0966-842X(00)01791-1

Majewski, J., Cohan, F.M., 1999. Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. Genetics 152, 1459–1474.

Mayr, E., 2014. Animal Species and Evolution.

Mayr, E., 1999. Systematics and the origin of species, from the viewpoint of a zoologist, 1st Harvard University Press pbk. ed. ed. Harvard University Press, Cambridge, Mass.

Mayr, E., 1975. Populations, species and evolution: an abridgement of Animal species and evolution, 4. print. ed. Belknap Press of Harvard Univ. Press, Cambridge, Mass.

Mayr, E., 1954. Change of genetic environment and evolution, in: Huxley, J.U.L.I.A.N., Hardy, A.C., Ford, E.B. (Eds.), Evolution as a Process. Allen and Unwin, London, UK, pp. 156–180.

Melendrez, M.C., Becraft, E.D., Wood, J.M., Olsen, M.T., Bryant, D.A., Heidelberg, J.F., Rusch, D.B., Frederick M. Cohan, Ward, D.M., 2016. Recombination Does Not Hinder Formation or Detection of Ecological Species of Synechococcus Inhabiting a Hot Spring Cyanobacterial Mat. Front. Microbiol. 6. https://doi.org/10.3389/fmicb.2015.01540

Mell, J.C., Redfield, R.J., 2014. Natural Competence and the Evolution of DNA Uptake Specificity. Journal of Bacteriology 196, 1471–1483. https://doi.org/10.1128/jb.01293-13

Messer, P.W., Petrov, D.A., 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends in Ecology & Evolution 28, 659–669. https://doi.org/10.1016/j.tree.2013.08.003

Meziti, A., Tsementzi, D., Rodriguez-R, L.M., Hatt, J.K., Karayanni, H., Kormas, K.A., Konstantinidis, K.T., 2019. Quantifying the changes in genetic diversity within sequence-discrete bacterial populations across a spatial and temporal riverine gradient. ISME J 13, 767–779. https://doi.org/10.1038/s41396-018-0307-6

Miller, S.R., Carvey, D., 2019. Ecological Divergence with Gene Flow in a Thermophilic Cyanobacterium. Microb Ecol 78, 33–41. https://doi.org/10.1007/s00248-018-1267-0

Minkin, I., Medvedev, P., 2020. Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. Nat Commun 11, 6327. https://doi.org/10.1038/s41467-020-19777-8

Miyazaki, K., Tomariguchi, N., 2019. Occurrence of randomly recombined functional 16S rRNA genes in Thermus thermophilus suggests genetic interoperability and promiscuity of bacterial 16S rRNAs. Sci Rep 9, 1–10. https://doi.org/10.1038/s41598-019-47807-z

Murray, C.S., Gao, Y., Wu, M., 2021. Re-evaluating the evidence for a universal genetic boundary among microbial species. Nat Commun 12, 4059. https://doi.org/10.1038/s41467-021-24128-2

Nagler, M., Podmirseg, S.M., Ascher-Jenull, J., Sint, D., Traugott, M., 2022. Why eDNA fractions need consideration in biomonitoring. Mol Ecol Resour 22, 2458–2470. https://doi.org/10.1111/1755-0998.13658

Norman, A., Hansen, L.H., Sørensen, S.J., 2009. Conjugative plasmids: vessels of the communal gene pool. Philosophical Transactions of the Royal Society B: Biological Sciences 364, 2275–2289. https://doi.org/10.1098/rstb.2009.0037

Novick, A., Doolittle, W.F., 2019. Horizontal persistence and the complexity hypothesis. Biol Philos 35, 2. https://doi.org/10.1007/s10539-019-9727-6

Oakley, B.B., Carbonero, F., van der Gast, C.J., Hawkins, R.J., Purdy, K.J., 2010. Evolutionary divergence and biogeography of sympatric niche-differentiated bacterial populations. ISME J 4, 488–497. https://doi.org/10.1038/ismej.2009.146

Olendzenski, L., Zhaxybayeva, O., Gogarten, P., 2006. Orthologs, Paralogs and Xenologs in Human and Other Genomes, in: Encyclopedia of Life Sciences. Wiley. https://doi.org/10.1038/npg.els.0005298

Oliverio, A.M., Power, J.F., Washburne, A., Cary, S.C., Stott, M.B., Fierer, N., 2018. The ecology and diversity of microbial eukaryotes in geothermal springs. ISME J 12, 1918–1928. https://doi.org/10.1038/s41396-018-0104-2

Oshima, K., Chiba, Y., Igarashi, Y., Arai, H., Ishii, M., 2012. Phylogenetic Position of Aquificales Based on the Whole Genome Sequences of Six Aquificales Species [WWW Document]. International Journal of Evolutionary Biology. https://doi.org/10.1155/2012/859264

Palys, T., Nakamura, L.K., Cohan, F.M., 1997. Discovery and Classification of Ecological Diversity in the Bacterial World: The Role of DNA Sequence Data. International Journal of Systematic and Evolutionary Microbiology 47, 1145–1156. https://doi.org/10.1099/00207713-47-4-1145

Pérez-Gracia, J.L., Gúrpide, A., Ruiz-Ilundain, M.G., Alegría, C.A., Colomer, R., García-Foncillas, J., Bermejo, I.M., 2010. Selection of extreme phenotypes: the role of clinical observation in translational research. Clin Transl Oncol 12, 174–180.

Petit, R.J., Excoffier, L., 2009. Gene flow and species delimitation. Trends in Ecology & Evolution 24, 386–393. https://doi.org/10.1016/j.tree.2009.02.011

Polz, M.F., Alm, E.J., Hanage, W.P., 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends in Genetics 29, 170–175. https://doi.org/10.1016/j.tig.2012.12.006

Power, J.F., Carere, C.R., Lee, C.K., Wakerley, G.L.J., Evans, D.W., Button, M., White, D., Climo, M.D., Hinze, A.M., Morgan, X.C., McDonald, I.R., Cary, S.C., Stott, M.B., 2018. Microbial biogeography of 925 geothermal springs in New Zealand. Nat Commun 9, 2876. https://doi.org/10.1038/s41467-018-05020-y

Preska Steinberg, A., Lin, M., Kussell, E., 2022. Core genes can have higher recombination rates than accessory genes within global microbial populations. eLife 11, e78533. https://doi.org/10.7554/eLife.78533

Queiroz, K. de, 2005a. Ernst Mayr and the modern concept of species. PNAS 102, 6600–6607. https://doi.org/10.1073/pnas.0502030102

Queiroz, K. de, 2005b. Different species problems and their resolution. BioEssays 27, 1263–1269. https://doi.org/10.1002/bies.20325

Raina, V., Nayak, T., Ray, L., Kumari, K., Suar, M., 2019. Chapter 9 - A Polyphasic Taxonomic Approach for Designation and Description of Novel Microbial Species, in: Das, S., Dash, H.R. (Eds.), Microbial Diversity in the Genomic Era. Academic Press, pp. 137–152. https://doi.org/10.1016/B978-0-12-814849-5.00009-5

Rainey, P.B., Travisano, M., 1998. Adaptive radiation in a heterogeneous environment. Nature 394, 69–72. https://doi.org/10.1038/27900

Ravenhall, M., Škunca, N., Lassalle, F., Dessimoz, C., 2015. Inferring Horizontal Gene Transfer. PLOS Computational Biology 11, e1004095. https://doi.org/10.1371/journal.pcbi.1004095

Reysenbach, A.-L., 2015. *Aquificales ord. nov*, in: Whitman, W.B., Rainey, F., Kämpfer, P., Trujillo, M., Chun, J., DeVos, P., Hedlund, B., Dedysh, S. (Eds.), Bergey's Manual of Systematics of Archaea and Bacteria. John Wiley & Sons, Ltd, Chichester, UK, pp. 1–1. https://doi.org/10.1002/9781118960608.obm00027

Rieseberg, L.H., Church, S.A., Morjan, C.L., 2004. Integration of populations and differentiation of species. New Phytologist 161, 59–69. https://doi.org/10.1046/j.1469-8137.2003.00933.x

Rodriguez-R, L.M., Konstantinidis, K.T., 2014. Bypassing Cultivation To Identify Bacterial Species: Culture-independent genomic approaches identify credibly distinct clusters, avoid cultivation bias, and provide true insights into microbial species. Microbe Magazine 9, 111–118. https://doi.org/10.1128/microbe.9.111.1

Rosen, M.J., Davison, M., Bhaya, D., Fisher, D.S., 2015. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science 348, 1019–1023. https://doi.org/10.1126/science.aaa4456

Rosselló-Mora, R., 2006. DNA-DNA Reassociation Methods Applied to Microbial Taxonomy and Their Critical Evaluation, in: Stackebrandt, E. (Ed.), Molecular Identification, Systematics, and Population Structure of Prokaryotes. Springer, Berlin, Heidelberg, pp. 23–50. https://doi.org/10.1007/978-3-540-31292-5_2

Rosselló-Móra, R., Urdiain, M., López-López, A., 2011. DNA–DNA Hybridization, in: Methods in Microbiology. Elsevier, pp. 325–347. https://doi.org/10.1016/B978-0-12-387730-7.00015-2

Sangwan, N., Xia, F., Gilbert, J.A., 2016. Recovering complete and draft population genomes from metagenome datasets. Microbiome 4, 8. https://doi.org/10.1186/s40168-016-0154-5

Schaller, D., Lafond, M., Stadler, P.F., Wieseke, N., Hellmuth, M., 2021. Indirect identification of horizontal gene transfer. J. Math. Biol. 83, 10. https://doi.org/10.1007/s00285-021-01631-0

Scott, B.J., 1994. Cyclic activity in the crater lakes of Waimangu hydrothermal system, New Zealand. Geothermics 23, 555–572. https://doi.org/10.1016/0375-6505(94)90019-1

Shapiro, B.J., David, L.A., Friedman, J., Alm, E.J., 2009. Looking for Darwin's footprints in the microbial world. Trends in Microbiology 17, 196–204. https://doi.org/10.1016/j.tim.2009.02.002

Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabó, G., Polz, M.F., Alm, E.J., 2012. Population Genomics of Early Events in the Ecological Differentiation of Bacteria. Science 336, 48–51. https://doi.org/10.1126/science.1218198

Shapiro, B.J., Leducq, J.-B., Mallet, J., 2016. What Is Speciation? PLOS Genetics 12, e1005860. https://doi.org/10.1371/journal.pgen.1005860

Shapiro, B.J., Polz, M.F., 2015. Microbial Speciation. Cold Spring Harb Perspect Biol 7, a018143. https://doi.org/10.1101/cshperspect.a018143

Shapiro, B.J., Polz, M.F., 2014. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol 22, 235–247. https://doi.org/10.1016/j.tim.2014.02.006

Sharp, C.E., Brady, A.L., Sharp, G.H., Grasby, S.E., Stott, M.B., Dunfield, P.F., 2014. Humboldt's spa: microbial diversity is controlled by temperature in geothermal environments. ISME J 8, 1166–1174. https://doi.org/10.1038/ismej.2013.237

Sites, J.W., Marshall, J.C., 2003. Delimiting species: a Renaissance issue in systematic biology. Trends in Ecology & Evolution 18, 462–470. https://doi.org/10.1016/S0169-5347(03)00184-8

Slatkin, M., 1987. Gene flow and the geographic structure of natural populations. Science 236, 787–792. https://doi.org/10.1126/science.3576198

Soucy, S., Olendzenski, L., Gogarten, J.P., 2013. Orthologues, Paralogues and Horizontal Gene Transfer in the Human Holobiont, in: ELS. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470015902.a0005298.pub3

Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A.D., Kämpfer, P., Maiden, M.C.J., Nesme, X., Rosselló-Mora, R., Swings, J., Trüper, H.G., Vauterin, L., Ward, A.C., Whitman, W.B., 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. International Journal of Systematic and Evolutionary Microbiology 52, 1043–1047. https://doi.org/10.1099/00207713-52-3-1043

Stackebrandt, E., Goebel, B.M., 1994. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. International Journal of Systematic and Evolutionary Microbiology 44, 846–849. https://doi.org/10.1099/00207713-44-4-846

Staley, J.T., 2006. The bacterial species dilemma and the genomic–phylogenetic species concept. Philos Trans R Soc Lond B Biol Sci 361, 1899–1909. https://doi.org/10.1098/rstb.2006.1914

Steensels, J., Gallone, B., Voordeckers, K., Verstrepen, K.J., 2019. Domestication of Industrial Microbes. Current Biology 29, R381–R393. https://doi.org/10.1016/j.cub.2019.04.025

Stepanauskas, R., 2012. Single cell genomics: an individual look at microbes. Current Opinion in Microbiology, Antimicrobials • Genomics 15, 613–620. https://doi.org/10.1016/j.mib.2012.09.001

Szabo, G., Preheim, S.P., Kauffman, K.M., David, L.A., Shapiro, J., Alm, E.J., Polz, M.F., 2013. Reproducibility of Vibrionaceae population structure in coastal bacterioplankton. ISME J 7, 509–519. https://doi.org/10.1038/ismej.2012.134

Takai, K., Horikoshi, K., Komatsu, T., 2001. Hydrogenobacter subterraneus sp. nov., an extremely thermophilic, heterotrophic bacterium unable to grow on hydrogen gas, from deep subsurface geothermal water. International Journal of Systematic and Evolutionary Microbiology 51, 1425–1435. https://doi.org/10.1099/00207713-51-4-1425

Takeuchi, N., Cordero, O.X., Koonin, E.V., Kaneko, K., 2015. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. BMC Biology 13, 20. https://doi.org/10.1186/s12915-015-0131-7

Thomas, C.M., Nielsen, K.M., 2005. Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. Nat Rev Microbiol 3, 711–721. https://doi.org/10.1038/nrmicro1234

Troell, K., Hallström, B., Divne, A.-M., Alsmark, C., Arrighi, R., Huss, M., Beser, J., Bertilsson, S., 2016. Cryptosporidium as a testbed for single cell genome characterization of unicellular eukaryotes. BMC Genomics 17, 471. https://doi.org/10.1186/s12864-016-2815-y

Vandamme, P., Pot, B., Gillis, M., Vos, P. de, Kersters, K., Swings, J., 1996. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol. Mol. Biol. Rev. 60, 407–438.

VanInsberghe, D., Arevalo, P., Chien, D., Polz, M.F., 2020. How can microbial population genomics inform community ecology? Philosophical Transactions of the Royal Society B: Biological Sciences 375, 20190253. https://doi.org/10.1098/rstb.2019.0253

Vaz-Moreira, I., Egas, C., Nunes, O.C., Manaia, C.M., 2011. Culture-dependent and culture-independent diversity surveys target different bacteria: a case study in a freshwater sample. Antonie van Leeuwenhoek 100, 245–257. https://doi.org/10.1007/s10482-011-9583-0

Viver, T., Conrad, R.E., Orellana, L.H., Urdiain, M., González-Pastor, J.E., Janet K. Hatt, Amann, R., Antón, J., Konstantinos T. Konstantinidis, Ramon Rosselló-Móra, 2021. Distinct ecotypes within a natural haloarchaeal population enable adaptation to changing environmental conditions without causing population sweeps. The ISME Journal 15, 1178–1191. https://doi.org/10.1038/s41396-020-00842-5

Volkova, V.V., Lu, Z., Besser, T., Gröhn, Y.T., 2014. Modeling the Infection Dynamics of Bacteriophages in Enteric Escherichia coli: Estimating the Contribution of Transduction to Antimicrobial Gene Spread. Applied and Environmental Microbiology 80, 4350–4362. https://doi.org/10.1128/AEM.00446-14

von Wintersdorff, C.J.H., Penders, J., van Niekerk, J.M., Mills, N.D., Majumder, S., van Alphen, L.B., Savelkoul, P.H.M., Wolffs, P.F.G., 2016. Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. Front. Microbiol. 7. https://doi.org/10.3389/fmicb.2016.00173

Vos, M., Didelot, X., 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J 3, 199–208. https://doi.org/10.1038/ismej.2008.93

Wäber, N.B., Hartmann, R.K., 2019. Aquificae, in: Schmidt, T.M. (Ed.), Encyclopedia of Microbiology (Fourth Edition). Academic Press, Oxford, pp. 226–233. https://doi.org/10.1016/B978-0-12-809633-8.20783-8

Wang, J., Li, Y., Pinto-Tomás, A.A., Cheng, K., Huang, Y., 2022. Habitat Adaptation Drives Speciation of a *Streptomyces* Species with Distinct Habitats and Disparate Geographic Origins. mBio 13, e02781-21. https://doi.org/10.1128/mbio.02781-21

Ward, L., Taylor, M.W., Power, J.F., Scott, B.J., McDonald, I.R., Stott, M.B., 2017. Microbial community dynamics in Inferno Crater Lake, a thermally fluctuating geothermal spring. ISME J 11, 1158–1167. https://doi.org/10.1038/ismej.2016.193

Ward, L.M., Hemp, J., Shih, P.M., McGlynn, S.E., Fischer, W.W., 2018. Evolution of Phototrophy in the Chloroflexi Phylum Driven by Horizontal Gene Transfer. Front. Microbiol. 9. https://doi.org/10.3389/fmicb.2018.00260

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., Moore, L.H., Moore, W.E.C., Murray, R.G.E., Stackebrandt, E., Starr, M.P., Truper, H.G.Y. 1987, 1987. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. International Journal of Systematic and Evolutionary Microbiology, 37, 463–464. https://doi.org/10.1099/00207713-37-4-463

Weltzer, M.L., Miller, S.R., 2013. Ecological Divergence of a Novel Group of Chloroflexus Strains along a Geothermal Gradient. Applied and Environmental Microbiology 79, 1353–1358. https://doi.org/10.1128/AEM.02753-12

Whitaker, R.J., Grogan, D.W., Taylor, J.W., 2005. Recombination shapes the natural population structure of the hyperthermophilic archaeon Sulfolobus islandicus. Mol. Biol. Evol. 22, 2354–2361. https://doi.org/10.1093/molbev/msi233

Wiedenbeck, J., Cohan, F.M., 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35, 957–976. https://doi.org/10.1111/j.1574-6976.2011.00292.x

Yang, S.-J., Kang, I., Cho, J.-C., 2016. Expansion of Cultured Bacterial Diversity by Large-Scale Dilution-to-Extinction Culturing from a Single Seawater Sample. Microb Ecol 71, 29–43. https://doi.org/10.1007/s00248-015-0695-3

Zarei-Baygi, A., Smith, A.L., 2021. Intracellular versus extracellular antibiotic resistance genes in the environment: Prevalence, horizontal transfer, and mitigation strategies. Bioresource Technology 319, 124181. https://doi.org/10.1016/j.biortech.2020.124181

CHAPTER 2: Gene flow boundaries delineate natural bacterial populations of Aquificota

2.1. Abstract

Defining and delineating a biologically meaningful definition of a bacterial species is impeded by our poor understanding of gene flow and the mechanisms that form an ecologically cohesive population in natural bacterial communities. Competing models of bacterial speciation proposed differing mechanisms (i.e. gene-specific and genome-wide sweep) for forming ecologically cohesive populations based on the prevalence of horizontal gene flow. Unfortunately, the prevalence and role of these mechanisms in natural bacterial communities are still not fully understood. Examining the general applicability of these mechanisms in natural bacterial communities requires an understanding of the role of horizontal gene flow in the formation of ecologically cohesive populations, especially in unlikely situations such as in chemolithoautotrophs. Horizontal gene flow and gene-specific sweeps are suggested to be most unlikely to occur in ecologically homogeneous populations such as chemolithoautotrophs as lack of metabolic flexibility can impart a high risk of deleterious consequences during horizontal gene flow. Therefore, metabolically limited bacteria are often assumed to have minimal horizontal gene flow and form ecologically cohesive populations through genome-wide sweeps. In this study, I use thermophilic chemolithoautotrophic hydrogen-oxidising *Aquificota* bacteria as a model to test this hypothesis. Using a novel high-throughput isolation technique with minimal enrichment, I isolated co-habiting closely related *Aquificota* bacteria. I sequenced their genomes to examine the role of horizontal gene flow in forming intra-species genomic variations. Phylogenomic analysis reveals three distinct populations of the undescribed genus *UBA11096* sp. Analysis of population heterogeneity and recent horizontal gene flow showed the populations have a high variation of intrapopulation heterogeneity driven by varying degrees of horizontal gene flow. Population structure analysis also showed that *Aquificota* populations were comprised of both clonal and recombining bacteria. The presence of both clonal lineages and horizontally-linked clusters of bacteria within a closely related taxa indicates that interplay between gene-specific and genome-wide sweeps simultaneously drives the formation of ecologically cohesive populations in natural bacterial communities. Additionally, the co-occurrence of both mechanisms in Aquificota suggests the generalisability of gene-specific sweeps as a mechanism of speciation across bacteria.

2.2. Introduction

The Biological Species Concept (BSC) is the most widely accepted concept of speciation for sexual metazoans (Barker and Wilson, 2010; Queiroz, 2005a). Unfortunately, there is no such universal definition for bacterial species (Hanage et al., 2005; Queiroz, 2005b). One major hurdle is our lack of understanding of the role of different gene flow mechanisms in the formation and delineation of ecologically cohesive bacterial populations (Mayr, 1999; Petit and Excoffier, 2009; Queiroz, 2005a). Without a clear understanding of the formation of ecologically cohesive populations, long-established ecological theories cannot be easily applied in microbial ecosystems (Shoemaker et al., 2017). In sexual metazoans, the formation of ecologically cohesive populations is linked to reproduction (Majewski, 2001) since adaptive genes are vertically passed from interbreeding parents to their offspring (Cohan, 2002; Eldredge, 1986; Mayr, 1982; Nosil, 2012). As such, divergence often accompanies ecological differentiation in sexual metazoans; thus, phenotypic and genomic similarities can be used to delineate ecologically cohesive populations (de Queiroz and Donoghue, 1988; Zuckerkandl and Pauling, 1965). However, identifying ecologically cohesive populations in bacteria is more complex since they can acquire adaptive genes through both vertical and horizontal gene flow (Cohan, 2002, p. 200; Cohan and Perry, 2007b; Koeppel et al., 2008; Shapiro et al., 2016). Current studies on the genome sequence diversity of natural bacterial communities show that bacteria form into sequence-discrete populations – an ecologically cohesive and genotypically distinct population akin to a eukaryotic species (Jain et al., 2018; Konstantinidis et al., 2022; Rodriguez-R et al., 2021). The process of their formation is still often debated (Cohan, 2016, 2002; Hanage et al., 2005; VanInsberghe et al., 2020). Even with the increasing observations of sequence-discrete populations, the role of horizontal gene flow in their formation in natural bacterial communities is still in question.

Sequence-discrete populations have been observed to form through both the interaction of mutation and selection (Becraft et al., 2015; Cohan and Perry, 2007b; Smith et al., 2006; Wiedenbeck and Cohan, 2011) or discontinuities in horizontal gene flow (Chase et al., 2019; Fraser et al., 2007; Friedman et al., 2013; Hanage et al., 2005; Shapiro and Polz, 2015). The prevailing ecotype theory suggests that horizontal gene flow is rare in natural bacterial communities (Cohan, 2016; Majewski and Cohan, 1999). Under this theory, the formation of sequence-discrete populations occurs through genome-wide sweeps in which periodic selection allows an adaptive mutant to survive and clonally expand within the population — forming a genetically and ecologically cohesive cluster or ecotype. In rare instances, recombination events allow an adaptive mutation to be shared across multiple ecotypes. Thus, multiple ecotypes survive similar selection events and lead to what can be observed as intrapopulation genomic diversity (Becraft et al., 2015; Bendall et al., 2016; Cohan, 2019, 2002). However, several studies have shown that horizontal gene flow can maintain genotypic and ecological cohesion in

bacterial populations through gene-specific sweeps (Cadillo-Quiroz et al., 2012; Fraser et al., 2009; Hanage et al., 2005; Shapiro et al., 2012; Shapiro and Polz, 2014). High levels of horizontal gene flow, particularly homologous recombination, allow gene-specific sweeps in which adaptive genes are preferentially shared with bacteria linked through horizontal gene transfer (HGT). A variation in gene flow (i.e., gene flow barriers) can limit the spread of adaptive genes and lead to the formation of sequence-discrete populations (Arevalo et al., 2019a; Cordero and Polz, 2014; Fraser et al., 2007; Hanage et al., 2005; Shapiro and Polz, 2015). Therefore, the competing ecological population theory suggests that ecologically distinct and sequence-discrete populations should only be delineated based on discontinuities in gene flow, termed ecological populations (either vertical or horizontal) (Cordero and Polz, 2014; Dykhuizen and Green, 1991; Fraser et al., 2009; Shapiro and Polz, 2014). Recent studies have shown evidence for gene-specific and genome-wide sweeps occurring in many bacterial communities (Arevalo et al., 2019a; Bendall et al., 2016; Chase et al., 2019; Olm et al., 2020). Their prevalence across all bacterial taxa and nature is still in question.

Unfortunately, a survey for the dominance of gene-specific or genome-wide sweeps in diverse natural bacterial communities would be an enormous challenge. In situ studies of ecology and evolution require large-scale sampling, spatially and temporally, to account for the complexity of the natural environment (Forde et al., 2008; Jessup et al., 2004). This is further complicated by the previous difficulties in observing microscopic bacteria's phenotypic characteristics and ecological niches (Achtman and Wagner, 2008; Palys et al., 1997; Shapiro et al., 2009). As such, model systems are used to create a simplified but generalisable understanding of biological systems and processes (Jessup et al., 2004). In studying natural environments, biological edge cases are ideal model systems as they operate in extreme parameters that test the limits of living systems and provide insights into uncommonly observed evolutionary or ecological processes (Pérez-Gracia et al., 2010). Therefore, elucidating the mechanisms of ecological differentiation requires an edge case model that would be able to demonstrate the importance of horizontal gene flow in microbial speciation, such as chemolithoautotrophic bacteria. Ecotype theory suggests that in ecologically homogeneous populations such as chemolithoautotrophs horizontal gene flow and gene level sweeps are unfavourable (Cohan, 2016). The narrow extreme conditions, coupled with lack of metabolic inflexibility, is argued to hamper horizontal gene flow as it could lead to recombination of disadvantageous or maladaptive genes in very entrenched genes such as the energy production pathway (Aris-Brosou, 2005; Jain et al., 1999). In addition to metabolic limitations, a model to test ecological population theory must also thrive in discrete habitats with multiple environmental factors undergoing a high degree of variability since these habitats favour microbial differentiation and the formation of ecologically distinct populations (Kassen and Rainey, 2004; Rainey and Travisano, 1998; Weltzer and Miller, 2013).

Earliest observations on the role of horizontal gene flow in population formation were seen in the facultatively chemoheterotrophic sulphur-oxidizing archaean isolates of *Sulfobolus islandicus* from hot springs, showing that purging of intrapopulation diversity after selection events was prevented by high homologous recombination (Whitaker et al., 2005). Since then, there have been many studies using autotrophic (e.g., *Chloroflexi*, *Prochlorococcus*), pathogenic (i.e. *Salmonella*), and free-living bacteria (e.g., *Streptomyces, Vibrio, Listeria* and etc.) forming ecological populations through horizontal gene flow (Arevalo et al., 2019a; Bendall et al., 2016; Bhatia et al., 2023; Chase et al., 2019; Chen et al., 2022; den Bakker et al., 2008; Hunt et al., 2008; Szabo et al., 2013; Wang et al., 2022; Ward et al., 2018). However, unlike most bacteria with multiple metabolic options, studies on the effect of horizontal gene flow in bacteria with limited metabolic redundancy are scarce. Highly specialised bacteria lacking the ability to switch resources or niches were supposed to have a lower recombination rate (Vos and Didelot, 2009a) due to recombination's higher chance of a deleterious effect (Cohan, 2016).

Here, I present chemolithoautotrophic *Aquificota* bacteria from geothermal hot springs in the Taupo Volcanic Zone as an ideal edge case for examining the mechanisms of microbial differentiation. Natural bacterial communities in geothermal hot springs have often been used to examine microbial diversification and dispersal theories (Becraft et al., 2020; He et al., 2023; Podar et al., 2020; Ward and Cohan, 2005). The extreme range and discrete nature of the geochemical and geophysical conditions (i.e. temperature and pH) of these springs limit the microbial diversity that can thrive (Louca, 2021; Oliverio et al., 2018; Sharp et al., 2014; Ward et al., 2017), and fluctuation in these conditions often result in shifts in microbial community structures (Power et al., 2023, p. 202; Wang et al., 2014; Ward et al., 2017) — ideal for studying microbial differentiation. Additionally, microbial communities in geothermal pools have higher speciation and mutation rates when stressed due to faster error-prone DNA repair mechanisms (Conrad et al., 2011; He et al., 2023; Li et al., 2014). One example of a highly specialized organism is found in these geothermal communities. Members of phylum *Aquificota*, a group of chemolithoautotrophic bacteria, are found to be dispersed and have a relatively high abundance across the geographically segregated geothermal springs in Taupo Volcanic Zone (Power et al., 2018). *Aquificota* bacteria primarily rely on hydrogen and sulphur oxidation for energy and obtain carbon through CO_2 fixation (Gupta and Lali, 2013; Reysenbach, 2015). *Aquificota* genomes are very compact, with some genomes observed to have minimal co-regulation of biosynthetic genes, implying their dependency on highly specialized chemolithoautotrophic processes (Swanson, 2001). The deleterious risk of recombination is further amplified when considering the reliance of many *Aquificota* on few non-redundant genes for hydrogen oxidation (Brugna-Guiral et al., 2003; Giovannelli et al., 2017; Greening et al., 2016).

To examine the role of gene-specific and genome-wide sweeps in natural bacterial communities, in this study, I investigated the population structure and prevalence for recent horizontal gene flow in closely related, minimally-enriched *Aquificota* bacteria isolated from Taupo Volcanic Zone hot springs. The degree of recent horizontal gene flow was measured using the PopCOGenT pipeline by measuring the fraction of low-diversity loci between genomes (Arevalo et al., 2019a). Using this method, reliable observation of natural bacterial diversity requires accurately capturing the variations in the closely related genomes, as homologous recombination is more prevalent with higher sequence similarity (Fraser et al., 2007; Majewski, 2001). Traditional isolation techniques often underrepresent natural bacterial diversity due to the domestication of bacteria during the enrichment process (Liu et al., 2017; Steensels et al., 2019; Vaz-Moreira et al., 2011; Xu et al., 2022). Meanwhile, high-throughput techniques such as metagenome-assembled genomes (MAGs) often form chimeric genomes brought by the batch effect (Orakov et al., 2021; Singleton et al., 2021), which hinders differentiation of the natural intrapopulation bacterial diversity. Even though single-cell amplified genomes (SAGs) can acquire genomes from individual bacteria, acquiring high-quality, non-discontiguous genomes reliably can be done by only a few laboratories due to the need for highly specialized instrumentation (Alneberg et al., 2018; Bowers et al., 2017; Stepanauskas, 2012; Troell et al., 2016). Hence, in this study, I developed a simple high-throughput targeted isolation technique through a direct serial dilution-to-extinction method (Henson et al., 2020; Yang et al., 2016) to acquire minimally-enriched samples of *Aquificota* bacteria that can capture the natural intrapopulation diversity. Using the genomes of these closely related *Aquificota*, this study sought to test the hypothesis that gene-specific sweeps can drive intrapopulation genomic diversity in edge-case bacteria like *Aquificota* and its population boundaries can be delineated through discontinuities in horizontal gene flow.

2.3. Methodology

2.3.1. Bacterial Isolation and Cultivation

2.3.1.1. Sites for water sampling

Geothermal features in Kuirau Park in Rotorua, Bay of Plenty, with a high abundance of different *Aquifictoa* species, were selected for this study (Figure 2.1). Based on the 1000 Springs project (Power et al., 2018), Kuirau Park geothermal feature 101 assigned as KRP1 and Kuirau Park geothermal feature 9 (assigned as KRP2) have an *Aquificota* abundance of 73.3% and 58.4%, respectively (1000springs.org.nz). Geothermal fluids were carefully sampled from the water column of the two features to minimize disturbance and contamination from sediments.

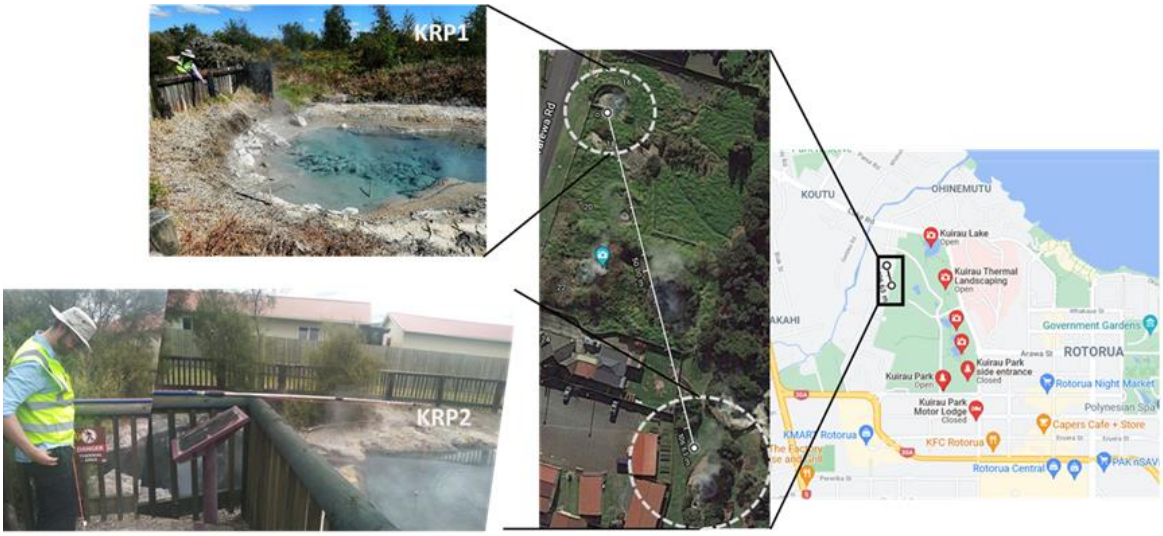

Figure 2.1 A map of the Kuirau Park and the sampling sites. Two springs in Kuirau Park (KRP1 and KRP2) were used for water sampling and isolation of *Aquificota* samples. The two springs are located on the western side of Kuirau Park, Rotorua and are separated 50m apart**.** *Imagery © 2024 Airbus, Bay of Plenty TA's, CNES/Airbus, Maxar Technologies, Waikato District Council, Map data © 2024 Google.*

2.3.2. Serial Dilution-to-Extinction

The wells of 96 deep-well plates were filled with *Aquificota*-targeted media (see SUPP. 1 Table 6.1) (Hetzer et al., 2008) and inoculated with the geothermal fluids within 2 hours of sampling. The initial dilution of the geothermal fluid was done with 1:1 dilution (2mL max. total volume) followed by subsequent serial dilution 1:4 dilution of inoculant to media. Serial dilution was done for eight iterations (Figure 2.2). The plates were then sealed using a gas-permeable membrane (Merck Breathe-Easy® sealing membrane) and placed in an environmental gas chamber (see SUPP. 1 Figure 2.6). The air mixture within the gas chamber was purged using N_2 gas and replaced with an 80:5:15 H_2:CO_2:air gas mixture for *Aquificota* cultivation (Hedlund et al., 2015; Hetzer et al., 2008; Reysenbach et al., 2001). Similar five-fold dilutions were done using test tubes with a total inoculum and media volume of 10 mL for each tube. The tubes were also purged and aerated using the prior gas mix. The tubes and the chamber with the plates were then incubated at 70°C for two weeks.

After the first passage, the growth of each well and tube was checked using a resazurin cell viability assay (Kim and Jang, 2017; Sandberg et al., 2009). Tubes and wells with bacterial growth changed from blue to pinkish hue (see SUPP. 1 Figure 2.6C). The last dilution that changed color was then used to inoculate the second passage. The samples were subjected to five five-fold serial dilutions and incubated using the same parameters as the initial passage.

After incubation and cell viability assay, the last dilution with growth was inoculated into 10 mL media in a test tube to increase their biomass for DNA extraction.

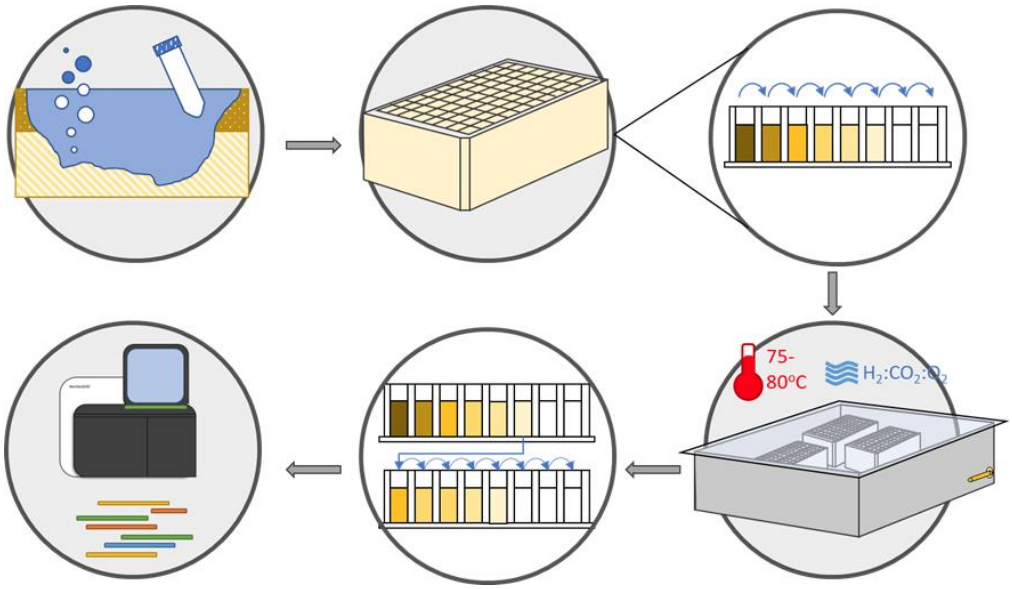

Figure 2.2 The serial dilution-to-extinction isolation and sequencing protocol for 96 deep-well plates. Geothermal fluid samples were directly inoculated into deep well plates or tubes with minimal enrichment. Serial dilution was performed for up to a max of 2*57 (156,250) dilution factor. The plates were put inside a modified gas chamber. The chamber and the tubes were degassed and refilled with an 80:5:15% ratio of H2:CO2:Air gas mixture and were then incubated for up to two weeks at 70°C. Growth was checked using a resazurin cell viability assay. The last well with growth is used for the subsequent passage and serial dilution-to-extinction. The final dilution with growth in the last passage was used for genome sequencing.

2.3.3. DNA Extraction and Genome Sequencing

Ten millilitres of bacterial cell suspension from the tube for increased biomass yield was pelleted at 4000 xg for 10 min and washed with PBS buffer. The biological materials were transferred into 1.5mL microcentrifuge tubes, and DNA extraction was performed using a MOBIO PowerFood™ extraction kit following the manufacturer's instructions. RNAse A treatment was also done to minimise RNA content during sequencing. Afterwards, the extracted DNA was cleaned and concentrated using the Zymo DNA Clean-up and Concentrator kit.

The isolates were validated as members of *Aquificota* using *16S rRNA* gene amplification and sequencing. The V4 region (515F and 806R region) was amplified using Earth Microbiome Project *16S rRNA* gene primers (Caporaso et al., 2012; Thompson et al., 2017) and was sequenced through capillary sanger sequencing. After quality control, the clean reads were identified as the closest type strain using BLAST against Genbank's *16S rRNA* gene RefSeq database (Schoch et al., 2020). Thirty confirmed *Aquificota* isolates were used for whole-genome sequencing in

Illumina™ HiSeq 2100 with a 2x150bp library size, and base calling was done using sequencer built-in software RTA.

Adapters were removed, and low-quality reads were filtered using Trimmomatic (ver. 0.40) (Bolger et al., 2014). The cleaned reads from each sample were then used for individual de novo assembly of whole genomes using SPAdes (Bankevich et al., 2012). The axenicity of the genomes was checked using simple quality control of assembled contigs. The average coverage (k-mer coverage) values reported by SPAdes were plotted against the length of the contigs. Contigs from the same organisms should appear in the plot to have a similar coverage distribution indifferent to the contig length. Contaminant contigs would appear to have a much different coverage (Douglass et al., 2019). The contaminant contigs found are removed from the final genome assembly. If the plot shows that contigs were distributed into two or three distinct clusters, the contigs are binned, and each bin is used as an individual genome for downstream analysis. The quality of the assembled genomes was then checked using checkM (Parks et al., 2015). Genomes with >2% contamination were not used for analysis.

2.3.4. Species delineation and sequence-discrete cluster analysis

2.3.4.1. Taxonomy Placement and Phylogenetic Tree Analysis

Confirmation of the identity of sequenced genomes and their placement among the *Aquificales* taxa was done using sequence analysis of the *16S rRNA* gene. The *16S rRNA* gene sequences were identified from the genome using the RNAMMER ver. 1.2 (Lagesen et al., 2007) and Barrnap ver. 0.9 (Seemann, 2018). The identified *16S rRNA* gene sequences from the genomes and *16S rRNA* gene sequences of closely related type-strains to the genomes from Genbank's RefSeq Ribosomal RNA database were used to identify the phylogenetic position of the genomes. The sequences were aligned using ClustalW (Thompson et al., 1994) and used to build an unrooted phylogenetic tree using maximum likelihood in MEGA7 (Kumar et al., 2016). Based on its lowest BIC score, phylogenetic tree reconstruction was done using Tamura 3-parameter with Gamma distribution as a substitution model, and was validated using the bootstrap method of 1000 trees (Tamura, 1992).

A parallel taxonomic placement of the genomes was also done using the Genome Taxonomic Database - Toolkit v1.5.0 (GTDB-TK) pipeline (Chaumeil et al., 2019; Parks et al., 2022). Using the GTDB-TK classify workflow, the placement of the genomes to the closest genomes within the GTDB database (release 202) was identified using the standard 120 bacterial gene markers. A rooted tree against the closest type strain in the database was also constructed using the concatenated amino acid sequences of the 120 bacterial gene markers in each sample.

The tree was constructed using a maximum likelihood method with the Le Gascuel with Gamma distribution model, based on the lowest BIC score during phylogenic tree modelling, and tested using a bootstrap method for 1000 trees (Le et al., 2012).

2.3.4.2. Species delineation using Average Nucleotide Identity

Genome comparison of the similarities of the samples was done using Average Nucleotide Identity (ANI) analysis. ANI pairwise comparisons between samples were plotted using the FastANI ver. 1.33 (Jain et al., 2018). The clustering of the genomes using ANI values was done using the pheatmap v1.0.12 package in R v4.2.1 (Kolde, 2018; R Core Team, 2022). Genomes with ANI values of more than >97% are classified as belonging to the same species (Stackebrandt and Goebel, 1994). The heat map colour gradient was set at 95% between white and blue and at 99% between orange and yellow to further see the differences between closely related genomes.

2.3.5. Discriminating horizontal gene flow and clonal expansion between genomes

The degree of recent gene flow between samples was analysed using the PopCOGenT pipeline (Arevalo et al., 2019a). The PopCOGenT pipeline utilizes a pairwise comparison of entire genomes to identify shared genomic regions in both the core and flexible genomes. The pipeline identifies this region by measuring the distribution of identical shared regions, termed locally collinear blocks (LCBs), between the paired genomes. LCBs are stretches of identical DNA sequences that have not recently experienced rearrangements and are absent of any SNPs that arise randomly throughout the genome due to mutation, vertical descent, or recent horizontal transfer. Genomes of recombining bacteria are observed to have a higher fraction of long LCBs than non-recombining bacteria. PopCOGenT measures the fraction and length of all LCBs between the paired genomes as the length bias. The length bias is compared to a genome size-adjusted null model to differentiate if the results are the effect of recombination or purifying selection. The length biases between genomes are used to construct population cluster networks. In this network, each unique genome is a node, and the weight of measured length bias is an edge between paired genomes. Paired genomes with a divergence of <0.0355 are considered highly clonal and collapsed as a single node in the network. Infomap, a flow-based method in network clustering, is used to discriminate sequence-discrete populations within the network (Edler et al., 2021).

2.4. Results

2.4.1. High Average Nucleotide Identity heterogeneity can be observed between closely-related Kuirau Park isolates

To investigate whether *Aquificota* populations in natural bacterial communities exhibit intra-species genomic variation, high-throughput targeted isolation of *Aquificota* was done using Kuirau Park 1 and 2 samples. The isolates were grown using the serial dilution-to-extinction method in both tubes and deep-well plates. From the 40 serial dilution setups, a total of 34 isolates were acquired from Kuirau Park sites 1 and 2 (19 and 15 isolates, respectively) after two rounds of passages (six weeks incubation) using the direct serial dilution-to-extinction technique in both tubes and deep-well plates. Cultures with less resolved or low-quality *16S rRNA gene* sanger sequences were filtered out due to potential contamination. Thirty isolates with high-quality reads were sent for sequencing. Out of the 30 genomes, one was unsuccessfully sequenced due to poor library construction, one returned as highly contaminated, and 28 cleaned genomes were acquired after separating contaminating contigs identified using a coverage-versus-length plot (Douglass et al., 2019). CheckM analysis shows a 99.2% ± 0.322 average completeness and 1.03% ± 0.185 contamination (see SUPP. 2 Table 6.2). The de-clustered genomes were then used for downstream comparative genomics and phylogenetic analysis.

Next, genomic variations between sequenced genomes were analysed using pairwise computation of the Average Nucleotide Identity (ANI). Heatmap clustering of the whole genome sequences showed that the genomes belong to three distinct populations (Figure 2.3). High intra-population homogeneity was observed between clustered genomes with mean intrapopulation ANI values of 99.5±0.21%, 99.8±0.182%, and 99.97% for Populations 1, 2, and 3, respectively. Meanwhile, interpopulation ANI showed a large distance between Population 3 and the other populations, with mean interpopulation ANI at 77.4±0.09%. In comparison, interpopulation ANI between Population 1 and 2 is only at 92.9±0.201%. Based on recent proposals for demarcating species at ANI >95% (Goris et al., 2007; Jain et al., 2018; Rodriguez-R and Konstantinidis, 2014), these three populations are distinct bacterial “species” (Figure 2.3). Unlike the species gap, no clear cut-offs can be used to define genus boundaries based on ANI values. However, members of the same genera have been found to have ANI similarity between 68-72% and even as high as 72-82% (Qin et al., 2014). As such, the three Aquificota populations could potentially belong to the same genus.

Even with the high homogeneity in these populations, genomes in Population 1 and 2 are not completely clonal and exhibit some heterogeneity, with ANI median values of 99.451% and 99.857% for Population 1 and 2, respectively (see SUPP. 3 Figure 6.2). Most pairwise comparisons in Population 1 have shown to fall between 99.353-99.536% ANI (Figure 2.3 - Light Orange), while Population 2 has ANI values mostly >99.769% (Figure 2.3 - Dark Orange & see SUPP. 3 Figure 6.2). Several clusters of genomes or subpopulations exhibited high ANI similarity between genomes (>99.9%), such as those of the G05, G06, and G07 or the G10, G26, and G27 groupings, which could be indicative that the isolates belong to the same clonal lineage. However, the observed heterogeneity between these subclusters can also be explained by a high mutation rate, weak purifying selection, gene-specific sweeps between the bacteria, or a

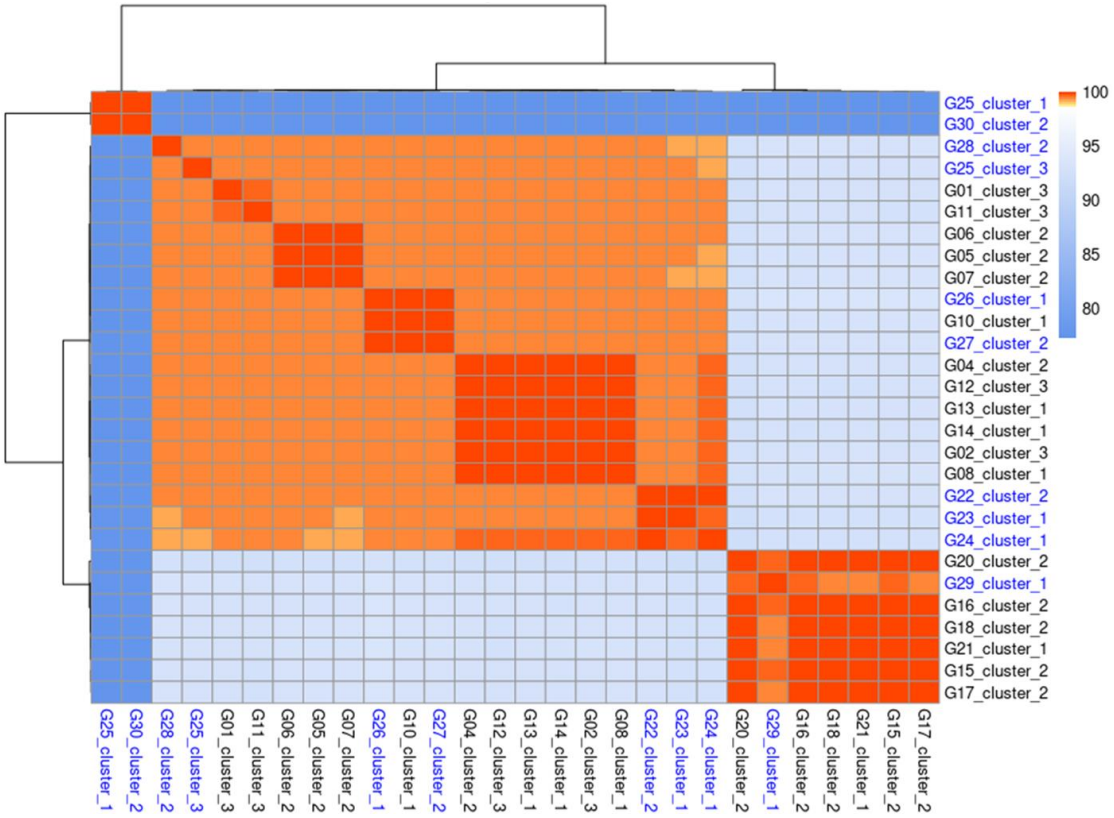

Figure 2.3. Heatmap of the pairwise ANI of the Kuirau Park genomes. The computed ANI grouped all the Kuirau Park genomes into three distinct populations with a mean intrapopulation ANI of 99.54±0.208%. Populations 1 and 2 exhibit high interpopulation ANI similarity at 92.9±0.201%, while Population 3 is distantly related to the other populations at 77.4±0.09% ANI. With ANI <95%, the three populations can be considered as distinct bacterial "species". High intrapopulation heterogeneity can be observed as most genomes have ANI similarity of 99.2-99.8% (light orange), while other members of the same cluster are extremely similar with ANI of >99.98% (dark orange) (see SUPP. 3 Figure 6.2). The heterogeneity of the isolates is not affected by whether the isolates were grown in tubes (blue labels) or with deep-well plates (black labels) (see SUPP. 4 Figure 6.3). For better visualisation, gradient in ANI colour was separated at 99% ANI with higher values coloured as yellow to dark orange while lower values as white to blue.

combination of these factors. Therefore, sequence similarity analysis alone cannot resolve the differences between these factors.

Given the heterogeneity observed in *Aquificota* genomes, I next validate that there is no bias between the sampling method used to isolate *Aquificota* bacteria that results in such heterogeneity. Both tube and deep-well plate methods produced genomes from Populations 1 and 2, while Population 3 was only isolated using the tube-based isolation method (Figure 2.3). The genomes acquired from deep-well plates and tube-grown *Aquificota* bacteria exhibit similar heterogeneity (see SUPP. 4 Figure 6.3), indicating that the differing volume used in the serial dilution-to-extinction method captures a similar level of diversity from the natural bacterial communities. Regardless of the isolation method, genomes from both methods can also form highly homogenous subpopulations, as shown by the high similarity between G10, G26 and G27 genomes (Figure 2.3). These results show that the serial dilution-to–extinction technique can capture intrapopulation and/or intra-species genomic heterogeneity while minimising enrichment in a high throughput manner.

2.4.2. Kuirau Park isolates were identified to be *Aquificota* members of the undescribed genus *UBA11096* sp.

To confirm that three isolated populations are distinct bacterial 'species' from the same genus, I identified their phylogenetic position using 16S rRNA gene sequence tree from microbial type strains from NCBI 16S RNA gene database and phylogenomic analysis using the markers and references from Genome Taxonomic Database (GTDB). The BLAST search of the extracted 16S rRNA gene against the NCBI 16S RNA gene database shows that 26 isolates from populations 1 and 2 are most closely related to *Hydrogenobacter subterraneus* HGP1 type-strain (99.046±0.133%). Meanwhile, Population 3's G30_cluster_2 is found closest to *Hydrogenobacter hydrogeniphilus* DSM2913 at 96.69%. Unfortunately, no 16S rRNA sequence was extracted from Population 3's G25_cluster_1 genome. The population's <97% 16S rRNA gene sequence similarity to the closest type-strain suggests it is potentially a novel undescribed species (Stackebrandt and Goebel, 1994).

The results of BLAST aligned with what was observed in the pplacer taxonomic analysis using the GTDB toolkit. Genomic comparison of the samples against the GTDB-classified Populations 1 and 2 as members of the unclassified genus UBA11096, with each population separated into two distinct species. In contrast, Population 3 is classified within the UBA11096 genus without any known closest species representative. Population 3's G30_cluster_2 and G25_cluster_1 match with the closest GTDB representative species *UBA11096* sp011006175, only

at 79.29% ± 0.0038 ANI similarity, which supports the rRNA gene sequence analysis that implies it is a putative novel bacterial "species". Meanwhile, Population 1 matched with undescribed species *UBA11096* sp011054805 (98.80% ± 0.0092 ANI similarity), and Population 2 matched into the separate undescribed *UBA11096* sp003534055 (99.06% ± 0.0103 ANI similarity). The difference in the naming convention between the two analyses is due to the lack of available genome associated with the described *Hydrogenobacter subterraneus* (Schoch et al., 2020; Takai et al., 2001), but analysis of the partial *16S rRNA* gene sequences of UBA11096 species also shows they are closely related to *H. subterraneus* (at 99.93%). Given the relationship between UBA11096 and *Hydrogenobacter subterraneus*, I will use the name of the UBA11096 species for the rest of this paper for consistency.

The phylogenetic tree construction using the *16S rRNA* gene sequences and genomic data (Figure 2.4) agrees that the Kuirau Park isolates belong to a branch of the *Aquificales* family closely related to *Hydrogenobacter*. The phylogenetic placement of the samples shows that the samples are clearly divided into three distinct populations (Figure 2.4B), corresponding to the three clusters found in ANI comparisons. Overall, phylogenomic analysis shows a higher resolution that group Population 2 into a distinct taxonomic clade than Population 1 isolates. Interestingly, aside from Population 2, the other taxonomic clades of isolated Aquificota bacteria can be found in Kuirau Park 1 and 2, indicating that the isolates are not geographically limited by the spring. The lack of a Population 2 representative in Kuirau Park 1 does not indicate its total absence but can be a result of insufficient sampling coverage. Notably, *UBA11096* sp. representative genomes and the three populations are taxonomically distant from other representative species of the genus *Hydrogenobacter*. This distance is significantly highlighted using the phylogenomic tree in which the samples and UBA11096 representative species' distance to the other representative species of *Hydrogenobacter* is comparable to the distance of *Hydrogenobacter* to *Thermocrinis* taxonomic clade. The distance between *UBA11096* taxonomic group supports the classification of UBA11096 as a distinct genus in GTDB and could indicate that the isolates are a putatively distinct novel genus, which could also include *Hydrogenobacter subterraneus* and other previously identified *Hydrogenobacter* sp. strains closely associated with *Hydrogenobacter subterraneus* (Eder and Huber, 2002; Hedlund et al., 2015).

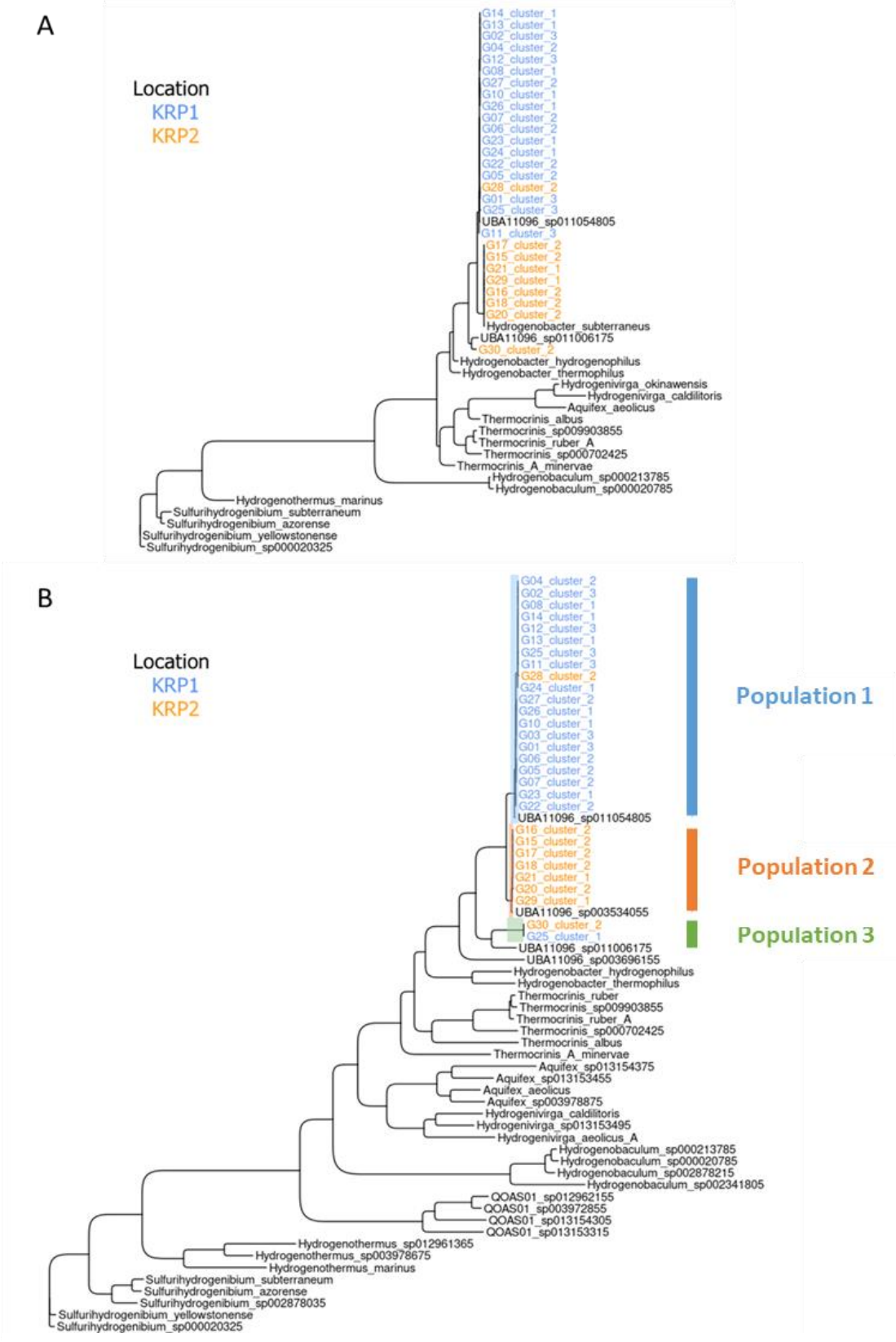

Figure 2.4. The phylogenetic position of Kuirau Park isolates in the Order *Aquificaceae*. (A) The phylogenetic position of the Kuirau Park isolates using 16S rRNA sequences and (B) 120 gene markers from the GTDB-TK pipeline shows that all isolates are clustered closely with the genus *Hydrogenobacter* and the unassigned taxa UBA11096 from GTDB. In the 16S tree (A), populations 1 and 2 cluster near *Hydrogenobacter subterraneus* and an undescribed UBA11096 species. The two populations appears indistinguisable suggesting the close relatedness between the and appears distinctly separate from sample (G30) . Phylogenetic analysis using the GTDB derived markers agrees that G30 (with the G25 isolates) cluster farther from the rest of the isolates which can be distinctively differentiated into two clades - *Clade 1* and *2* (B).

2.4.3. Variation in homologous recombination delineates the *Aquificota* populations in Kuirau Park

To differentiate the effect of horizontal gene flow from those of mutation and purifying selection on both inter- and intrapopulation heterogeneity, I examine the recent horizontal gene flow between the genomes using the PopCOGenT pipeline (Arevalo et al., 2019a; VanInsberghe et al., 2020). The pipeline offers a higher understanding of intrapopulation heterogeneity by measuring the genome fraction containing low diversity loci between two genomes, a hallmark of recent recombination. It then performs a cluster analysis to identify genomes that are connected through recent or a high degree of horizontal gene flow. PopCOGenT analysis of the *Aquificota* genomes showed that the isolated bacteria comprised two gene-flow clusters with three distinct sub-clusters linked through varying degrees of horizontal gene flow (Figure 2.5). The identified sub-clusters correspond to the three populations previously observed in ANI and the phylogenetic analysis. Cluster 1 comprises a single population previously identified as Population 3 in ANI and phylogenomic analysis, which appears as a single node composed of two genomes in a clonal complex (G25_cluster_1 and G30_cluster_2). Meanwhile, cluster 2 consists of two distinct subclusters, corresponding to previously labelled Populations 1 and 2. Cluster 2 comprises multiple nodes linked with varying degrees of horizontal gene flow. The degree of horizontal gene flow varies widely from very strong, as indicated by the connection between the nodes G26 and G27 with G10, to weak gene flow that divides cluster2 into the two distinct sub-clusters (Figure 2.5 – Blue and Orange Nodes). It is worth highlighting that there is no recently observed horizontal gene flow between Cluster 1 and 2. This result supports the observed distant relation of Population 3 with the rest of the isolates in phylogenetic analysis and ANI. Additionally, the congruence of the PopCOGenT clusters with those of the ANI and phylogenomics and the clear demarcations of the strength of horizontal gene flow further provide credence to the idea that the sequence-discrete populations are genetically and ecologically distinct populations – an ecological population (Arevalo et al., 2019b; Shapiro and Polz, 2014).

The intrapopulation heterogeneity, previously observed in ANI, is correlated with both high horizontal gene flow or clonal complexes within the ecological population. For example, the high homogeneity of G10, G27, and G26 genomes in ANI (Figure 2.3) is driven by the observed strong horizontal gene flow between them (Figure 2.5). At the same time, the homogenous group of G02, G04, G08, G12, G13 and G14 appears as a single clonal complex in PopCOGenT. This result suggests that high genomic similarity between genomes can be driven by horizontal gene flow and clonal expansion. In contrast, a comparison of heterogeneity between clonal nodes within Population 1 or 2 exhibits higher intra-population heterogeneity, implying that observed intra-population heterogeneity in an *Aquificota* population could be driven by varying degrees

of horizontal gene flow. Additionally, these variations in the formation of ecological population heterogeneity in *Aquificota* isolates clearly illustrate that bacterial populations can be composed of both clonal lineages and/or clusters of isolates linked through horizontal gene flow. The co-occurrence of both clonal lineages and horizontally linked clusters in edge-case populations like *Aquificota* implies that bacterial populations in natural communities are not solely driven by either genome-wide or gene-specific sweeps. Instead, genome-wide and gene-specific sweep interplay drives the formation of ecologically distinct and sequence-discrete populations.

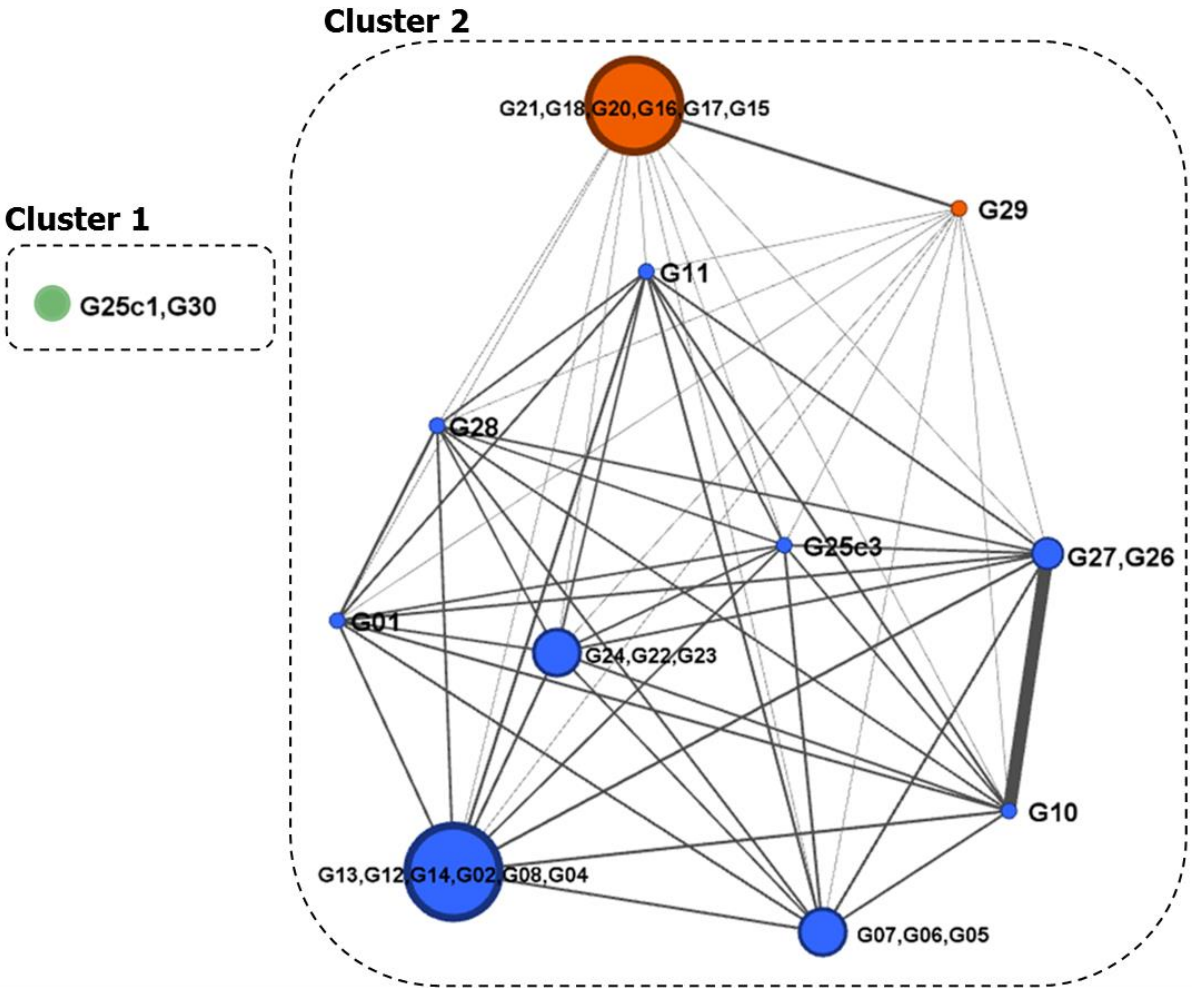

Figure 2.5. The Kuirau Park isolates are linked by gene flow networks. PopCOGenT clusters the Kuirau Park isolates into two distinct clusters (Cluster 1 and Cluster 2). Cluster 2 is divided into two distinct subclusters (blue and orange) composed of multiple nodes linked through varying degrees of horizontal gene flow with observed lower degree of horizontal gene flow between populations than within each recombining population 1 (blue) and 2 (orange) (see SUPP. 5 Figure 6.4A). The size of each node corresponds to the number of isolates within each node. The thickness of each edge corresponds to the degree of horizontal gene flow between nodes

2.4.4. *Aquificota* bacteria from Canada and New Zealand are linked through a homologous recombination network

The formation of sequence-discrete clusters through gene-specific sweeps is not limited to the *Aquificota* bacteria isolated from Kuirau Park. A recent study on *Aquificota* SAGs from a similar site in Canada shows that closely related members of *Hydrogenobacter* sp. from Canadian hot springs have a highly panmictic genetic structure (Bowers et al., 2022). The *Hydrogenobacter* sp.-related isolates in Canada exhibit an elevated nucleotide diversity and a very rapid decay in SNP linkage disequilibrium of SAGs, which is indicative of recent recombination in those populations. Given the similarity of the Canadian *Hydrogenobacter sp.* isolates to Kuirau Park isolates in taxonomic classification (data not shown) and their recombining nature, a question arises on how similar the roles of genome-wide and gene-specific sweeps in the formation of diversity in the Canadian isolates. Additionally, as allopatry plays a role in population dynamics in higher-order metazoans (Amor et al., 2014; Breusing et al., 2020; Cabej, 2012; Zink, 2012), it is crucial to investigate the similarity with microbial populations by observing the role of geographical distance/barrier on the differentiation of natural microbial populations. Therefore, I ran ANI and PopCOGenT analysis with both the *Aquificota* genomes from Kuirau Park and Canada.

The ANI analysis of the Canadian and Kuirau Park isolates confirmed that the Canadian isolates are taxonomically related to the Kuirau Park isolates (Figure 2.6A). At a range of 96.13-98.63% (mean 98.14±0.52%) ANI similarity, the Canadian isolates (SAGs with completeness >60%) clustered together with members of Kuirau Park Population 1 and, thus, are also a member of the putative novel genus of UBA11096 (Figure 2.6A). The ANI also showed the high diversity reported in Bower's data, with the Canadian isolates having ANI values ranging from 96.91-99.53% (mean 98.75 ± 0.66%) similarity. In comparison, all Kuirau Park 1 isolates have intrapopulation ANI of >99.23% (mean 99.50 ± 0.18%) similarity. However, this higher observed heterogeneity was more prevalent in Canadian isolates with lower genome completeness (60-75% completeness) (see SUPP. 5 Figure 6.4B).

Even with the higher diversity in ANI, analysis of recent recombination shows the Canadian isolates form a single cluster of singleton nodes, composed of one genome, linked by horizontal gene flow (Figure 2.6B). Supporting the results of ANI, the Canadian isolates (SAG completeness >80%) are linked through horizontal gene flow with the Kuirau Park isolates' Population 1 forming a distinct subcluster with a length bias (339.16 ± 56.61) greater than between the two populations of Kuirau Park Cluster 1 (33.81 ± 5.34). The length bias within the Canadian isolates (901.44 ± 293.27) and within Kuirau Park Population 1 (1880.46 ± 1693.73) are higher than those observed between them (see SUPP. 5 Figure 6.5), which implies that recent recombination between the two sites is rarer and supports the effect of geographical distance as

a hindrance to recombination (Greenlon et al., 2019). The analysis of recent recombination using SAGs with varying genome completeness shows that observed length bias is affected by genome completeness with SAGs >80% completeness (901.44 ± 293.27) is higher than those with lower completeness, SAG with 50-80% completeness (613.52 ± 106.74) and SAG with <50% completeness (454.97 ± 171.11) (see SUPP. 5 Figure 6.4B). This discrepancy in observed length bias and genome completeness, coupled with high ANI diversity observed in SAGs with lower genome completeness, highlights the need for good quality and highly complete genomes in studying recent recombination in sequence-discrete populations and the formation of intrapopulation diversity (Van Rossum et al., 2020). In addition, the minimal difference in genomic diversity using varying volumes of inocula in the serial dilution-to-extinction method (see SUPP. 3 Figure 6.3) and the high bar of entry in acquiring high-quality SAGs – both in cost and need for specialist manpower – suggests the advantage of using minimal enrichment methods as a cost-effective alternative method in studying intrapopulation genomic diversity and the mechanisms of horizontal gene flow in natural bacterial populations.

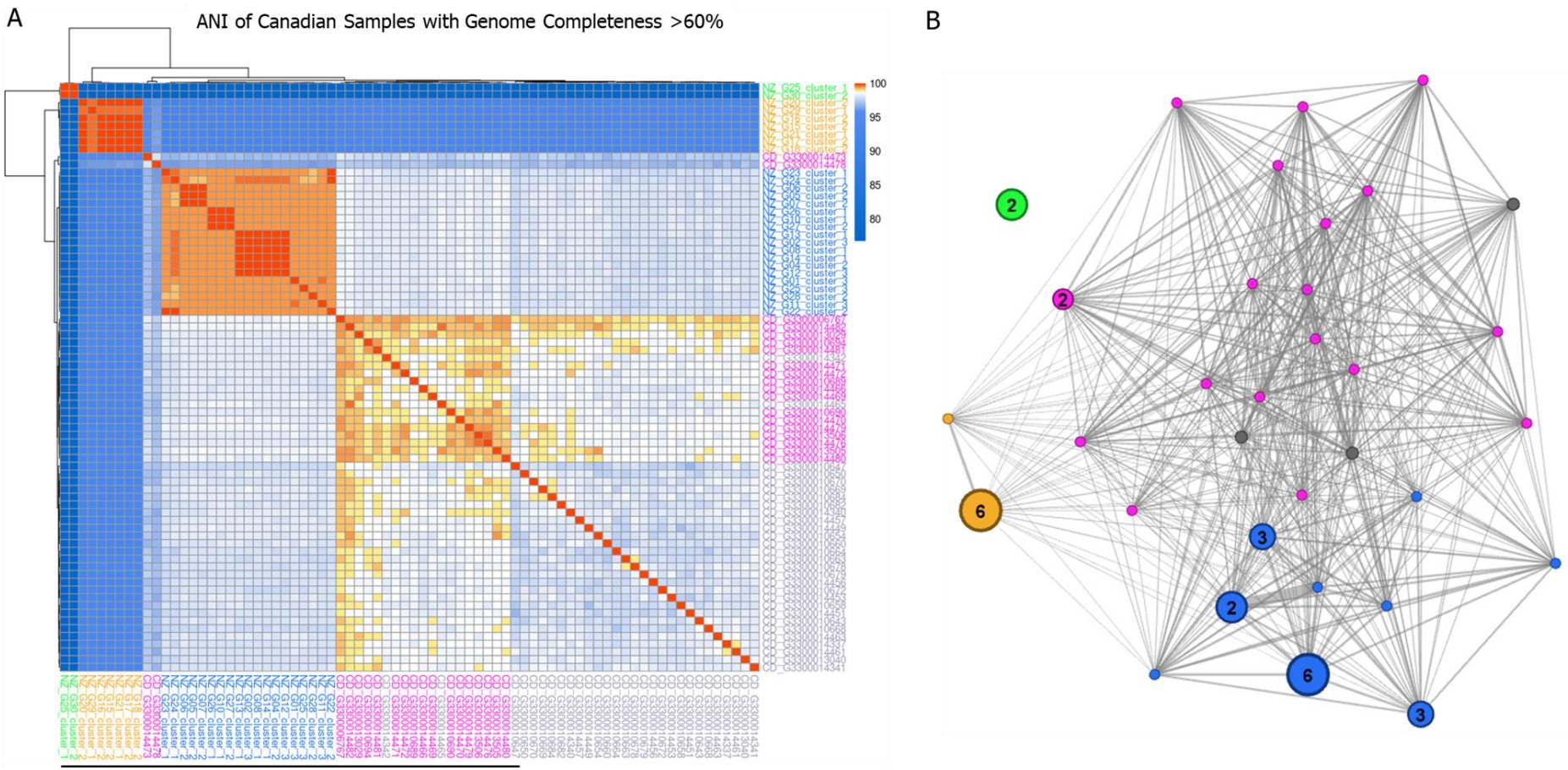

Figure 2.6. Horizontal gene flow drives intrapopulation diversity in Canadian Aquificota . (A) ANI of the Canadian SAGs with >50% and >80% genome completeness (grey and purple texts, respectively) clustered together with genomes isolated from Kuirau Park Population 1 (blue) and are distantly related with Population 2 (orange text) and Population 3 (green text) of Kuirau Park isolates. ANI homogeneity is observed to be higher in Kuirau Park isolates than the Canadian isolates (see SUPP. 5 Figure 6.5), with SAGs that have higher completeness(>80%) showing higher ANI similarity between isolates. (B) Similarly, PopCOGenT analysis of selected genomes (A - Black Bar) shows the Canadian *Aquificota* isolates (Purple and Grey nodes) to belong to one cluster that is horizontally linked with Kuirau Park Population 1 and 2 indicating recent recombination (see SUPP. 5 Figure 6.5). Node size in the PopCOGenT network graphs (B) indicates the size of clonal complexes. The edge thickness indicates the strength of horizontal gene flow between nodes. Cut-off ANI distance is set at 99%, with higher values colored yellow to dark orange.

2.5. Discussion

Traditional delineation of microbial populations often relies on comparing the phenotypic and genotypic similarities between organisms without regard to natural boundaries (Brenner et al., 2005; Raina et al., 2019; Stackebrandt et al., 2002; Vandamme et al., 1996; Wayne et al., 1987). With diversification and speciation in bacteria previously relied on the assumption that genome sweeps generally applied to all bacteria (Cohan, 2002; Majewski, 2001; Vogan and Higgs, 2011), genomes are assumed to diverge and speciate from closely related organisms from a common ancestor which radiated through the vertical descent of varying adaptive mutants after different selection events (Cohan and Koeppel, 2008; Sakoparnig et al., 2021). As such, taxonomic distance is primarily assumed to equate to actual evolutionary distance and, thus, can be measured using proteomic and genomic sequence similarity (Zuckerkandl and Pauling, 1965), such as those featuring DNA-DNA hybridisation (Johnson, 1980; Rosselló-Mora, 2006), genomic sequence identity (Jain et al., 2018; Olm et al., 2020; Van Rossum et al., 2020), and/or barcoding markers (Mende et al., 2013; Rotimi et al., 2018). The dominance of vertical gene flow also implies that the increased intrapopulation genomic diversity is a transient phenomenon not yet purged between selection events (Cohan, 2016). However, our results show that the propagation of bacterial diversity and formation of microbial populations does not solely rely on the vertical descent and clonal expansion of adaptive mutants. Analysis of recent horizontal gene flow in the natural populations of *Aquificota* shows intra-species genomic diversity (Figure 2.3) that is associated with varying degrees of horizontal gene flow (Figure 2.5). These results contradict the traditional notion that horizontal gene flow is rare in natural microbial populations (Cohan, 2016, 2002; Majewski, 2001) and support the substantial role of gene-specific sweeps in bacterial speciation (Chase et al., 2019; Crits-Christoph et al., 2020; Shapiro et al., 2012).

Varying degrees of horizontal gene flow can affect bacterial populations differently (Fraser et al., 2007; Van Rossum et al., 2020). A strong horizontal gene flow is a homogenising force that allows genes, including those coding for adaptive traits, to be preferably shared between organisms (Shapiro and Polz, 2015). Strong or frequent gene flow between organisms can appear as high genomic similarity even in ANI (Figure 2.3 & Figure 2.5 – G10, G26, G27). Previous studies show that a shift in this horizontal gene flow, through any means of constructing gene flow barriers (i.e. temporal, geographical barriers or changes in biological factors) and with selection, can result in speciation and differentiation of organisms to sequence-discrete and ecologically distinct populations (Fraser et al., 2007; Vogan and Higgs, 2011). Thus, a weaker or less frequent horizontal gene flow observed in natural bacterial populations – such as the two populations of *Aquificota* in Cluster 2 (Figure 2.5) – can be a hallmark of ecological

and genomic differentiation similar to those observed in sympatric speciation in sexual eukaryotes (Shapiro et al., 2012). The weaker or less recent gene flow between the two populations may result in genes and novel mutations being preferentially shared only within each population. The increasing preference of bacteria to recombine within a select group further leads to a downward spiral decrease in the rate of horizontal gene flow as the flanking sequences become more varied between groups or the appearance of physiological differences inhibiting recombination (Majewski, 2001; Shapiro and Polz, 2015).

It should be noted that diversity in *Aquificota* is not solely driven by gene-specific sweep and horizontal gene flow. The higher genomic diversity (Figure 2.3) observed in the clonal complex of G22, G23and G24 (Figure 2.5) compared to other clonal complexes can imply the accumulation of mutation within the lineage, which results in increased genomic diversity. However, determining whether this phenomenon is a transient or a more stable increase in diversity is difficult to ascertain. Interestingly, the lack of observed recent horizontal gene flow between highly clonal genomes in cluster 1 (Figure 2.5) and the other cluster implies that mechanisms of formation of sequence-discrete populations vary even in closely related organisms such as in these *Aquificota* populations. The factors determining whether a genome has a higher propensity to undergo genome-wide or gene-specific sweep remain unknown.

Even with the supposed fitness cost of horizontal gene flow in bacteria with compact genome size and limited metabolic redundancy (Oliveira et al., 2016), the studied *Aquificota* populations show high horizontal connectivity with each other. One likely factor affecting the preference for *Aquificota* bacteria to undergo gene-specific sweep is the intrinsic genomic characteristics of the population that affect their propensity to undergo homologous recombination and/or horizontal gene flow. Factors such as genome size and free-living lifestyle have been observed to affect the abundance of homologous recombination (González-Torres et al., 2019). González-Torres (2019) study also showed a correlation between high competence brought by com regulon, the presence of restriction-modification systems and high CRISPR-Cas gene content with increased homologous recombination. The association of bacterial host defence system with homologous recombination could indicate that other factors regulating non-homologous recombination (i.e. mobile genetic elements) also play a role in homologous recombination. The high amount of mobile genetic elements (MGEs) and CRISPR-Cas spacers in Canadian '*Hydrogenobacter sp.*' SAGs (Bowers et al., 2022) could support that non-homologous recombination mechanism affecting horizontal gene flow in *Aquificota* bacteria (Nazarian et al., 2018; Oliveira et al., 2016). However, it was primarily assumed that homologous recombination drives gene-specific sweeps within natural bacterial populations (Bendall et al., 2016; Fraser et al., 2007; Whitaker et al., 2005), but PopCOGenT *a priori* approach in identifying

recent horizontal gene flow does not discriminate on the source of horizontally swept genes (Arevalo et al., 2019a). Therefore, further analysis of the swept regions between populations is needed to see the extent of the role of MGEs in forming ecologically distinct recombining populations (VanInsberghe et al., 2020).

Clustering Canadian SAGs and genomes from New Zealand into a single recombining cluster indicates a history of horizontal gene flow between the two subpopulations. The lower observed length bias between the two geographically distant subpopulations than within each locality (see SUPP. 5 Figure 6.5) may suggest a less recent horizontal gene flow. This lower rate of horizontal gene flow might be due to infrequent and irregular geothermal spring eruptions followed by dust dispersal between the sites, resulting in sizeable aeolian dispersal of *Aquificota* bacteria across large geographical distances (Smith et al., 2013). A continuous hop between host organisms of mobile genetic elements of even homologous recombination-driven sequences can also lead to an observed decrease in apparent recombination between bacteria (Greenlon et al., 2019). However, host hopping is less likely to occur for metabolically limited and habitat-constrained bacteria like Aquificota as it requires specific habitat to grow. Previously observed long-distance dispersal of eukaryotic and bacterial organisms after a volcanic eruption (Herbold et al., 2014; Van Eaton et al., 2013) suggests that rather than multiple dispersal events, *Aquificota* populations could have been likely deposited to Canada and/or New Zealand springs after a single large eruption from a volcanic or geothermal source.

Nevertheless, the geographical distance and transmission event between the source and destination act as a gene flow barrier—leading to the accumulation of mutation and its preferential distribution with the bacteria within the respective localities. The distance may result in the observed decay in genome length bias after a long time of no new colonisation of new bacteria in each locality. Nonetheless, the historical horizontal linkage and genomic similarity between the Canadian and New Zealand populations remain, which suggests the potential for active recombination as soon as biotic interaction between the two populations returns (Shapiro et al., 2012). This potential for long-distance gene flow also implies that geothermal springs are not geographical islands in terms of microbial ecology as previously thought (Louca, 2021), and the geographical distance does not limit the formation of ecologically distinct and sequence-discrete populations of *Aquificota*.

2.6. Conclusion

The presence of recombining and clonal population structure in sequence-discrete populations of closely related *Aquificota* bacteria challenges the premise of ecotype theory that metabolically restricted bacteria do not form ecologically cohesive populations through horizontal gene flow (gene-specific sweeps). The co-occurrence of clonal and recombining populations within co-habiting populations implies that genome-wide or gene-specific sweeps do not solely drive bacterial speciation in natural bacterial communities. Instead, the interplay between the two populations can lead to the formation of ecologically cohesive populations. Therefore, delineating sequence-discrete populations in natural bacterial communities must consider the mechanisms of speciation from which genetic heterogeneity is derived – as a niche-defining adaptive trait can be conferred through both horizontal and vertical gene flow. With the increasing observation of intra-species diversity driven by horizontal gene flow (Bendall et al., 2016; Cadillo-Quiroz et al., 2012; Chase et al., 2019; Hanage et al., 2005; Konstantinidis et al., 2022; Viver et al., 2021; Wang et al., 2022; Whitaker et al., 2005), it would be ideal and prudent to demarcate ecologically distinct and sequence-discrete population through discontinuities in total gene flow (both vertical and horizontal gene flow). This criterion would allow bacterial populations or species to be treated akin to the eukaryotic species and the application of ecological theories developed for eukaryotic species in bacterial populations.

2.7. Acknowledgement

This research was supported by the Marsden Research Fund, administered by the Royal Society of New Zealand (17-UOW-026). Access to Kuirau Park was granted by the Rotorua City Council.

2.8. References

Achtman, M., Wagner, M., 2008. Microbial diversity and the genetic nature of microbial species. Nature Reviews Microbiology 6, 431–440. https://doi.org/10.1038/nrmicro1872

Alneberg, J., Karlsson, C.M.G., Divne, A.-M., Bergin, C., Homa, F., Lindh, M.V., Hugerth, L.W., Ettema, T.J.G., Bertilsson, S., Andersson, A.F., Pinhassi, J., 2018. Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. Microbiome 6, 173. https://doi.org/10.1186/s40168-018-0550-0

Amor, M.D., Norman, M.D., Cameron, H.E., Strugnell, J.M., 2014. Allopatric Speciation within a Cryptic Species Complex of Australasian Octopuses. PLOS ONE 9, e98982. https://doi.org/10.1371/journal.pone.0098982

Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J., Polz, M.F., 2019a. A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. Cell 178, 820-834.e14. https://doi.org/10.1016/j.cell.2019.06.033

Arevalo, P., VanInsberghe, D., Polz, M.F., 2019b. A Reverse Ecology Framework for Bacteria and Archaea, in: Polz, M.F., Rajora, O.P. (Eds.), Population Genomics: Microorganisms, Population Genomics. Springer International Publishing, Cham, pp. 77–96. https://doi.org/10.1007/13836_2018_46

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 19, 455–477. https://doi.org/10.1089/cmb.2012.0021

Barker, M.J., Wilson, R.A., 2010. COHESION, GENE FLOW, AND THE NATURE OF SPECIES. The Journal of Philosophy 107, 61–79.

Becraft, E.D., Wood, J.M., Cohan, F.M., Ward, D.M., 2020. Biogeography of American Northwest Hot Spring A/B′-Lineage Synechococcus Populations. Frontiers in Microbiology 11.

Becraft, E.D., Wood, J.M., Rusch, D.B., Kühl, M., Jensen, S.I., Bryant, D.A., Roberts, D.W., Frederick M. Cohan, Ward, D.M., 2015. The molecular dimension of microbial species: 1. Ecological distinctions among, and homogeneity within, putative ecotypes of Synechococcus inhabiting the cyanobacterial mat of Mushroom Spring, Yellowstone National Park. Front. Microbiol. 6. https://doi.org/10.3389/fmicb.2015.00590

Bendall, M.L., Stevens, S.L., Chan, L.-K., Malfatti, S., Schwientek, P., Tremblay, J., Schackwitz, W., Martin, J., Pati, A., Bushnell, B., Froula, J., Kang, D., Tringe, S.G., Bertilsson, S., Moran, M.A., Shade, A., Newton, R.J., McMahon, K.D., Malmstrom, R.R., 2016. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. The ISME Journal 10, 1589–1601. https://doi.org/10.1038/ismej.2015.241

Bhatia, R.P., Kirit, H.A., Lewis, C.M., Jr., Sankaranarayanan, K., Bollback, J.P., 2023. Evolutionary barriers to horizontal gene transfer in macrophage-associated Salmonella. Evolution Letters 7, 227–239. https://doi.org/10.1093/evlett/qrad020

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., Tringe, S.G., Ivanova, N.N., Copeland, A., Clum, A., Becraft, E.D., Malmstrom, R.R., Birren, B., Podar, M., Bork, P., Weinstock, G.M., Garrity, G.M., Dodsworth, J.A., Yooseph, S., Sutton, G., Glöckner, F.O., Gilbert, J.A., Nelson, W.C., Hallam, S.J., Jungbluth, S.P., Ettema, T.J.G., Tighe, S., Konstantinidis, K.T., Liu, W.-T., Baker, B.J., Rattei, T., Eisen, J.A., Hedlund, B., McMahon, K.D., Fierer, N., Knight, R., Finn, R., Cochrane, G., Karsch-Mizrachi, I., Tyson, G.W., Rinke, C., Genome Standards Consortium, Lapidus, A., Meyer, F., Yilmaz, P., Parks, D.H., Eren, A.M., Schriml, L., Banfield, J.F., Hugenholtz, P., Woyke, T., 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. 35, 725–731. https://doi.org/10.1038/nbt.3893

Bowers, R.M., Nayfach, S., Schulz, F., Jungbluth, S.P., Ruhl, I.A., Sheremet, A., Lee, J., Goudeau, D., Eloe-Fadrosh, E.A., Stepanauskas, R., Malmstrom, R.R., Kyrpides, N.C., Dunfield, P.F., Woyke, T., 2022. Dissecting the dominant hot spring microbial populations based on community-wide sampling at single-cell genomic resolution. ISME J 16, 1337–1347. https://doi.org/10.1038/s41396-021-01178-4

Brenner, D.J., Staley, J.T., Krieg, N.R., 2005. Classification of Procaryotic Organisms and the Concept of Bacterial Speciation, in: Brenner, D.J., Krieg, N.R., Staley, J.T., Garrity, G.M. (Eds.), Bergey's Manual® of Systematic Bacteriology: Volume Two: The Proteobacteria, Part A Introductory Essays. Springer US, Boston, MA, pp. 27–32. https://doi.org/10.1007/0-387-28021-9_4

Breusing, C., Johnson, S.B., Tunnicliffe, V., Clague, D.A., Vrijenhoek, R.C., Beinart, R.A., 2020. Allopatric and Sympatric Drivers of Speciation in Alviniconcha Hydrothermal Vent Snails. Mol Biol Evol 37, 3469–3484. https://doi.org/10.1093/molbev/msaa177

Brugna-Guiral, M., Tron, P., Nitschke, W., Stetter, K.-O., Burlat, B., Guigliarelli, B., Bruschi, M., Giudici-Orticoni, M.T., 2003. [NiFe] hydrogenases from the hyperthermophilic bacterium Aquifex aeolicus: properties, function, and phylogenetics. Extremophiles 7, 145–157. https://doi.org/10.1007/s00792-002-0306-3

Cabej, N.R., 2012. 1Species and Allopatric Speciation, in: Cabej, N.R. (Ed.), Epigenetic Principles of Evolution. Elsevier, London, pp. 707–723. https://doi.org/10.1016/B978-0-12-415831-3.00018-5

Cadillo-Quiroz, H., Didelot, X., Held, N.L., Herrera, A., Darling, A., Reno, M.L., Krause, D.J., Whitaker, R.J., 2012. Patterns of Gene Flow Define Species of Thermophilic Archaea. PLOS Biology 10, e1001265. https://doi.org/10.1371/journal.pbio.1001265

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M., Gormley, N., Gilbert, J.A., Smith, G., Knight, R., 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6, 1621–1624. https://doi.org/10.1038/ismej.2012.8

Chase, A.B., Arevalo, P., Brodie, E.L., Polz, M.F., Karaoz, U., Martiny, J.B.H., 2019. Maintenance of Sympatric and Allopatric Populations in Free-Living Terrestrial Bacteria. mBio 10. https://doi.org/10.1128/mBio.02361-19

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., Parks, D.H., 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics btz848. https://doi.org/10.1093/bioinformatics/btz848

Chen, Z., Wang, X., Song, Y., Zeng, Q., Zhang, Y., Luo, H., 2022. Prochlorococcus have low global mutation rate and small effective population size. Nat Ecol Evol 6, 183–194. https://doi.org/10.1038/s41559-021-01591-0

Cohan, F.M., 2019. Transmission in the Origins of Bacterial Diversity, From Ecotypes to Phyla, in: Microbial Transmission. John Wiley & Sons, Ltd, pp. 311–343. https://doi.org/10.1128/9781555819743.ch18

Cohan, F.M., 2016. Bacterial Speciation: Genetic Sweeps in Bacterial Species. Current Biology 26, R112–R115. https://doi.org/10.1016/j.cub.2015.10.022

Cohan, F.M., 2002. What are Bacterial Species? Annual Review of Microbiology 56, 457–487. https://doi.org/10.1146/annurev.micro.56.012302.160634

Cohan, F.M., Koeppel, A., 2008. The Origins of Ecological Diversity in Prokaryotes. Current Biology 18, R1024–R1034. https://doi.org/10.1016/j.cub.2008.09.014

Cohan, F.M., Perry, E.B., 2007. A Systematics for Discovering the Fundamental Units of Bacterial Diversity. Current Biology 17, R373–R386. https://doi.org/10.1016/j.cub.2007.03.032

Conrad, T.M., Lewis, N.E., Palsson, B.Ø., 2011. Microbial laboratory evolution in the era of genome-scale science. Molecular Systems Biology 7, 509. https://doi.org/10.1038/msb.2011.42

Cordero, O.X., Polz, M.F., 2014. Explaining microbial genomic diversity in light of evolutionary ecology. Nature Reviews Microbiology 12, 263–273. https://doi.org/10.1038/nrmicro3218

Crits-Christoph, A., Olm, M.R., Diamond, S., Bouma-Gregson, K., Banfield, J.F., 2020. Soil bacterial populations are shaped by recombination and gene-specific selection across a grassland meadow. The ISME Journal 14, 1834–1846. https://doi.org/10.1038/s41396-020-0655-x

de Queiroz, K., Donoghue, M.J., 1988. Phylogenetic Systematics and the Species Problem. Cladistics 4, 317–338. https://doi.org/10.1111/j.1096-0031.1988.tb00518.x

den Bakker, H.C., Didelot, X., Fortes, E.D., Nightingale, K.K., Wiedmann, M., 2008. Lineage specific recombination rates and microevolution in Listeria monocytogenes. BMC Evolutionary Biology 8, 277. https://doi.org/10.1186/1471-2148-8-277

Douglass, A.P., O'Brien, C.E., Offei, B., Coughlan, A.Y., Ortiz-Merino, R.A., Butler, G., Byrne, K.P., Wolfe, K.H., 2019. Coverage-Versus-Length Plots, a Simple Quality Control Step for de Novo Yeast Genome Sequence Assemblies. G3 (Bethesda) 9, 879–887. https://doi.org/10.1534/g3.118.200745

Dykhuizen, D.E., Green, L., 1991. Recombination in Escherichia coli and the definition of biological species. Journal of Bacteriology 173, 7257–7268. https://doi.org/10.1128/jb.173.22.7257-7268.1991

Eder, W., Huber, R., 2002. New isolates and physiological properties of the Aquificales and description of Thermocrinis albus sp. nov. Extremophiles 6, 309–318. https://doi.org/10.1007/s00792-001-0259-y

Edler, D., Holmgren, A., Rosvall, Ma., 2021. The MapEquation software package.

Eldredge, N., 1986. Hierarchic Interactions: The Evolutionary Process, in: Eldredge, N. (Ed.), Unfinished Synthesis: Biological Hierarchies and Modern Evolutionary Thought. Oxford University Press, p. 0. https://doi.org/10.1093/oso/9780195036336.003.0010

Forde, S.E., Beardmore, R.E., Gudelj, I., Arkin, S.S., Thompson, J.N., Hurst, L.D., 2008. Understanding the limits to generalizability of experimental evolutionary models. Nature 455, 220–223. https://doi.org/10.1038/nature07152

Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G., Hanage, W.P., 2009. The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity. Science 323, 741–746. https://doi.org/10.1126/science.1159388

Fraser, C., Hanage, W.P., Spratt, B.G., 2007. Recombination and the Nature of Bacterial Speciation. Science 315, 476–480. https://doi.org/10.1126/science.1127573

Friedman, J., Alm, E.J., Shapiro, B.J., 2013. Sympatric Speciation: When Is It Possible in Bacteria? PLoS ONE 8, e53539. https://doi.org/10.1371/journal.pone.0053539

Giovannelli, D., Sievert, S.M., Hügler, M., Markert, S., Becher, D., Schweder, T., Vetriani, C., 2017. Insight into the evolution of microbial metabolism from the deep-branching bacterium, Thermovibrio ammonificans. eLife 6, e18990. https://doi.org/10.7554/eLife.18990

González-Torres, P., Rodríguez-Mateos, F., Antón, J., Gabaldón, T., 2019. Impact of Homologous Recombination on the Evolution of Prokaryotic Core Genomes. mBio 10, e02494-18. https://doi.org/10.1128/mBio.02494-18

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., Tiedje, J.M., 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence

similarities. International Journal of Systematic and Evolutionary Microbiology 57, 81–91. https://doi.org/10.1099/ijs.0.64483-0

Greening, C., Biswas, A., Carere, C.R., Jackson, C.J., Taylor, M.C., Stott, M.B., Cook, G.M., Morales, S.E., 2016. Genomic and metagenomic surveys of hydrogenase distribution indicate H2 is a widely utilised energy source for microbial growth and survival. ISME J 10, 761–777. https://doi.org/10.1038/ismej.2015.153

Greenlon, A., Chang, P.L., Damtew, Z.M., Muleta, A., Carrasquilla-Garcia, N., Kim, D., Nguyen, H.P., Suryawanshi, V., Krieg, C.P., Yadav, S.K., Patel, J.S., Mukherjee, A., Udupa, S., Benjelloun, I., Thami-Alami, I., Yasin, M., Patil, B., Singh, S., Sarma, B.K., von Wettberg, E.J.B., Kahraman, A., Bukun, B., Assefa, F., Tesfaye, K., Fikre, A., Cook, D.R., 2019. Global-level population genomics reveals differential effects of geography and phylogeny on horizontal gene transfer in soil bacteria. Proc Natl Acad Sci U S A 116, 15200–15209. https://doi.org/10.1073/pnas.1900056116

Hanage, W.P., Fraser, C., Spratt, B.G., 2005. Fuzzy species among recombinogenic bacteria. BMC Biology 3, 6. https://doi.org/10.1186/1741-7007-3-6

He, Q., Wang, S., Feng, K., Michaletz, S.T., Hou, W., Zhang, W., Li, F., Zhang, Y., Wang, D., Peng, X., Yang, X., Deng, Y., 2023. High speciation rate of niche specialists in hot springs. ISME J. https://doi.org/10.1038/s41396-023-01447-4

Hedlund, B.P., Reysenbach, A.-L., Huang, L., Ong, J.C., Liu, Z., Dodsworth, J.A., Ahmed, R., Williams, A.J., Briggs, B.R., Liu, Y., Hou, W., Dong, H., 2015. Isolation of diverse members of the Aquificales from geothermal springs in Tengchong, China. Frontiers in Microbiology 6.

Henson, M.W., Lanclos, V.C., Pitre, D.M., Weckhorst, J.L., Lucchesi, A.M., Cheng, C., Temperton, B., Thrash, J.C., 2020. Expanding the Diversity of Bacterioplankton Isolates and Modeling Isolation Efficacy with Large-Scale Dilution-to-Extinction Cultivation. Appl Environ Microbiol 86, e00943-20. https://doi.org/10.1128/AEM.00943-20

Herbold, C.W., Lee, C.K., McDonald, I.R., Cary, S.C., 2014. Evidence of global-scale aeolian dispersal and endemism in isolated geothermal microbial communities of Antarctica. Nat Commun 5, 3875. https://doi.org/10.1038/ncomms4875

Hetzer, A., McDonald, I.R., Morgan, H.W., 2008. Venenivibrio stagnispumantis gen. nov., sp. nov., a thermophilic hydrogen-oxidizing bacterium isolated from Champagne Pool, Waiotapu, New Zealand. International Journal of Systematic and Evolutionary Microbiology 58, 398–403. https://doi.org/10.1099/ijs.0.64842-0

Hunt, D.E., David, L.A., Gevers, D., Preheim, S.P., Alm, E.J., Polz, M.F., 2008. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science 320, 1081–1085. https://doi.org/10.1126/science.1157890

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., Aluru, S., 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9, 5114. https://doi.org/10.1038/s41467-018-07641-9

Jessup, C.M., Kassen, R., Forde, S.E., Kerr, B., Buckling, A., Rainey, P.B., Bohannan, B.J.M., 2004. Big questions, small worlds: microbial model systems in ecology. Trends in Ecology & Evolution 19, 189–197. https://doi.org/10.1016/j.tree.2004.01.008

Johnson, J.L., 1980. The use of DNA homology in bacterial taxonomy and identification. Clinical Microbiology Newsletter 2, 1–3. https://doi.org/10.1016/S0196-4399(80)80006-7

Kassen, R., Rainey, P.B., 2004. The Ecology and Genetics of Microbial Diversity. Annu. Rev. Microbiol. 58, 207–231. https://doi.org/10.1146/annurev.micro.58.030603.123654

Kim, H.J., Jang, S., 2017. Optimization of a resazurin-based microplate assay for large-scale compound screenings against Klebsiella pneumoniae. 3 Biotech 8, 3. https://doi.org/10.1007/s13205-017-1034-9

Koeppel, A.F., Perry, E.B., Sikorski, J., Krizanc, D., Warner, A., Ward, D.M., Rooney, A.P., Brambilla, E., Connor, N., Ratcliff, R.M., Nevo, E., Frederick M. Cohan, 2008. Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics. PNAS 105, 2504–2509. https://doi.org/10.1073/pnas.0712205105

Kolde, R., 2018. pheatmap: Pretty Heatmaps.

Konstantinidis, K.T., Viver, T., Conrad, R.E., Venter, S.N., Rossello-Mora, R., 2022. Solar salterns as model systems to study the units of bacterial diversity that matter for ecosystem functioning. Current Opinion in Biotechnology 73, 151–157. https://doi.org/10.1016/j.copbio.2021.07.028

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Molecular Biology and Evolution 33, 1870–1874. https://doi.org/10.1093/molbev/msw054

Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.-H., Rognes, T., Ussery, D.W., 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35, 3100–3108. https://doi.org/10.1093/nar/gkm160

Le, S.Q., Dang, C.C., Gascuel, O., 2012. Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates. Molecular Biology and Evolution 29, 2921–2936. https://doi.org/10.1093/molbev/mss112

Li, S.-J., Hua, Z.-S., Huang, L.-N., Li, J., Shi, S.-H., Chen, L.-X., Kuang, J.-L., Liu, J., Hu, M., Shu, W.-S., 2014. Microbial communities evolve faster in extreme environments. Sci Rep 4, 6205. https://doi.org/10.1038/srep06205

Liu, B., Eydallin, G., Maharjan, R.P., Feng, L., Wang, L., Ferenci, T. 2017, 2017. Natural Escherichia coli isolates rapidly acquire genetic changes upon laboratory domestication. Microbiology 163, 22–30. https://doi.org/10.1099/mic.0.000405

Louca, S., 2021. The rates of global bacterial and archaeal dispersal. ISME J 1–9. https://doi.org/10.1038/s41396-021-01069-8

Majewski, J., 2001. Sexual isolation in bacteria. FEMS Microbiology Letters 199, 161–169. https://doi.org/10.1111/j.1574-6968.2001.tb10668.x

Majewski, J., Cohan, F.M., 1999. Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. Genetics 152, 1459–1474.

Mayr, E., 1999. Systematics and the origin of species, from the viewpoint of a zoologist, 1st Harvard University Press pbk. ed. ed. Harvard University Press, Cambridge, Mass.

Mayr, E., 1982. The Growth of Biological Thought: Diversity, Evolution, and Inheritance. Harvard University Press.

Mende, D.R., Sunagawa, S., Zeller, G., Bork, P., 2013. Accurate and universal delineation of prokaryotic species. Nat Methods 10, 881–884. https://doi.org/10.1038/nmeth.2575

Nazarian, P., Tran, F., Boedicker, J.Q., 2018. Modeling Multispecies Gene Flow Dynamics Reveals the Unique Roles of Different Horizontal Gene Transfer Mechanisms. Frontiers in Microbiology 9.

Nosil, P., 2012. Ecological Speciation. OUP Oxford.

Oliveira, P.H., Touchon, M., Rocha, E.P.C., 2016. Regulation of genetic flux between bacteria by restriction–modification systems. Proceedings of the National Academy of Sciences 113, 5658–5663. https://doi.org/10.1073/pnas.1603257113

Oliverio, A.M., Power, J.F., Washburne, A., Cary, S.C., Stott, M.B., Fierer, N., 2018. The ecology and diversity of microbial eukaryotes in geothermal springs. ISME J 12, 1918–1928. https://doi.org/10.1038/s41396-018-0104-2

Olm, M.R., Crits-Christoph, A., Diamond, S., Lavy, A., Matheus Carnevali, P.B., Banfield, J.F., 2020. Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. mSystems 5, e00731-19. https://doi.org/10.1128/mSystems.00731-19

Orakov, A., Fullam, A., Coelho, L.P., Khedkar, S., Szklarczyk, D., Mende, D.R., Schmidt, T.S.B., Bork, P., 2021. GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biology 22, 178. https://doi.org/10.1186/s13059-021-02393-0

Palys, T., Nakamura, L.K., Cohan, F.M., 1997. Discovery and Classification of Ecological Diversity in the Bacterial World: The Role of DNA Sequence Data. International Journal of Systematic and Evolutionary Microbiology 47, 1145–1156. https://doi.org/10.1099/00207713-47-4-1145

Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., Hugenholtz, P., 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Research 50, D785–D794. https://doi.org/10.1093/nar/gkab776

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25, 1043–1055. https://doi.org/10.1101/gr.186072.114

Pérez-Gracia, J.L., Gúrpide, A., Ruiz-Ilundain, M.G., Alegría, C.A., Colomer, R., García-Foncillas, J., Bermejo, I.M., 2010. Selection of extreme phenotypes: the role of clinical observation in translational research. Clin Transl Oncol 12, 174–180.

Petit, R.J., Excoffier, L., 2009. Gene flow and species delimitation. Trends in Ecology & Evolution 24, 386–393. https://doi.org/10.1016/j.tree.2009.02.011

Podar, P.T., Yang, Z., Björnsdóttir, S.H., Podar, M., 2020. Comparative Analysis of Microbial Diversity Across Temperature Gradients in Hot Springs From Yellowstone and Iceland. Front. Microbiol. 11, 1625. https://doi.org/10.3389/fmicb.2020.01625

Power, J.F., Carere, C.R., Lee, C.K., Wakerley, G.L.J., Evans, D.W., Button, M., White, D., Climo, M.D., Hinze, A.M., Morgan, X.C., McDonald, I.R., Cary, S.C., Stott, M.B., 2018. Microbial biogeography of 925 geothermal springs in New Zealand. Nat Commun 9, 2876. https://doi.org/10.1038/s41467-018-05020-y

Power, J.F., Lowe, C.L., Carere, C.R., McDonald, I.R., Cary, S.C., Stott, M.B., 2023. Temporal dynamics of geothermal microbial communities in Aotearoa-New Zealand. Frontiers in Microbiology 14.

Qin, Q.-L., Xie, B.-B., Zhang, X.-Y., Chen, X.-L., Zhou, B.-C., Zhou, J., Oren, A., Zhang, Y.-Z., 2014. A Proposed Genus Boundary for the Prokaryotes Based on Genomic Insights. J Bacteriol 196, 2210–2215. https://doi.org/10.1128/JB.01688-14

Queiroz, K. de, 2005a. Ernst Mayr and the modern concept of species. PNAS 102, 6600–6607. https://doi.org/10.1073/pnas.0502030102

Queiroz, K. de, 2005b. Different species problems and their resolution. BioEssays 27, 1263–1269. https://doi.org/10.1002/bies.20325

R Core Team, 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raina, V., Nayak, T., Ray, L., Kumari, K., Suar, M., 2019. Chapter 9 - A Polyphasic Taxonomic Approach for Designation and Description of Novel Microbial Species, in: Das, S., Dash, H.R. (Eds.), Microbial Diversity in the Genomic Era. Academic Press, pp. 137–152. https://doi.org/10.1016/B978-0-12-814849-5.00009-5

Rainey, P.B., Travisano, M., 1998. Adaptive radiation in a heterogeneous environment. Nature 394, 69–72. https://doi.org/10.1038/27900

Reysenbach, A.-L., Huber, R., Stetter, K.O., Ishii, M., Kawasumi, T., Igarashi, Y., Eder, W., L'Haridon, S., Jeanthon, C., 2001. Phylum BI. Aquificae phy. nov., in: Boone, D.R., Castenholz, R.W., Garrity, G.M. (Eds.), Bergey's Manual® of Systematic Bacteriology: Volume One : The Archaea and the Deeply Branching and Phototrophic Bacteria. Springer New York, New York, NY, pp. 359–367. https://doi.org/10.1007/978-0-387-21609-6_18

Rodriguez-R, L.M., Jain, C., Conrad, R.E., Aluru, S., Konstantinidis, K.T., 2021. Reply to: "Re-evaluating the evidence for a universal genetic boundary among microbial species." Nat Commun 12, 4060. https://doi.org/10.1038/s41467-021-24129-1

Rodriguez-R, L.M., Konstantinidis, K.T., 2014. Bypassing Cultivation To Identify Bacterial Species: Culture-independent genomic approaches identify credibly distinct clusters, avoid cultivation bias, and provide true insights into microbial species. Microbe Magazine 9, 111–118. https://doi.org/10.1128/microbe.9.111.1

Rosen, M.J., Davison, M., Bhaya, D., Fisher, D.S., 2015. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science 348, 1019–1023. https://doi.org/10.1126/science.aaa4456

Rosselló-Mora, R., 2006. DNA-DNA Reassociation Methods Applied to Microbial Taxonomy and Their Critical Evaluation, in: Stackebrandt, E. (Ed.), Molecular Identification, Systematics, and Population Structure of Prokaryotes. Springer, Berlin, Heidelberg, pp. 23–50. https://doi.org/10.1007/978-3-540-31292-5_2

Rotimi, A.M., Pierneef, R., Reva, O.N., 2018. Selection of marker genes for genetic barcoding of microorganisms and binning of metagenomic reads by Barcoder software tools. BMC Bioinformatics 19, 309. https://doi.org/10.1186/s12859-018-2320-1

Sakoparnig, T., Field, C., van Nimwegen, E., 2021. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. eLife 10, e65366. https://doi.org/10.7554/eLife.65366

Sandberg, M.E., Schellmann, D., Brunhofer, G., Erker, T., Busygin, I., Leino, R., Vuorela, P.M., Fallarero, A., 2009. Pros and cons of using resazurin staining for quantification of viable Staphylococcus aureus biofilms in a screening assay. Journal of Microbiological Methods 78, 104–106. https://doi.org/10.1016/j.mimet.2009.04.014

Schoch, C.L., Ciufo, S., Domrachev, M., Hotton, C.L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J.P., Sun, L., Turner, S., Karsch-Mizrachi, I., 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford) 2020, baaa062. https://doi.org/10.1093/database/baaa062

Seemann, T., 2018. Barrnap 0.9: Rapid ribosomal RNA prediction.

Shapiro, B.J., David, L.A., Friedman, J., Alm, E.J., 2009. Looking for Darwin's footprints in the microbial world. Trends in Microbiology 17, 196–204. https://doi.org/10.1016/j.tim.2009.02.002

Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabó, G., Polz, M.F., Alm, E.J., 2012. Population Genomics of Early Events in the Ecological Differentiation of Bacteria. Science 336, 48–51. https://doi.org/10.1126/science.1218198

Shapiro, B.J., Leducq, J.-B., Mallet, J., 2016. What Is Speciation? PLOS Genetics 12, e1005860. https://doi.org/10.1371/journal.pgen.1005860

Shapiro, B.J., Polz, M.F., 2015. Microbial Speciation. Cold Spring Harb Perspect Biol 7, a018143. https://doi.org/10.1101/cshperspect.a018143

Shapiro, B.J., Polz, M.F., 2014. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol 22, 235–247. https://doi.org/10.1016/j.tim.2014.02.006

Sharp, C.E., Brady, A.L., Sharp, G.H., Grasby, S.E., Stott, M.B., Dunfield, P.F., 2014. Humboldt's spa: microbial diversity is controlled by temperature in geothermal environments. ISME J 8, 1166–1174. https://doi.org/10.1038/ismej.2013.237

Shoemaker, W.R., Locey, K.J., Lennon, J.T., 2017. A macroecological theory of microbial biodiversity. Nat Ecol Evol 1, 1–6. https://doi.org/10.1038/s41559-017-0107

Singleton, C.M., Petriglieri, F., Kristensen, J.M., Kirkegaard, R.H., Michaelsen, T.Y., Andersen, M.H., Kondrotaite, Z., Karst, S.M., Dueholm, M.S., Nielsen, P.H., Albertsen, M., 2021. Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. Nat Commun 12, 2009. https://doi.org/10.1038/s41467-021-22203-2

Smith, D.J., Timonen, H.J., Jaffe, D.A., Griffin, D.W., Birmele, M.N., Perry, K.D., Ward, P.D., Roberts, M.S., 2013. Intercontinental Dispersal of Bacteria and Archaea by Transpacific Winds. Applied and Environmental Microbiology 79, 1134–1139. https://doi.org/10.1128/AEM.03029-12

Smith, N.H., Kremer, K., Inwald, J., Dale, J., Driscoll, J.R., Gordon, S.V., van Soolingen, D., Hewinson, R.G., Smith, J.M., 2006. Ecotypes of the Mycobacterium tuberculosis complex. J Theor Biol 239, 220–225. https://doi.org/10.1016/j.jtbi.2005.08.036

Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A.D., Kämpfer, P., Maiden, M.C.J., Nesme, X., Rosselló-Mora, R., Swings, J., Trüper, H.G., Vauterin, L., Ward, A.C., Whitman, W.B., 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. International Journal of Systematic and Evolutionary Microbiology 52, 1043–1047. https://doi.org/10.1099/00207713-52-3-1043

Stackebrandt, E., Goebel, B.M., 1994. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology.

International Journal of Systematic and Evolutionary Microbiology 44, 846–849. https://doi.org/10.1099/00207713-44-4-846

Steensels, J., Gallone, B., Voordeckers, K., Verstrepen, K.J., 2019. Domestication of Industrial Microbes. Current Biology 29, R381–R393. https://doi.org/10.1016/j.cub.2019.04.025

Stepanauskas, R., 2012. Single cell genomics: an individual look at microbes. Current Opinion in Microbiology, Antimicrobials • Genomics 15, 613–620. https://doi.org/10.1016/j.mib.2012.09.001

Swanson, R.V., 2001. Genome of Aquifex aeolicus, in: Methods in Enzymology, Hyperthermophilic Enzymes Part A. Academic Press, pp. 158–169. https://doi.org/10.1016/S0076-6879(01)30373-7

Szabo, G., Preheim, S.P., Kauffman, K.M., David, L.A., Shapiro, J., Alm, E.J., Polz, M.F., 2013. Reproducibility of Vibrionaceae population structure in coastal bacterioplankton. ISME J 7, 509–519. https://doi.org/10.1038/ismej.2012.134

Takai, K., Horikoshi, K., Komatsu, T., 2001. Hydrogenobacter subterraneus sp. nov., an extremely thermophilic, heterotrophic bacterium unable to grow on hydrogen gas, from deep subsurface geothermal water. International Journal of Systematic and Evolutionary Microbiology 51, 1425–1435. https://doi.org/10.1099/00207713-51-4-1425

Tamura, K., 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol 9, 678–687. https://doi.org/10.1093/oxfordjournals.molbev.a040752

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22, 4673–4680.

Thompson, L.R., Sanders, J.G., McDonald, D., Amir, A., Ladau, J., Locey, K.J., Prill, R.J., Tripathi, A., Gibbons, S.M., Ackermann, G., Navas-Molina, J.A., Janssen, S., Kopylova, E., Vázquez-Baeza, Y., González, A., Morton, J.T., Mirarab, S., Zech Xu, Z., Jiang, L., Haroon, M.F., Kanbar, J., Zhu, Q., Jin Song, S., Kosciolek, T., Bokulich, N.A., Lefler, J., Brislawn, C.J., Humphrey, G., Owens, S.M., Hampton-Marcell, J., Berg-Lyons, D., McKenzie, V., Fierer, N., Fuhrman, J.A., Clauset, A., Stevens, R.L., Shade, A., Pollard, K.S., Goodwin, K.D., Jansson, J.K., Gilbert, J.A., Knight, R., 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551, 457–463. https://doi.org/10.1038/nature24621

Troell, K., Hallström, B., Divne, A.-M., Alsmark, C., Arrighi, R., Huss, M., Beser, J., Bertilsson, S., 2016. Cryptosporidium as a testbed for single cell genome characterization of unicellular eukaryotes. BMC Genomics 17, 471. https://doi.org/10.1186/s12864-016-2815-y

Van Eaton, A.R., Harper, M.A., Wilson, C.J.N., 2013. High-flying diatoms: Widespread dispersal of microorganisms in an explosive volcanic eruption. Geology 41, 1187–1190. https://doi.org/10.1130/G34829.1

Van Rossum, T., Ferretti, P., Maistrenko, O.M., Bork, P., 2020. Diversity within species: interpreting strains in microbiomes. Nat Rev Microbiol 18, 491–506. https://doi.org/10.1038/s41579-020-0368-1

Vandamme, P., Pot, B., Gillis, M., Vos, P. de, Kersters, K., Swings, J., 1996. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol. Mol. Biol. Rev. 60, 407–438.

VanInsberghe, D., Arevalo, P., Chien, D., Polz, M.F., 2020. How can microbial population genomics inform community ecology? Philosophical Transactions of the Royal Society B: Biological Sciences 375, 20190253. https://doi.org/10.1098/rstb.2019.0253

Vaz-Moreira, I., Egas, C., Nunes, O.C., Manaia, C.M., 2011. Culture-dependent and culture-independent diversity surveys target different bacteria: a case study in a freshwater sample. Antonie van Leeuwenhoek 100, 245–257. https://doi.org/10.1007/s10482-011-9583-0

Viver, T., Conrad, R.E., Orellana, L.H., Urdiain, M., González-Pastor, J.E., Janet K. Hatt, Amann, R., Antón, J., Konstantinos T. Konstantinidis, Ramon Rosselló-Móra, 2021. Distinct ecotypes within a natural haloarchaeal population enable adaptation to changing environmental conditions without causing population sweeps. The ISME Journal 15, 1178–1191. https://doi.org/10.1038/s41396-020-00842-5

Vogan, A.A., Higgs, P.G., 2011. The advantages and disadvantages of horizontal gene transfer and the emergence of the first species. Biol Direct 6, 1. https://doi.org/10.1186/1745-6150-6-1

Vos, M., Didelot, X., 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J 3, 199–208. https://doi.org/10.1038/ismej.2008.93

Wang, J., Li, Y., Pinto-Tomás, A.A., Cheng, K., Huang, Y., 2022. Habitat Adaptation Drives Speciation of a Streptomyces Species with Distinct Habitats and Disparate Geographic Origins. mBio 13, e02781-21. https://doi.org/10.1128/mbio.02781-21

Wang, S., Dong, H., Hou, W., Jiang, H., Huang, Q., Briggs, B.R., Huang, L., 2014. Greater temporal changes of sediment microbial community than its waterborne counterpart in Tengchong hot springs, Yunnan Province, China. Sci Rep 4. https://doi.org/10.1038/srep07479

Ward, D.M., Cohan, F.M., 2005. Microbial diversity in hot spring cyanobacterial mats: pattern and prediction. Geothermal biology and geochemistry in Yellowstone National Park 185–201.

Ward, L., Taylor, M.W., Power, J.F., Scott, B.J., McDonald, I.R., Stott, M.B., 2017. Microbial community dynamics in Inferno Crater Lake, a thermally fluctuating geothermal spring. ISME J 11, 1158–1167. https://doi.org/10.1038/ismej.2016.193

Ward, L.M., Hemp, J., Shih, P.M., McGlynn, S.E., Fischer, W.W., 2018. Evolution of Phototrophy in the Chloroflexi Phylum Driven by Horizontal Gene Transfer. Front. Microbiol. 9. https://doi.org/10.3389/fmicb.2018.00260

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., Moore, L.H., Moore, W.E.C., Murray, R.G.E., Stackebrandt, E., Starr, M.P., Truper, H.G.Y. 1987, 1987. Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. International Journal of Systematic and Evolutionary Microbiology, 37, 463–464. https://doi.org/10.1099/00207713-37-4-463

Weltzer, M.L., Miller, S.R., 2013. Ecological Divergence of a Novel Group of Chloroflexus Strains along a Geothermal Gradient. Applied and Environmental Microbiology 79, 1353–1358. https://doi.org/10.1128/AEM.02753-12

Whitaker, R.J., Grogan, D.W., Taylor, J.W., 2005. Recombination shapes the natural population structure of the hyperthermophilic archaeon Sulfolobus islandicus. Mol. Biol. Evol. 22, 2354–2361. https://doi.org/10.1093/molbev/msi233

Wiedenbeck, J., Cohan, F.M., 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35, 957–976. https://doi.org/10.1111/j.1574-6976.2011.00292.x

Xu, Y.-F., Dong, X.-M., Luo, C., Ma, S.-N., Xu, J.-L., Cui, Y.-D., 2022. Nitrogen Enrichment Reduces the Diversity of Bacteria and Alters Their Nutrient Strategies in Intertidal Zones. Frontiers in Marine Science 9.

Yang, S.-J., Kang, I., Cho, J.-C., 2016. Expansion of Cultured Bacterial Diversity by Large-Scale Dilution-to-Extinction Culturing from a Single Seawater Sample. Microb Ecol 71, 29–43. https://doi.org/10.1007/s00248-015-0695-3

Zink, R.M., 2012. The Geography of Speciation: Case Studies from Birds. Evo Edu Outreach 5, 541–546. https://doi.org/10.1007/s12052-012-0411-4

Zuckerkandl, E., Pauling, L., 1965. Molecules as documents of evolutionary history. Journal of Theoretical Biology 8, 357–366. https://doi.org/10.1016/0022-5193(65)90083-4

CHAPTER 3: Adaptive mechanisms of metabolically limited bacteria in response to ecological perturbations

3.1. Abstract

Defining and delineating bacterial species is hampered by the lack of consensus on the mechanisms driving bacterial speciation. Previous observations of the co-occurrence of clonal and recombinogenic populations in metabolically limited *Aquificota* bacteria suggest that gene-specific and genome-level sweeps simultaneously facilitate bacterial speciation. The presence of gene-specific sweeps in metabolically limited bacteria challenges the *ecotype* theory, which postulates that an ecologically homogeneous population responds to ecological shifts solely through genome-wide sweeps due to the rarity of horizontal gene flow. Therefore, in this study, I examined whether *Aquificota* populations, due to their limited metabolic flexibility, exclusively undergo genome-wide sweeps in response to ecological perturbations regardless of whether they are clonal or recombining populations. Longitudinal observations of Aquificota populations from natural bacterial communities were done to investigate the changes in their abundance, intrapopulation diversity, and population structure in response to ecological shifts. Comparative genome analysis revealed six distinct populations of *Aquificota* that exhibited varying degrees of horizontal gene flow, with three populations (*Sulfurhydrogenibium* sp., *UBA11096* sp. cluster 1 Green population, and *Hydrogenobacter thermophilus*) appearing to have a clonal population structure. The other three populations (*Venenivibrio sp.* and UBA11096 sp. cluster 2 Blue and Orange populations) exhibited a recombining population structure. Analysis of the population structure and intrapopulation heterogeneity of *Aquificota* populations suggested that clonal and recombining populations of *Aquificota* employ different mechanisms to respond to ecological perturbations. Clonal populations of *Aquificota* appeared to have undergone a genome-wide sweep even before the study and thus consistently appeared as highly homogeneous populations. Meanwhile, recombining populations of *Aquificota* did not undergo a genome-wide sweep as a response to ecological perturbations but exhibited a persistence of the dominant genomic background across time within the population. The co-occurrence of different adaptation mechanisms in closely related metabolically limited bacteria challenges the applicability of ecotype theory in recombining populations and supports the generalisability of the *ecological population* theory as a model of bacterial speciation.

3.2. Introduction

Defining bacterial species is hampered by the lack of consensus on the role of horizontal gene flow in bacterial speciation. The current prevailing theory of bacterial speciation, the ecotype theory, assumes the rarity of horizontal gene flow in natural bacterial populations and postulates the formation of ecologically cohesive populations ('ecotype') as driven by genome-wide sweeps and periodic selection (Atwood et al., 1951; Cohan, 2001; Cohan and Perry, 2007; Smith et al., 2006). However, a competing theory, dubbed the ecological population theory, acknowledges the prevalence of horizontal gene flow (Bobay and Ochman, 2017; Vos and Didelot, 2009) and posits that ecologically cohesive populations can also form through gene-specific sweeps (Cadillo-Quiroz et al., 2012; Hanage et al., 2005; Shapiro et al., 2012; Takeuchi et al., 2015; Whitaker et al., 2005). Analysis of horizontal gene flow in genomes from the same named species in the NCBI databases has shown that more than half of the named species experience horizontal gene flow, suggesting that gene-level sweep could be more prevalent than expected (Bobay and Ochman, 2017). Unfortunately, the relative occurrences of gene-specific sweeps and genome-wide sweeps in natural bacterial communities are not yet well elucidated (Bendall et al., 2016; Utter et al., 2022). Previous studies have been primarily limited to heterotrophic bacteria (Arevalo et al., 2019; den Bakker et al., 2008; Hanage et al., 2005) or those with metabolic redundancies (Kashtan et al., 2014; Whitaker et al., 2005).

On the other hand, evidence of genome-wide sweeps in natural bacterial communities was initially observed in members of *Chlorobium* sp. (Bendall et al., 2016). Metagenomic studies of lake bacterial communities across the span of eight years have shown five metagenome-assembled genomes (MAGs) exhibiting a decrease in diversity across their genome. One genome, Chlorobium-111, lost diversity in nearly all gene loci, consistent with the ecotype theory of speciation (Cohan, 2016). In the same study, five MAGs with a reduction of diversity in only a few genomic regions were also observed, thereby suggesting a gene-specific sweep within the populations. Another MAG (Polynucleobacter-258) was also observed to have no diversity in a 24kbp region, suggesting a gene-specific sweep in a recombining population (Bendall et al., 2016). However, Cohan (2016) argued that rather than due to homologous recombination, the observed sweeps in narrow regions of the genome are brought by ecological heterogeneity within the population. In their 'Adapt Globally-Act Locally' model of ecotype theory (Majewski and Cohan, 1999), a general adaptive mutation can be shared through infrequent recombination with other ecotypes harbouring their distinct adaptive mutations. A narrow sweep can occur when selective pressure acts on the general adaptive mutation, resulting in diverse genomic backgrounds (Cohan, 2016; Majewski and Cohan, 1999). This theory suggests that narrow sweeps are more prevalent in bacteria with diverse metabolic flexibility, which allows rapid shift and adaptation to new ecological niches (Chen et al., 2021; Kopac et al., 2014). In contrast, metabolically constrained bacteria are more likely to be composed of a single ecotype (Cohan, 2016). However, this model has been criticised as less than optimal since infrequent horizontal gene flow in numerous highly diverse ecotypes can be argued as less probable than a single

recombining population undergoing a gene-specific sweep (Bendall et al., 2016; Shapiro, 2016). As such, the debate over the roles of gene-specific and genome-wide sweeps on bacterial speciation continues with multiple studies aligning to either ecotype (Utter et al., 2022; Viver et al., 2021; Wang et al., 2020) and ecological population theory (Arevalo et al., 2019a; Bao et al., 2016; Lopes et al., 2018; Marttinen and Hanage, 2017; Miller and Carvey, 2019) for bacterial speciation and delineation.

Understanding speciation in bacteria necessitates the examination of the roles of gene-specific and genome-wide sweep in natural bacterial communities. Our previous study (Chapter 2) observed a correlation between intrapopulation heterogeneity and the degree of horizontal gene flow in *Aquificota* populations. Highly homogeneous populations of *Aquificota* exhibited a clonal population structure, whereas higher intrapopulation heterogeneity was observed in recombining populations of *Aquificota*. The co-occurrence of clonal lineages and recombining populations suggests that gene-specific and genome-wide sweeps both drive bacterial speciation in these closely related populations with similar metabolic limitations. Gene-specific sweeps in metabolically limited populations such as *Aquificota* challenge the model of bacterial speciation postulated by the ecotype theory. However, since the previous study (Chapter 2) focused only on a single time point, it raised the question that the observed diversity driven by gene flow may be a temporary diversification event between periodic selections (Cohan, 2016; Wiedenbeck and Cohan, 2011).

Therefore, in this study, I examined the roles of gene-specific and genome-wide sweeps in bacterial diversification through a longitudinal study observing the changes in the population diversity and structure of *Aquificota* bacteria in response to ecological shifts. Analysis of the intrapopulation heterogeneity and population structure of *Aquificota* bacteria were expanded to four additional time points spanning a year. Utilising the same Kuirau Park geothermal springs in Chapter 2, I isolated cohabiting closely related *Aquificota* bacteria to test the null hypothesis that ecologically and genotypically cohesive populations of *Aquificota*, regardless of whether they are clonal or recombining populations, undergo genome-wide sweeps in response to environmental shifts due their metabolically limited lifestyle. Recent horizontal gene flow analysis was used to construct the population structure of isolated *Aquificota* bacteria at each time point. Geochemical and metagenomic analyses were used to identify ecological shifts within the natural bacterial communities. Metagenomic data were also used to observe intrapopulation heterogeneity and abundance changes in natural *Aquificota* populations as they responded to ecological shifts.

3.3. Methodology

3.3.1. Bacterial Isolation and Cultivation

Geothermal features in Kuirau Park in Rotorua, Bay of Plenty, with a high abundance of different *Aquificota* species, were selected for this study (Figure 3.1). Based on the 1000 Springs project (Power et al., 2018), Kuirau Park geothermal feature 101 (designated as KRP1) and Kuirau Park geothermal feature 9 (KRP2) have *Aquificota* relative abundances of 73.3% and 58.4%, respectively (1000springs.org.nz). Kuirau Park geothermal feature 9 (KRP2) has a short (~0.25m) outflow channel leading to an extinct geothermal feature (KRP2B). Unlike KRP1 and KRP2, KRP2B has no underground water source; thus, temperature and water levels depend on run-offs and KRP2 overflow (Figure 3.1).

Figure 3.1 Local map of Kuira Park geothermal spring sites. Within the wider Kuirau Park geothermal area, Kuirau Park sites 1 and 2 are located 112m apart from each other. Kuirau Park site 1 (KRP1) is a single pool with its own underground water source, while Kuirau Park 2 is composed of two pools, KRP2 and KRP2B, with an underground water source only located in KRP2. KRP2B water level is dependent on KRP2 via a ~0.25m channel. *Imagery © 2024 Airbus, Bay of Plenty TA's, CNES/Airbus, Maxar Technologies, Waikato District Council, Map data © 2024 Google.*

Geothermal fluids were carefully sampled from the water column of the two features to minimise disturbance and contamination from sediments. Initial sampling was done in July 2019 and November 2019 as part of an initial survey of different geothermal features for *Aquificota* abundance. Following the observations of fluctuating water levels of the sites between July 2019 and January 2020, subsequent sampling of geothermal fluids was done in January 2020, March 2020, July 2020, and September 2020. Geothermal fluids for bacterial isolation were only sampled from June 2019 to July 2020 due to the limited capacity for genome sequencing. Meanwhile, geothermal fluid sampling for metagenome sequencing was only done at points from January 2020 to September 2020. Additionally, KRP2B was sampled for bacterial isolation and metagenome studies only from March 2020 onwards since it was added to the study later.

3.3.2. Geochemical Analysis

Geothermal fluids for geochemical analysis were sampled from November 2019 to July 2020 alongside samples for bacterial isolation. Temperature was measured on-site using a multi-mode probe reader. Fluid samples were collected, and pH was measured when the temperature of the samples was around 20°C. Geothermal fluids were filtered and collected for geochemical analysis. Fluids for total metal analysis were analysed using ICP-MS following acid digestion. Meanwhile, inorganic nutrients such as nitrate, nitrite, and ammonia, alongside dissolved reactive phosphorus, were measured using flow injection analysis. Sulfate and chloride concentrations were analysed through ion exchange chromatography.

3.3.3. Serial Dilution-to-Extinction

Bacterial isolation was performed similarly using the proven serial dilution-to-extinction method (Figure 3.2) and cultivation for *Aquificota* using deep-well plates (see Chapter 2). Collected geothermal fluid was used to inoculate 96 deep-well plates (2 mL max. total volume per well). Inoculation of the initial dilution was done at 1:1 dilution, while subsequent serial dilution was done at 1:4 dilution of inoculant to media for up to eight iterations. The plates were sealed using a gas-permeable membrane (Merck Breathe-Easy® sealing membrane) and placed in a custom-built environmental chamber with a gas mixture of 80:5:15 H_2:CO_2:air gas mixture for *Aquificota* cultivation (Hedlund et al., 2015; Hetzer et al., 2008; Reysenbach et al., 2001). After two weeks of incubation at 70°C, growth was checked using a resazurin cell viability assay (Kim and Jang, 2017; Sandberg et al., 2009). The last dilution with growth was used as an inoculant for the subsequent passage, and a similar process was repeated for the second and third passages. The last dilution of the third passage was again grown in plates/tubes to increase their biomass for DNA extraction.

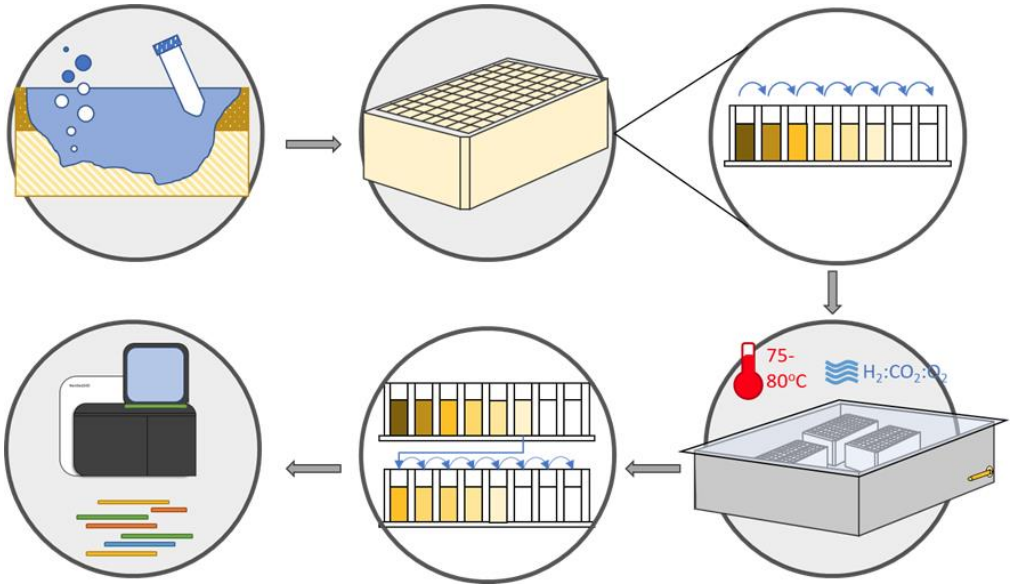

Figure 3.2 The serial dilution-to-extinction isolation and sequencing protocol for 96 deep-well plates. Geothermal fluid samples were directly inoculated into deep well plates or tubes with minimal enrichment. Serial dilution was performed for up to a max of 257 (156,250) dilution factor. The plates were put inside a modified gas chamber. The chamber and the tubes were degassed and refilled with an 80:5:15% ratio of H2:CO2:Air gas mixture and were then incubated for up to two weeks at 70°C. Growth was checked using a resazurin cell viability assay. The last well with growth is used for the subsequent passage and serial dilution-to-extinction. The final dilution with growth in the last passage was used for genome sequencing.

3.3.4. Genome extraction and sequencing

3.3.4.1. Genomic and Metagenomic DNA Extraction

Bacterial suspension grown from the last dilution with the growth of the serial-dilution to extinction step was pelleted and washed with PBS buffer. DNA extraction was performed for the June 2019 samples using a MOBIO PowerFood™ extraction kit following the manufacturer's instructions with the RNAse A treatment step. The extracted DNA was cleaned and concentrated using the Zymo DNA Clean-up and Concentrator Kit. Meanwhile, a modified lysozyme-SDS DNA extraction protocol was used for the rest of the samples (Goldenberger et al., 1995). Bacterial cells were centrifuged and washed with TE buffer pH8.0, which were then re-suspended in 150 µL lysozyme buffer (10mg/mL in 10 mM Tris-HCl, with 0.1 M NaCl, 1 mM EDTA, and 1% v/v TRITON® X-100) and incubated at 37°C for 30 minutes. Afterwards, 150 µL of lysis buffer (50 mM tris-HCl pH 7.5, 50 mM EDTA, 10 mM CaCl2, 1% w/v SDS, 0.1% w/v Triton® X-100, and 1 mM DTT) was added and incubated for 90 minutes. The DNA was precipitated using a salt alcohol precipitation method (1/10 volume of 3 M sodium acetate pH 5.2 with two volumes of 96% ice-cold ethanol) at -20°C for at least an hour. The DNA was washed with 70% ethanol twice and resuspended with nuclease-free water.

For metagenomic DNA, extraction was done using a modified Phenol-Chloroform-Isoamyl Alcohol extraction protocol for sterivex filters (Brazelton and Pendleton, 2021) with a freeze-thaw cycle step. Two litres of geothermal fluid was passed through a sterivex filter (0.2µm) with both female and male luer-lok in either end. Sterivex filters were stored at -80°C until extraction. Before extraction, the outflow end was closed with a luer-lok cap and sealed with parafilm. DNA extraction buffer (see SUPP. 7 Table 7.1) was added to the intake end of the sterivex filter until it filled the filter chamber, leaving the intake's neck empty (~1.4-1.8mL). The filter was then incubated at -80°C for 12 minutes and then quickly transferred to a 70°C bath for 15 minutes and repeatedly done for 90 minutes. Afterwards, 36 µL of lysozyme (50 mg/mL) were added into the sterivex and incubated at 37°C for 30 min with constant agitation, followed by the addition of 45 µL of Proteinase K and 90 µL 20% SDS. The sterivex were then incubated at 65°C for 90 minutes with constant agitation. Afterwards, the fluid was removed from the filter and centrifuged to remove CTAB-associated material. Extract fluid was transferred into a clean tube and added with equal volume buffered Phenol-Chloroform-Isoamyl Alcohol (25:24:1) mixture. The extract was gently shaken before centrifuging at 14,000 xg for 1 min, and the aqueous layer was transferred into a new tube with an equal volume of chloroform-isoamyl alcohol (24:1). After shaking and centrifuging the tube, the aqueous supernatant is used for ethanol precipitation and washing similar to that used in genome extraction, albeit with incubation at -20°C for overnight. The quality and concentration were checked using spectrophotometry of the 260/280 Absorbance ratio and Qubit fluorescence DNA high sensitivity kit. The total DNA was normalised to 100ng for each sequencing run.

3.3.4.2. Sequencing, pre-processing, and assembly

From Kuirau Park-grown isolates, 192 were used for genome sequencing and analysis across different time points. Extracted DNA from the isolates were sequenced using Illumina HiSeq 2100 with a 2x150bp library size, and base calling was done using the sequencer's built-in software RTA. Acquired reads were analysed using FastQC (ver. 0.12.0), and low-quality reads and adapters were removed using Trimmomatic (ver. 0.40) (Bolger et al., 2014). Clean reads were assembled into contigs using the de novo assembler SPAdes (Bankevich et al., 2012). The axenicity of the genomes was checked by plotting k-mer coverage against contig length to remove contaminants and separate co-occurring contigs into genome bins (Douglass et al., 2019). Genome completeness and strain heterogeneity were checked using checkM (Parks et al., 2015). Genomes with contamination >2% were not used for downstream analysis.

The metagenomic DNA was also sequenced using the Illumina HiSeq 2100 with a 2x150bp library size. Sequenced metagenomic reads were base-called on the built-in RTA software. The reads were further cleaned using the BBMap (ver. 39.01) pipeline with BBDuk,

which was used for adapter trimming, contaminant removal, and filtering low-quality reads (Bushnell, 2014). Tadpole was used for error correction of the reads. The quality was confirmed afterwards using FastQC.

3.3.5. Population structure analysis

3.3.5.1. Population Delineation

Genomes were initially identified through comparison against the Genome Taxonomic Database (R202) using the GTDB-TK (Chaumeil et al., 2019; Parks et al., 2022). Identified *Aquificales* and *Hydrogenothermales* genomes were used for Average nucleotide identity to delineate population boundary using the FastANI algorithm (Jain et al., 2018) and visualised using pheatmap v1.0.12 package in R V4.2.1 (Kolde, 2018; R Core Team, 2022). Genomes with 95% ANI similarity are classified as distinct populations for subsequent studies. An *a priori* delineation of populations was also done using gene flow boundary delineation using PopCOGenT (Arevalo et al., 2019a). The PopCOGenT pipeline utilises a pairwise comparison of entire genomes. Genomes from all sites and time points belonging to the same GTDB genus were run together to determine the extent of horizontal gene flow between closely related genomes. Additionally, the genomes were also segregated per time point, and PopCOGenT was run only with genomes within those time points to observe how population structure and horizontal gene flow change across time points.

3.3.5.2. Swept Region Annotation

Previously identified recombining clusters of UBA11096 sp. cluster 2 using PopCOGenT analysis were used for swept gene analysis. Swept gene analysis has been limited to genomes isolated from July 2019 and July 2020. Swept genes from the flexible genome analysis were done using the modified pipeline of PopCOGenT (Arevalo et al., 2019a). A gene is identified as a part of the flexible genome when it is swept within a population or PopCOGenT subcluster (e.g., UBA11096 sp. cluster Blue or Orange population) but is not found throughout the other populations within its cluster. The PopCOGenT pipeline identifies the clusters of genes that swept a population. A custom Python script was used to create a workflow that fetches swept gene sequences from a representative genome and performs functional annotation and other downstream analysis. The sequences were translated into predicted proteins using Prodigal ver. 4.8.4 (Hyatt et al., 2010), which were afterwards clustered per population into clusters of orthologous genes (COGs) using mmseqs2 (ver. 15.6f452) (Steinegger and Söding, 2017). Functional analysis of each cluster was done using the Diamond ver. 2.1.8.162 BLASTp module

(Buchfink et al., 2021) against the Uniprot database 2023_1 release (Coudert et al., 2023) and mapped to COG categories using eggNOG mapper ver. 5 (Huerta-Cepas et al., 2017).

Similar to swept genes identified in the flexible genome, identification of swept genes in the core genome is limited to the recombining *UBA11096* sp. cluster 2 Blue and Orange populations from July 2019 and July 2020. The swept genes from the core genome were analysed using the modified PopCOGenT Core Genome pipeline in a Docker container (glbcabria/popcogent-revised:1.3.2). The PopCOGenT Core Genome pipeline identifies a swept gene in the core genome through identification of low within-population nucleotide diversity regions that also form a monophyletic clade. As such, population-differentiating SNPs in these regions are identified as SNPs present within all members of the population but not in other populations. Open reading frame and protein prediction were identified from swept regions using Prodigal ver. 4.8.4. The predicted genes were annotated using the Diamond BLASTp module against the Uniprot database release 2023_1, and COG categories were identified using the eggNOG mapper (ver. eggNOG5 database).

Comparisons of genes between the swept regions of the UBA11096 sp. cluster 2 Blue and Orange populations from different time points were performed using the OrthoVenn pipeline ver. 3 (Sun et al., 2023). Non-clustered predicted proteins were clustered using the OrthoMCL algorithm with an e-value of $1x10^{-2}$ and an inflation factor of 1.50. The clustered proteins were visualised using a Venn diagram generated by ComplexUpset ver. 1.3.5 (Krassowski, 2020; Lex et al., 2014) and the ggplot2 package in R ver. 4.3.1 (R Core Team, 2022). Statistical analyses were done using the base R packages.

3.3.6. Genome extraction and sequencing

3.3.6.1. Abundance estimation and community diversity

Community diversity and abundance estimation of *Aquificota* samples were analysed using fractional scaling (FracMinHash) through Sourmash (ver. 4.8.4). For abundance estimation, high-quality genomes with high completeness and low contamination from each established population were selected as representative genomes for each identified *Aquificota* population. MinHash signatures for the representative genomes and the metagenomic reads were created through random sampling of k-mer content with a k-mer size of 31 with a fractional scaling of 1/1000 and accounting abundance as stated in the user guide (Pierce et al., 2019). The abundance of the population representatives was measured through a containment query of the sourmash signatures of the representative genomes against the metagenome reads using the *sourmash gather* method. Meanwhile, community diversity and composition were measured using the metagenomic read signatures as containment queries against the GTDB representative

genome database and the Genbank (March 2022) genomes using *sourmash gather* and *sourmash tax* methods.

Statistical analyses of community diversity and site geochemistry were all conducted in the R ver 4.3.1 (R Core Team, 2022), with plots generated using the ggplot2 ver. 3.4.4 (Wickham, 2016) and data analysis was performed using the tidyverse ver 2.0 (Wickham et al., 2019). The alpha diversity was measured using the Shannon Index and species richness, which were visualised using the Vegan v2.6-4 (Dixon, 2003) and Phyloseq v1.44.0 (McMurdie and Holmes, 2013) packages in R. The community structure of the geothermal pools across time points was assessed through robust Aitchinson distances of robust centre log-ratio (rCLR) on sequence abundance through phyloseq and vegan packages following compositional data analysis guidelines (Gloor et al., 2017). Statistical analysis of the correlation between environmental variables and microbial diversity was measured through the Mantel Test in Microeco package ver 1.1.0 (Liu et al., 2021). Environmental factors were also plotted using a distance-based Redundancy Analysis (RDA) through the Microeco package in R.

3.3.6.2. Single Nucleotide Polymorphism Analysis

The prevalence of SNPs in the *Aquificota* populations from the community metagenomic data was measured following the GATK-for-Microbes pipeline using GATK v4.2.6.0 (Benjamin et al., 2019; Van der Auwera et al., 2013). Metagenomic reads that matched >95% identity to the representative genome of each *Aquificota* population were used for SNP analysis to minimise the contamination between populations. The read mapping filtering cut-off was delineated using Minimap2 by mapping reads of the acquired *Aquificota* genomes between and within population representatives (see SUPP. 8 Figure 7.1). Afterwards, reads were downsampled based on the lowest coverage of a sample per population. Downsampling was done in triplicates. The reads were then mapped into reference genomes using BWA-mem, and duplicates were marked using PICARD (v2.26.10). SNPs and Indel detection were then done using the Mutect2 (GATK v4.2.6.1) short variant discovery caller, followed by FilterMutectCalls to remove artefacts and errors. Statistical analysis and Pearson correlation of the SNPs with the relative abundance was done using base packages in R.

3.4. Results and Discussion

3.4.1. Community diversity and water level shifts over time

Both Kuirau Park Site 1 and 2 (Figure 3.1) have shown drastic changes in water levels over a span of a year. It was visually observed that the water level was low (Figure 3.3A) during the July 2019 sampling and increased to the maximum height during January 2020 (Figure 3.3A). Since KRP2B has no direct underground water source and relies on KRP2 outflows, its water height appears distinctly lower than other pools during periods of low water levels. However, KRP2B fills up to the same heights as KRP2 during high water levels due to the absence of any notable outflow channels from KRP2B. However, no significant temperature and pH changes were associated with the rise of water level in KRP1 and KRP2 (Figure 3.3B, see SUPP. 9 Table 7.2). The correlation between these cyclic changes in water levels and other nearby geothermal features in Kuirau Park and Rotorua geothermal areas has not yet been established. However, cyclic patterns in water level shifts and temperature have been observed previously in Inferno Crater Lake, 21 km southeast of Kuirau Park (Scott, 1994; Ward et al., 2017). Water level and temperature were shown to undergo 20-day cycling between high and low levels (Scott, 1994), with the shift in temperature associated with changes in water level brought by the influx and retreat of geothermal fluids (Vandemeulebrouck et al., 2008). Shift in the microbial populations in the Inferno crater has been observed to result in population turnover between the archaeon *Sulfolobus-like spp.* and *Hydrogenobaculum* sp. as a response to thermal cycling (Ward et al., 2017). Therefore, a shift in Kuirau Park's water level may elicit similar perturbations that affect its community and intrapopulation diversity.

Therefore, I investigated the microbial diversity in Kuirau Park 1 and 2 across the different time points through compositional data analysis of the metagenomic reads (Gloor et al., 2017). The diversity and abundance of the microbial community were assessed with FracMinHash using metagenomic reads as containment queries against the GenBank reference genome dataset (O'Leary et al., 2016; Pierce et al., 2019). The alpha diversity and Shannon index of observed species showed that both KRP1 and KRP2 microbial communities remained relatively constant across time points from January 2020 to July 2020 but dropped in September 2020 (Figure 3.3C). Comparisons between time points of community structure showed a significant change in relative abundances between September 2020 and prior time points in KRP1 and KRP2, implying a perturbation of microbial diversity during or prior to September 2020 (Figure 3.3D). A significant change was also observed between January 2020 and July 2020 in KRP2, suggesting a shift in microbial diversity.

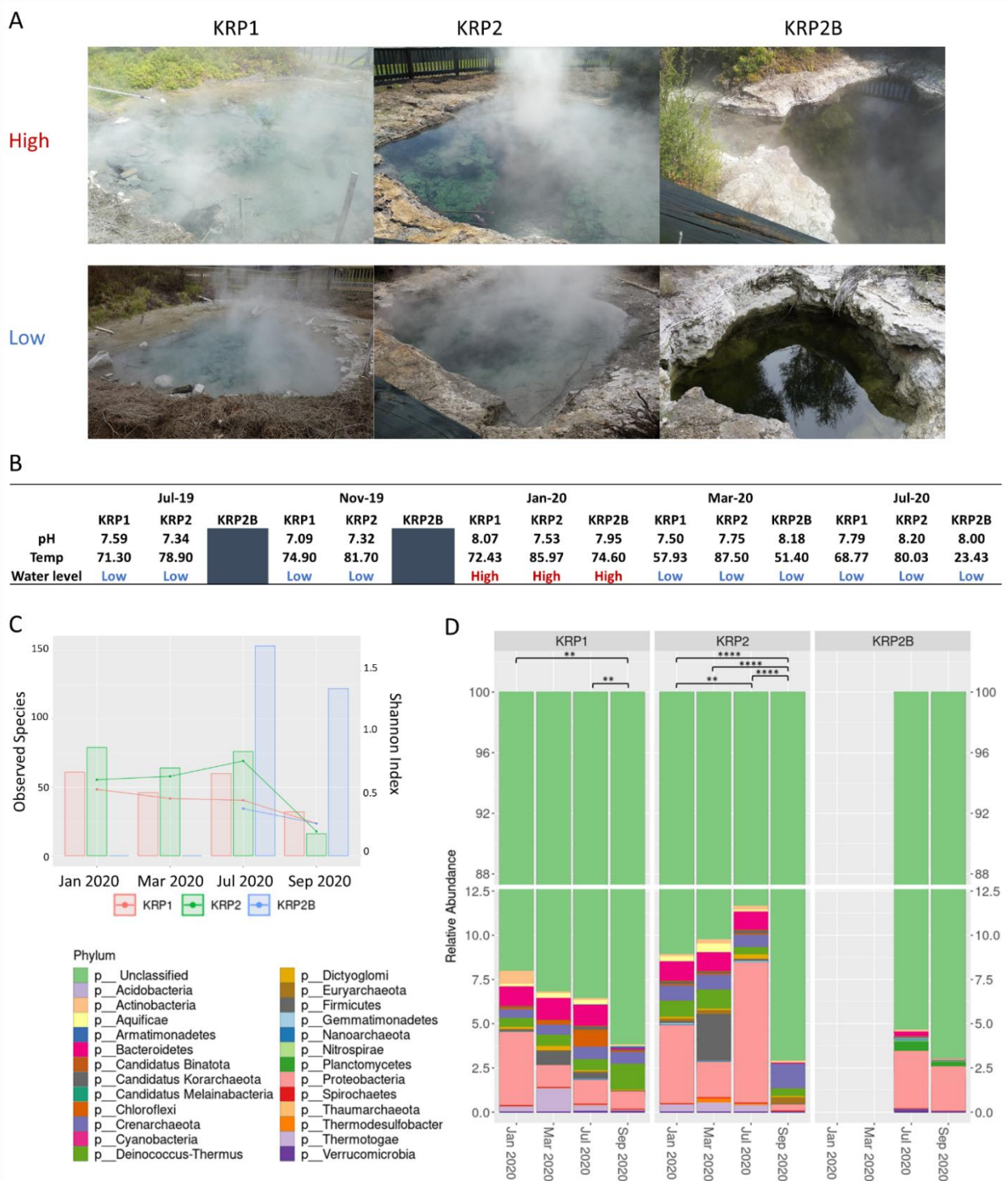

Figure 3.3 Water level and microbial diversity shifts in Kuirau Park pools. The water levels of Kuira Park sites 1 and 2 have been visually observed to change between high and low levels (A), albeit with no known confirmed period between states, as it was only recorded in a single time point in this study (B). Water level shows no significant correlation between pH and Temperature values (B, see SUPP. 9). Meanwhile, alpha diversity within the sites remained constant from Jan 2020 to Jul 2020 but dropped during September 2020, indicating a perturbation in microbial diversity in each pool (C). The shift in the microbial community was observed in the significant changes in the metagenomic read abundance between September 2020 and previous time points (D). No geochemical data were collected in September 2020 and KRP2B for July and November 2019 time points. Statistical analysis was done between time points on each site using the Wilcoxon signed-rank test. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1.

Previous studies of microbial communities in Kuirau Park observed that KRP1 and KRP2 were dominated by *Aquificota* bacteria at 70% and 90% relative abundances, respectively (Power et al., 2018). However, the current analysis showed a maximum of 0.25% relative abundance of *Aquificota* metagenomic signatures in the site (Figure 3.3D), which may be due to differences in DNA extraction, sequencing, data analysis methods, and the six-year time difference. The phylum *Proteobacteria* dominated most sites across different time points, followed by *Bacteroidetes* and *Firmicute*. Beta diversity analysis using robust Aitchinson distances showed the microbial community to be significantly different across time points at each site (p=0.006, PerMANOVA) and between sites for each time point (p=0.001, PerMANOVA), but no significant difference between water levels (see SUPP. 9 Table 7.3). Analysis of the impact of other measured geochemical parameters using the Mantel test showed no significant association with microbial diversity (see SUPP. 9 Table 7.5). Investigation of the effect of measured geochemical parameters on members of *Aquificota* bacteria also showed no significant correlation (see SUPP. 9 Table 7.6). However, redundancy analysis showed a potential relationship between *Aquificales'* relative abundance and water level, while a negative relationship can be observed between *Hydrogenothermales* bacteria and temperature (see SUPP. 9 Figure 7.2). The lack of significant association between microbial diversity and environmental parameters indicates that unmeasured geochemical parameters or hydrological processes may have driven the observed significant changes in the microbial community.

3.4.2. Both clonal and recombining populations can arise from distinct but closely related *Aquificota* populations

To investigate whether the observed perturbation in the microbial communities affects the intrapopulation diversity of *Aquificota* bacteria, first, I isolated and sequenced 179 genomes of *Aquificota* bacteria and examined whether they exhibit similar intra-species genomic variation associated with varying degrees of horizontal gene flow as previously observed (Chapter 2). Two sets from prior time points used in metagenomic studies, July 2019 (Chapter 2) and November 2019, were included in the genomes sequenced. No geochemical and metagenomic data were available for those time points, and no isolates were sequenced for September 2020 due to COVID-19 restrictions and logistical complications. Comparison of genomic similarity of all sequenced genomes grouped them into six distinct populations (Figure 3.4A) using the traditional cut-off for species at 95% ANI similarity (Goris et al., 2007; Jain et al., 2018). The six populations were classified as members of *Hydrogenothermales* and *Aquificales* families and categorised into four distinct genera using the Genome Taxonomic Database (GtDB) pipeline (see SUPP. 10). Three populations belonging to *UBA11096* clustered with genomes from the previous study (Chapter 2). At the same time, the other three populations were classified as related to *Hydrogenobacter thermophilus*, *Sulfurihydrogenibium* sp., and

Venenivibrio. Based on previous clustering of recent horizontal gene flows, members of *UBA11096* population 1 and 2 from Chapter 2 are referred to as *UBA11096* sp. cluster 2 Blue and Orange populations, whereas *UBA11096* population 3 is referred to as *UBA11096* sp. cluster 1 Green population from here on.

Table 3.1 Average Nucleotide Identity of the six *Aquificota* populations

Population	Mean ANI	Standard Deviation	Min Value	Max Value
UBA11096 sp. cluster 2 Blue	99.214	0.520	95.306	99.996
UBA11096 sp. cluster 2 Orange	99.423	0.342	98.508	99.999
UBA11096 sp. cluster 1 Green	99.977	0.021	99.864	99.999
H. thermophilus	99.997	0.003	99.989	99.999
Sulfurihydrogenibium sp.	99.936	0.070	99.684	99.998
Venenivibrio	97.616	1.839	89.242	99.999

In agreement with the previous study (Chapter 2), significant intra-species genomic variation was observed in *UBA11096* sp. cluster 2 Blue and Orange populations, while the *UBA11096* sp. cluster 1 Green population appeared as a highly homogeneous population. *H. thermophilus* and *Sulfurihydrogenibium* sp. populations also appeared highly homogeneous (mean ANI >99.9%), implying that they are highly clonal populations similar to the *UBA11096* sp. cluster 1 Green population (Table 3.1). Meanwhile, *Venenivibrio* genomes exhibited heterogeneity with ANI values ranging from 89.242-99.999% (97.615±1.838% mean sd). However, one genome behaves erratically as it matches at >97% ANI to most genomes but shows lower ANI (89-95% ANI) to a select few. Since most genomes clustered within the ANI cut-off, *Venenivibrio* genomes were treated as a single population for the rest of the study. Analysis of genome similarity showed that more heterogeneous populations exhibited a substantial gap in ANI distribution between 99.5% and 99.9% ANI value (see SUPP. 11, Figure 7.3). A gap in ANI similarity (99.2-99.8%) has been observed in other cohabiting archaea and bacteria and has been attributed to the substantial gene-content differences between 'ecotypes' (Rodriguez-R et al., 2023a; Viver et al., 2024, 2021).

Since intra-species genomic variation in UBA11096 sp. has been previously associated with varying degrees of horizontal gene flow (Chapter 2), all the sequenced genomes were analysed for recent horizontal gene flow using PopCOGenT. The genomes were grouped and analysed per genus, based on GtDB classification, since PopCOGenT is designed primarily to analyse horizontal gene flow (i.e. homologous recombination) in closely related taxa (Arevalo et al., 2019a). Recent horizontal gene flow analysis confirmed the results in Chapter 2 in which highly homologous populations are composed of clonal lineages, while a recombining population structure is observed in bacteria with higher intra-species heterogeneity. Genomes of *H.*

thermophilus and *Sulfurihydrogenibium* sp. and *UBA11096* sp. cluster 1 Green population appeared as a single node, whereas *UBA11096* sp. cluster 2 Blue and Orange population exhibited a high degree of horizontal gene flow within each population with a few lower intensity linkages to other populations (Figure 3.4B).

Meanwhile, the *Venenivibrio* population was shown to be composed of a few clonal lineages connected by horizontal gene flow, albeit at a much lower degree of flow, as indicated by the lower genome length bias, the fraction of shared genomic region between bacteria (302.42±29.71). *Venenivibrio*'s mean length bias was significantly lower than those observed in intrapopulation linkages of *UBA11096* sp. cluster 2 (1636.67±651.1) and comparable to the genome length bias between UBA11096 sp. cluster 2 Blue population in Canada and New Zealand (Chapter 2). Since the frequency of horizontal gene flow is often associated with sequence similarity (Friedman et al., 2013; Hanage et al., 2005; Shapiro et al., 2012), the weaker horizontal gene flow between *Venenivibrio* members, coupled with lower ANI values, imply that the population is in the process of diverging and forming a new species. Differentiation of co-occurring Venenivibrio bacteria within the same habitat suggests sympatric speciation is potentially driven by decreasing horizontal gene flow or recombination (Friedman et al., 2013). A recent study has observed contrasting growth conditions and lifestyles between *Venenivibrio* and *Sulfurihydrogenibium*, forming alternate environmental niches and potentially leading to sympatric speciation (Power et al., 2024). Similarly, the formation of alternate environmental niches within co-habiting Venenivibrio could also drive the increasing intrapopulation heterogeneity (Baquero et al., 2021) and lower degree of horizontal gene flow within them, leading to speciation (Shapiro et al., 2012).

Lastly, the observed ANI intra-population heterogeneity at recombining *UBA11096* sp. populations suggests that varying degrees of horizontal gene flow could explain the 'ANI gap' at 99.4-99.8% (Rodriguez-R et al., 2023). Analysis of very closely-related genomes in the database has shown a significant gap in the distribution of ANI values at 99.4-99.8%. The gap is prevalent across bacterial taxa, including clonal and recombining bacteria. Genomes with ANI values higher than the 'ANI gap' have been correlated with sequence-types or discrete sequence clusters (Rodriguez-R et al., 2023a; Viver et al., 2024). These discrete sequence clusters are suggested to be composed of clonal complexes that have undergone similar evolutionary and ecological responses. The mechanisms in the formation of these sequence types remain speculative and were suggested to be established by either gene-level or genome-wide sweeps (Viver et al., 2024, 2021). Therefore, the results here could provide evidence that preferential homologous recombination could drive the formation of the 'ANI Gap'. With the analysis of more cohabiting bacterial genome sequences and studies, the role of horizontal gene flow in the ANI gap and early diversification will become more apparent.

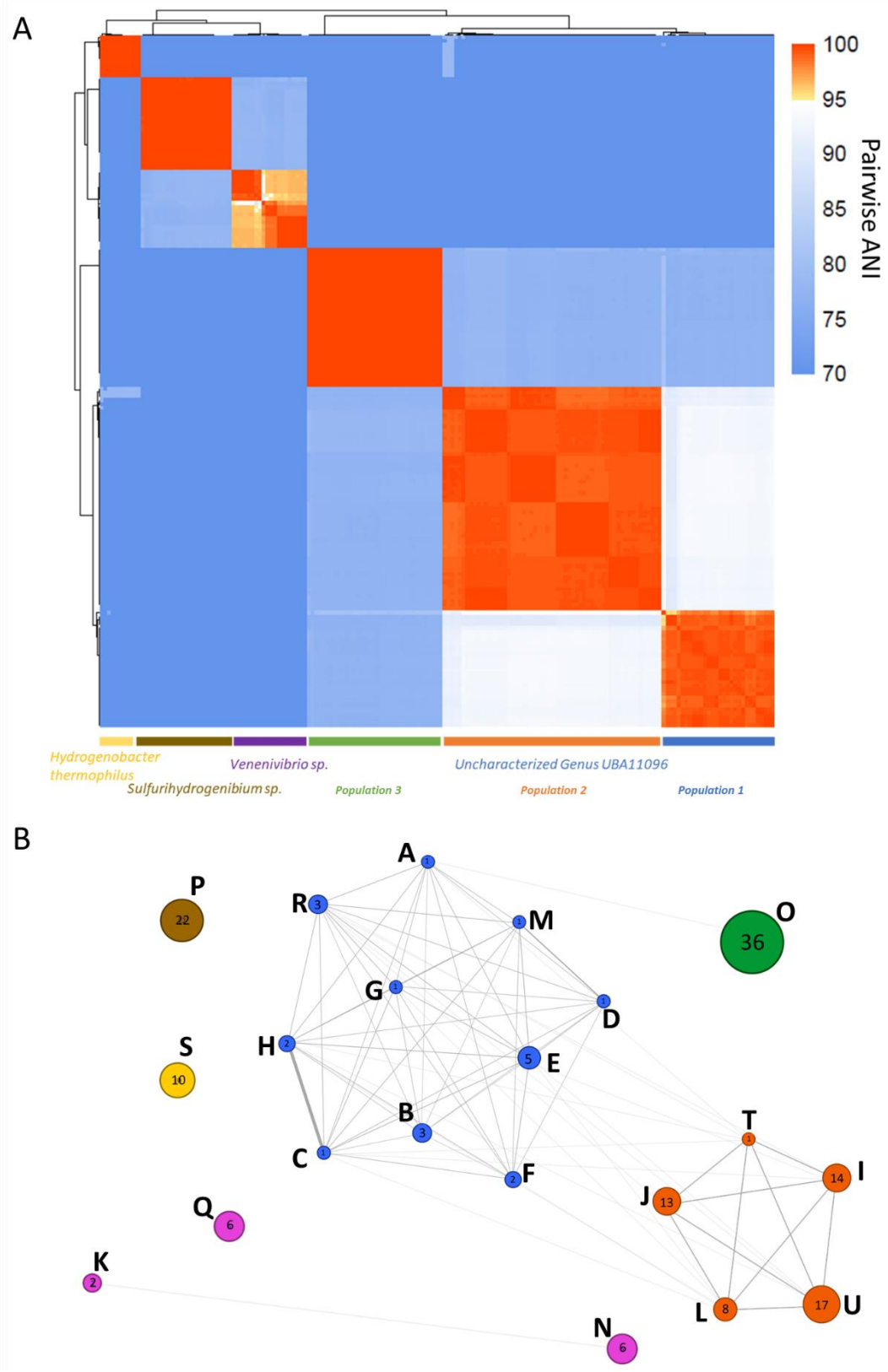

Figure 3.4 ANI and PopCOGenT analysis delineates *Aquificota* genomes in six distinct populations. The computed ANI grouped the Kuirau Park clustered genomes into six distinct populations (A), with ANI similarity of >95% within each population. (B) PopCOGenT analysis of high-quality genomes with checkM contamination <2% (see SUPP. 12 Figure 7.4) shows different degrees of horizontal gene flow across the different populations. Each edge's thickness and colour intensity indicate the degree of recent horizontal gene flow. Each circle represents a clonal group or distinct genomic background (ANI >99.9%) with colour matching its population as shown in ANI heatmap (A). Each circle is labelled with its clonal group ID and the number of genomes grouped in the clonal complex. The metadata for each clonal group is in SUPP. 13 (Table 7.8).

3.4.3. Recombining populations exhibit intrapopulation turnover of dominant genomic background across time

Given the observed shift in microbial community and intrapopulation diversity of recombining *Aquificota* populations (i.e., *UBA11096* sp. cluster 2 and *Venenivibrio*), it is likely that perturbations affecting the microbial community also drove intrapopulation heterogeneity in *Aquificota* populations. To examine if there was any shift in *Aquificota* populations and understand how *Aquificota* bacteria responded to change, I analysed recent horizontal gene flows for each population by time points to examine changes in their population structure. As expected, the highly homogeneous populations of *Sulfurihydrogenibium* sp. and *UBA11096* sp. cluster 1 Green population appeared consistently clonal across time, with the other highly homologous *Hydrogenobacter thermophilus* population appearing only during March 2020 as a single clonal group (Figure 3.5). The more heterogeneous recombining *UBA11096* sp. cluster 2 Blue and Orange populations exhibited a more diverse population structure over time. A higher degree of linkages (recent horizontal gene flow) was observed in July 2019 and July 2020 than in the intervening period in which only the *UBA11096* sp. cluster 2 Orange population was isolated and formed a single clonal group at each time point. Meanwhile, *Venenivibrio* appeared as unlinked clonal groups separated between the March and July 2020 time points. It could be posited that the segregation of the K and N clonal groups of *Venenivibrio* at different time points acted as a gene flow barrier, leading to the observed lower degree of horizontal gene flow (Figure 3.4) (Friedman et al., 2013; Shapiro et al., 2012).

At first glance, the shift in *UBA11096* sp. cluster 2 Orange population structure from a recombining population structure to a population consisting of a single clonal group during the intervening periods may indicate that the population underwent a genome-wide sweep (Cohan, 2016; Shapiro and Polz, 2014; Takeuchi et al., 2015). However, closer inspection of the clonal group IDs (Figure 3.5) suggested a likely turnover within the population of the dominant clonal groups or unique genomic backgrounds across time. Different clonal groups of UBA11096 sp. cluster Orange population appeared at each time point, implying a shift in dominant genomic background within the population rather than an outright genome-wide sweep in which one clonal group survived selection and expanded to take over the population. In addition to shifting dominant genomic background, both *UBA11096* sp. cluster 2 Blue and Orange populations also had clonal groups that appeared at both July 2019 and July 2020 time points and, thus, providing support into an intrapopulation turnover of dominant genomic background and the lack of genome-wide sweep.

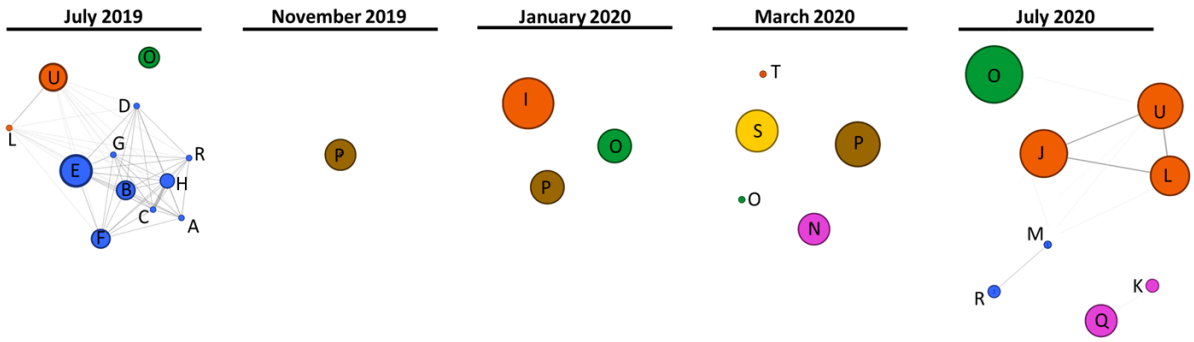

Figure 3.5 Horizontal gene flow analysis of *Aquificota* populations exhibits a shift in population structure across time. A higher degree of horizontal gene flow was observed in July 2019 and July 2020 than in the intervening period for members of recombining UBA11096 cluster 2 Blue and Orange populations. Meanwhile, the interim time points (November 2019 – March 2020) have been dominated by distinct clonal groups from both clonal and recombining *Aquificota* populations. It should be noted that the clonal groups (see SUPP. 13 Table 7.8) of UBA11096 sp. cluster 2 Blue and Orange populations change for each time point with only clonal groups U, L and R appearing in more than one-time point, implying a turnover in the dominant genomic background in the population across time.

Next, I examined whether the population structure shift mirrored intrapopulation heterogeneity in *Aquificota* bacteria found in the metagenomic data. Through the stringent mapping of metagenomic reads to representative genomes of each *Aquificota* population, I investigated changes in single nucleotide polymorphisms (SNPs), which represented the intrapopulation heterogeneity of each *Aquificota* population and could provide insight into the presence of genome-wide sweep or lack thereof in natural bacterial communities. Independent measurement of the relative abundance of *Aquificota* was also performed since the previous metagenomic analysis of relative abundance does not include representative genomes that belong to the studied *Aquificota* populations. Analysis of SNPs in the highly homologous and clonal populations of *UBA11096* sp. cluster 1 Green population and *H. thermophilus* showed minimal change in intrapopulation heterogeneity across time, even with significant changes in observed read relative abundances (Figure 3.6). The constantly low intrapopulation diversity of homologous populations reflected their unchanging population structure, which consisted of a single clonal group (Figure 3.4).

The absence of a total purge in SNP diversity in these clonal populations did not indicate a lack of genome-wide sweep. The clonal population structure and the constantly low SNP diversity independent, albeit drastic changes in abundance in these clonal populations, may suggest that adaptive clones have been selected and reached fixation before the study. Subsequently, these adaptive organisms remain at low neutral diversity as constant strong purifying selection minimises the chance for diversification (Cohan, 2001). Given that the geothermal Kuirau Park sites likely encounter putatively cyclic patterns in environmental

conditions similar to other geothermal features in the region (Scott, 1994; Ward et al., 2017), periodic selection may only increase cohesion in the population's ecological traits (Cohan, 2005; Shapiro and Polz, 2014). Furthermore, the decrease in diversity coupled with the increase in abundance in *Sulfurihydrogenibium* sp. (Figure 3.6) suggests a potential expansion of an adaptive clone of *Sulfurihydrogenibium* sp. during March 2020, a sign of genome-wide sweep.

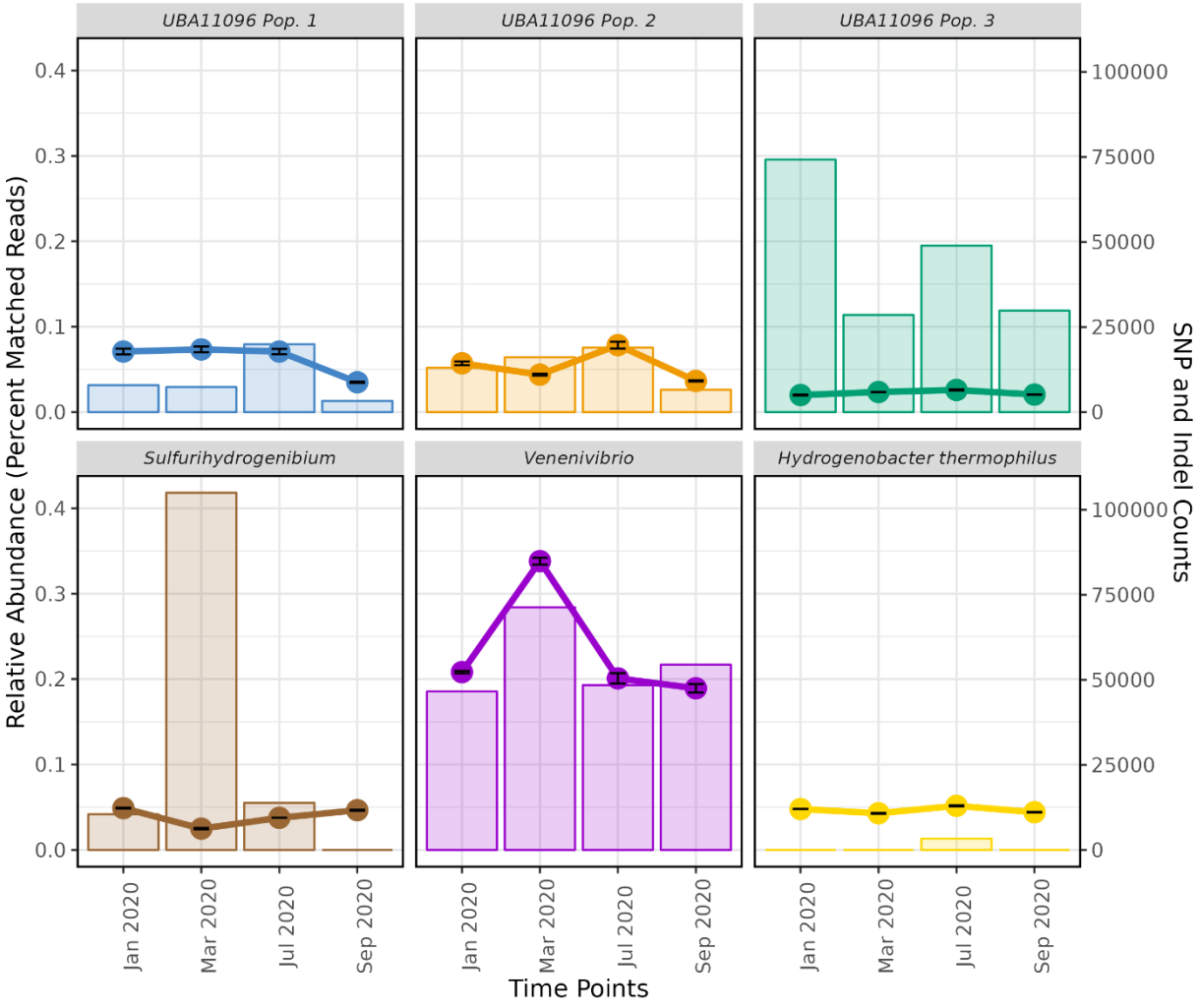

Figure 3.6 Read and SNP abundance of natural *Aquificota* populations. Read abundance (bars) was measured by containment analysis of metagenomic reads to the representative genomes of each population. SNPs and Indel (line) were calculated using the filtered reads with >95% identity to the representative genomes. The observed changes in the whole population are similar to those observed within individual KRP1 and KRP2 sites (see SUPP. 14 Figure 7.5), with clonal populations' SNP diversity remaining constant across time points and recombining UBA11096 sp. cluster population's SNP diversity consistent with their abundance. Meanwhile, *Venenivibrio* also shows a similar pattern in the number of SNPs that mirrors its abundance**.** Calculation of pearson's correlation showed a positive correlation between SNP counts and relative abundance in *Venenivibrio* sp, *UBA11096* sp. Blue and Orange populations, a weak positive correlation in *Hydrogenobacter thermophilus* and a negative correlation in *Sulfurihydrogenibium* sp. and *UBA11093* sp. Green population (See SUPP 7.8 Figure 7.6).

Meanwhile, the more heterogeneous *Venenivibiro* sp. and *UBA11096* sp. cluster 2 Blue and Orange populations were observed to have the highest intrapopulation diversity (i.e., highest SNP abundance) concurrent with their highest relative abundance, suggesting that their intrapopulation heterogeneity was linked with population size (Bobay and Ochman, 2018). Interestingly, *Venenivibrio* sp. exhibited the highest count of SNPs among the *Aquificota* population, reflecting the high intrapopulation diversity measured through ANI. The high intrapopulation diversity and abundance of *Venenivibrio* in the metagenomic data suggested that coexisting diverging subpopulations of *Venenivibrio* exist in Kuirau Park, which undergo sympatric speciation and potentially temporal niche-partitioning as they have been proposed to be associated in higher co-occurring phylotype (Dini-Andreote et al., 2014), such as diverging populations.

For *UBA11096* sp. cluster populations, the persistence of observed genotypes between July 2019 and July 2020 (Figure 3.5) and the minimal changes in neutral diversity (Figure 3.6), suggest that the population likely undergoes gene-specific sweep. In contrast to unchanging diversity in more clonal populations, the increase in SNPs in *UBA11096* sp. cluster 2 for July 2020 represented an increase in heterogeneity that may suggest an already diverse gene pool that has been suppressed or lies dormant (Polz et al., 2013) at previous time points, which were able to increase in relative abundance at the July 2020 time point due to potentially more tolerant environmental conditions (Lasky, 2019; Pompei et al., 2023). Therefore, the increase in intrapopulation heterogeneity, the abundance, and the observed shift in the dominant genomic background suggest that *UBA11096* sp. populations did not likely undergo genome-wide sweeps but intrapopulation turnover of the dominant genomic background between July 2019 and July 2020. A turnover in the dominant genomic background within a population is likely to exist in recombining populations due to the higher initial genetic diversity brought by gene-specific sweeps (Lasky, 2019; Polz et al., 2013; Pompei et al., 2023). Gene-specific sweeps allow diverse genomic backgrounds to persist along the shared ecological niche dimension, which lowers the burden of carrying genes as these can be dispersed across diverse genetic variants (Kupczok et al., 2018; Lasky, 2019). Therefore, The diverse genomic background acts as a genetic reservoir that can facilitate the shift of previously standing neutral genes or alleles to adaptive alleles when subjected to selective pressure (Barton, 2001; Lasky, 2019; Wilson et al., 2017). Additionally, genetic reservoirs protect potentially beneficial but seldomly used genes from being lost from the population (Granato and Kümmerli, 2017; Morris et al., 2012).

One interesting question is what differentiates the changes in diversity, or lack thereof, observed between the clonal and recombining populations of *Aquificota,* especially populations of *UBA11096* sp.? Given the taxonomic relatedness and cohabitation of these *Aquificota* populations, they should experience similar environmental perturbations that apply selective

pressure. I posit that the difference is due to the predisposition of recombining populations' adaptations to undergo 'soft' selective sweeps rather than 'hard' sweeps. 'Soft' selective sweeps primarily occur when adaptive genes/alleles appear faster than those genes can reach fixation within the community, allowing multiple adaptive alleles to co-occur through the population simultaneously. Meanwhile, 'hard' selective sweeps, in which only a single adaptive allele swept through the population, occur when the time for the appearance of novel adaptive alleles is long enough to reach fixation within the population (Hermisson and Pennings, 2017; Messer and Petrov, 2013). Novel adaptive alleles often arise when new or more substantial selective pressure purges the diversity of previously adaptive clones. 'Soft' sweeps happen when multiple adaptive mutations appear from previously neutral or deleterious standing mutations or adaptations from multiple *de novo* mutations in large populations (Hermisson and Pennings, 2017). The large genetic reservoir in recombining populations supplants the need for large population sizes since standing adaptive alleles/genes can arise simultaneously in the presence of novel or stronger selective pressure (Lasky, 2019). Therefore, it can be argued that gene-specific sweeps in bacteria often result in 'soft' selective sweeps because they allow the preservation of diverse genomic backgrounds. The large repertoire of potentially adaptive genes/alleles in 'soft' sweeps has been attributed to the rapid adaptation of the population (Hermisson and Pennings, 2017; Messer and Petrov, 2013; Wilson et al., 2017), which was evident in the observed high SNP diversity of recombining populations. Additionally, 'soft' selective sweeps have been suggested to drive evolutionary rescue—in which population decline driven by environmental change is rescued from extinction by genetic adaptation (Wilson et al., 2017)—which is also exemplified by the intrapopulation turnover in recombining *UBA11096* sp. cluster 2 population.

The changes in relative abundance and intrapopulation heterogeneity in metagenomic data and the shift in dominant genomic background contradict the null hypothesis that metabolically limited populations of *Aquificota* bacteria respond to environmental shifts solely through genome-wide sweeps. These results imply that even chemolithoautotrophic populations of *Aquificota* respond differently to environmental perturbations. A non-recombining population would likely undergo a genome-wide sweep, while gene-specific sweeps are more likely to occur in recombining populations (Shapiro, 2016). Therefore, the applicability of the ecological population theory to *Aquificota* bacteria may suggest the potential generalisability of the ecological population theory as a model of speciation in wider bacterial taxa.

3.4.4. Swept regions have similar functions between populations.

Since gene-specific sweep facilitates the spread of adaptive genes preferentially within recombining populations, functional analysis of the horizontally swept regions within the population allows us to infer biotic and abiotic factors applying selective pressure towards the

population (Arevalo et al., 2019a, 2019b; Shapiro and Polz, 2014). To examine the factors affecting the recombining *UBA11096* sp. populations, I identified and annotated the swept genes from the core and flexible genome of the *UBA11096* sp. cluster 2 population. Analysis of the core genome using the PopCOGenT pipeline identifies the population-differentiating alleles in genomic regions that are common within both populations (i.e., the Blue and Orange populations). Population-differentiating alleles are determined to be alleles that are common within the population but vary in other populations. In contrast, swept regions in the flexible genome are genes that have swept within the population but not to the rest of the cluster (Arevalo et al., 2019a). Clustering of orthologous genes (COGs) from July 2019 and July 2020 showed that *UBA11096* sp. cluster 2 Blue and Orange populations mostly shared common genes within their swept regions (Figure 3.7A). A comparison of the COGs from the same populations but different time points showed large numbers of shared COGs between different time points, with 144 and 73 COGs shared within Blue and Orange populations, implying that these populations experienced similar selective pressures between time points (see SUPP. 16 Figure 7.7). However, the highest number of COGs was observed in the Orange population from July 2019 and the Blue population from July 2020, mainly consisting of genes from the core swept region. The commonality between the swept genes from the two populations suggests that they may have experienced similar selective pressure from their environment.

Annotation of the COGs showed a similar distribution of the swept genes from the core genome across functional categories between the different populations (p=0.62) (Figure 3.7B see SUPP. 15 Table 3.11). The swept genes were mainly distributed across replication (L), cell wall/membrane (M), motility(N), and protein turnover (O) for both populations and, additionally, with the highest relative abundance found in translational processes (J) (see SUPP. 15 Table 3.10). Interestingly, a contrasting trend in the changes in the numbers of swept genes in the core genome was observed between the populations in which the number of swept genes increases in Blue populations while there is an observed decrease in the Orange population (Figure 3.7C). Given the similarity in the functional categories in the swept genes, this contrasting trend may suggest that the population experienced similar selective pressure, albeit at different time points. Unfortunately, since the swept genes observed in the core genome belonged to broad categories of housekeeping and regulating functions, pinpointing specific environmental or biotic pressure may be difficult. Given that these genes are related to division, motility, and biofilm formation (see SUPP. 16 Table 7.13), one potential theory is that recent gene flows allow synchronisation within the population responding to the same environmental cue to undergo cell division, biofilm formation, dissociation, or colonisation of their environment. The temporal synchronisation of bacterial processes allows quorum and coupling of metabolic activity, thereby increasing population and community survival (Li and Nair, 2012; Ottesen et al., 2013).

The swept genes of *UBA11096* sp. cluster 2 Blue populations in the flexible genome were distributed similarly to the swept regions of its core genome (p=0.35) (Figure 3.7B, see SUPP. 15 Table 7.11). Meanwhile, *UBA11096* sp. cluster 2 Orange populations had very few COGs identified in its flexible genome. Among the five COGs found to be shared across all populations and time points, CRISPR-associated protein Cas3 is the only annotated cluster, suggesting a role of horizontal gene flow in maintaining bacterial host defence system. Interestingly, one difference between the core and flexible genome was the high abundance of swept genes in the flexible genome of *UBA11096 sp*. cluster 2 Blue population in the energy production and conversion category (C) (Figure 3.7B). The swept genes were components of the electron transport chain (e.g., cytochrome-c, cytochrome-b6, ferredoxin-2) and nitrogen-fixation pathway (e.g., *nifE, nifN*) genes (see SUPP. 15 Table 7.9 and SUPP. 16 Table 7.14). Other nitrogen fixation genes (*nifB*, *nifH, nifD, nifK, nifX, nifW, nifT)* were also swept within the flexible genome of *UBA11096* sp. cluster 2 Blue population. Members of *Aquificota* phyla, specifically of genera *Thermocrinis* and *Hydrogenobacter*, were among the most prevalent taxa with *nifH* genes in high-temperature geothermal pools (Hamilton et al., 2014; Loiacono et al., 2012; Song et al., 2022). The taxonomically close relationship of *UBA11096 sp*. with *Thermocrinis* and *Hydrogenobacter* genera and the presence of the main nitrogenase gene cassette (*nifH, nifD, nifK*) in the swept genes suggests that members of *UBA11096 sp*. play a role in nitrogen fixation within the geothermal pool ecosystem. Additionally, the presence of energy production and nitrogen-fixing genes only in the swept regions of the *UBA11096 sp*. cluster 2 Blue population indicated that these processes could be one of the population-defining traits that distinguish the two closely recombinogenic *UBA11096 sp*. cluster 2 populations. Further examination of the differences in enzymatic kinetics of these genes ascertains their role as putative population-defining genes.

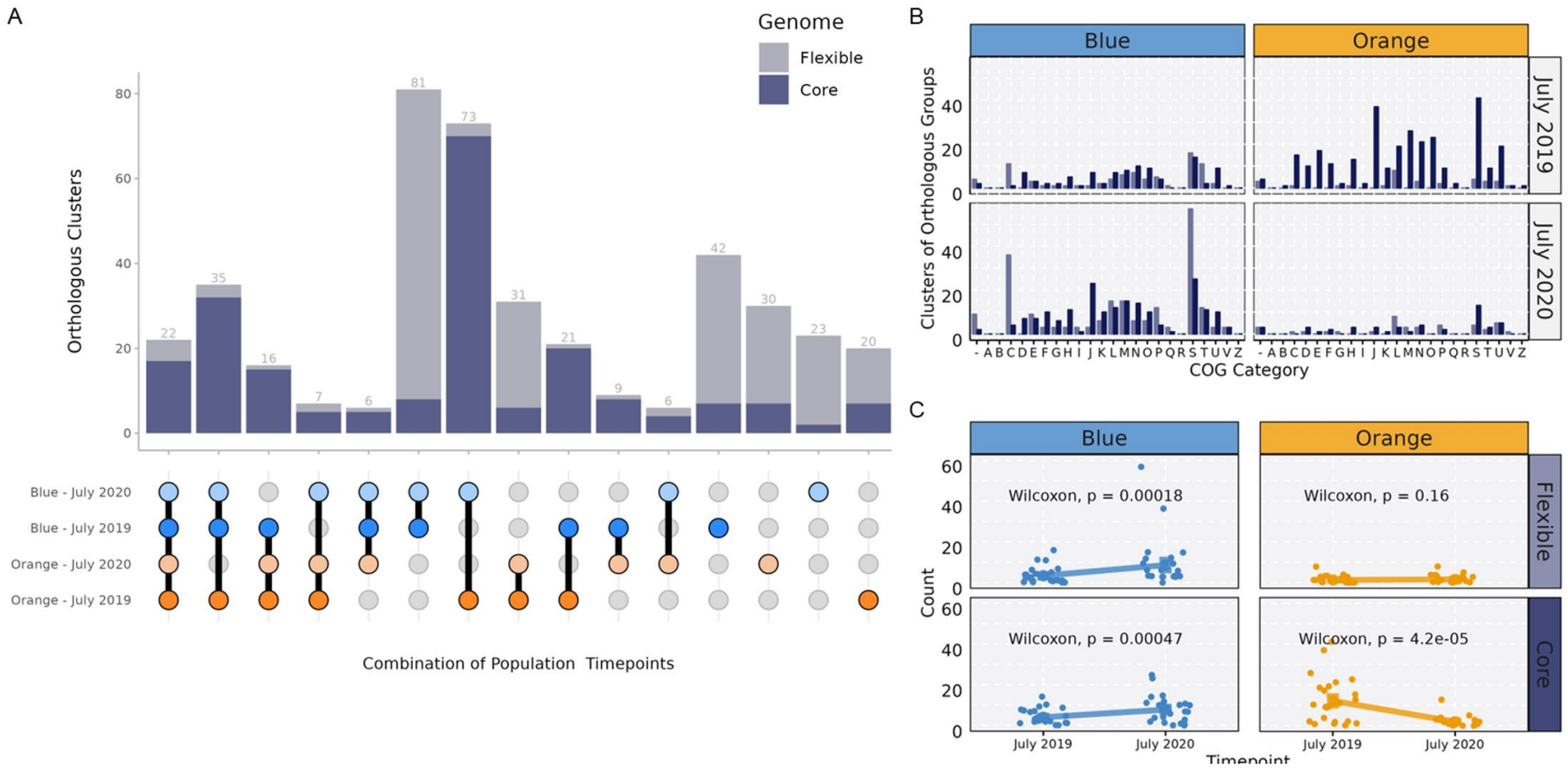

Figure 3.7 Clustering of orthologous genes between swept regions of *UBA11096* sp. Orange and Blue populations at different time points. Upset plot of the predicted proteins from both the flexible- and core-swept regions of the *UBA11096 sp.* cluster 2 Blue and Orange populations shows that most COGs are commonly shared among populations at different time points (A), with 307 COGs shared at least between two populations, while 115 are unique to each set. The highest similarity is between the flexible swept region of the *UBA11096 sp.* cluster 2 Blue populations at different time points and the core-swept region of the Blue population during July 2020 and the Orange population during July 2019. COGs in both populations (B) are distributed primarily in [L] 'Replication, recombination and repair', [M] 'Cell wall/membrane biogenesis', [N] 'Cell motility', and [J] 'Translation' categories (see SUPP. 15). Additionally, a comparison of the number of COGs in the Core Genome shows a contrasting trend between populations (C, see SUPP. 16), in which there is a significant decrease in swept genes observed in UBA11096 sp. cluster 2 Orange core genome while it increases in the Blue population. Statistical analysis is done using the Wilcoxon Signed-Rank test

3.5. Conclusion

In this study, I examined the genetic changes in metabolically limited populations of *Aquificota* bacteria across time. In contrast to the prevailing ecotype theory that posits ecologically homogenous populations, such as chemolithoautotrophs, solely respond to perturbations through genome-wide sweeps (Cohan, 2016), this study demonstrated that metabolically limited bacteria do not necessarily respond to environmental perturbation through genome-wide sweeps. The more clonal populations of *Aquificota* appeared to have undergone a genome-wide sweep prior to the start of the study, as shown by the unchanging intrapopulation diversity across time. Meanwhile, observed changes in intrapopulation diversity and intrapopulation turnover of the dominant genomic background in recombining *Aquificota* populations suggested that recombining populations likely responded to environmental perturbations through gene-specific sweeps. The co-occurrence of apparent gene-specific and genome-wide sweep between closely related populations with similar metabolic limitations raises the question of ecotype theory's applicability in recombining populations. The presence of gene-specific sweep in chemolithoautotrophic *Aquificota* bacteria indicates the potential generalisability of the ecological population theory as a model of bacterial speciation since it explicitly acknowledges the role of gene-specific sweeps in bacterial diversification.

3.6. Acknowledgement

This research was supported by the Marsden Fund, administered by the Royal Society of New Zealand (17-UOW-026). Access to Kuirau Park was granted by the Rotorua City Council.

3.7. References

Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J., Polz, M.F., 2019a. A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. Cell 178, 820-834.e14. https://doi.org/10.1016/j.cell.2019.06.033

Arevalo, P., VanInsberghe, D., Polz, M.F., 2019b. A Reverse Ecology Framework for Bacteria and Archaea, in: Polz, M.F., Rajora, O.P. (Eds.), Population Genomics: Microorganisms, Population Genomics. Springer International Publishing, Cham, pp. 77–96. https://doi.org/10.1007/13836_2018_46

Atwood, K.C., Schneider, L.K., Ryan, F.J., 1951. Periodic Selection in Escherichia Coli. Proc Natl Acad Sci U S A 37, 146–155.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 19, 455–477. https://doi.org/10.1089/cmb.2012.0021

Bao, Y.-J., Shapiro, B.J., Lee, S.W., Ploplis, V.A., Castellino, F.J., 2016. Phenotypic differentiation of Streptococcus pyogenes populations is induced by recombination-driven gene-specific sweeps. Sci Rep 6, 36644. https://doi.org/10.1038/srep36644

Baquero, F., Coque, T.M., Galán, J.C., Martinez, J.L., 2021. The Origin of Niches and Species in the Bacterial World. Frontiers in Microbiology 12.

Barton, N.H., 2001. Adaptation at the edge of a species' range. Integrating ecology and evolution in a spatial context.

Bendall, M.L., Stevens, S.L., Chan, L.-K., Malfatti, S., Schwientek, P., Tremblay, J., Schackwitz, W., Martin, J., Pati, A., Bushnell, B., Froula, J., Kang, D., Tringe, S.G., Bertilsson, S., Moran, M.A., Shade, A., Newton, R.J., McMahon, K.D., Malmstrom, R.R., 2016. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. The ISME Journal 10, 1589–1601. https://doi.org/10.1038/ismej.2015.241

Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., Lichtenstein, L., 2019. Calling Somatic SNVs and Indels with Mutect2.

Bobay, L.-M., Ochman, H., 2018. Factors driving effective population size and pan-genome evolution in bacteria. BMC Evolutionary Biology 18, 153. https://doi.org/10.1186/s12862-018-1272-4

Bobay, L.-M., Ochman, H., 2017. Biological species are universal across Life's domains. Genome Biol Evol 9, 491–501. https://doi.org/10.1093/gbe/evx026

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brazelton, W., Pendleton, H.L., 2021. RNA Extraction from Sterivex Filters.

Buchfink, B., Reuter, K., Drost, H.-G., 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18, 366–368. https://doi.org/10.1038/s41592-021-01101-x

Bushnell, B., 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner.

Cadillo-Quiroz, H., Didelot, X., Held, N.L., Herrera, A., Darling, A., Reno, M.L., Krause, D.J., Whitaker, R.J., 2012. Patterns of Gene Flow Define Species of Thermophilic Archaea. PLOS Biology 10, e1001265. https://doi.org/10.1371/journal.pbio.1001265

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., Parks, D.H., 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics btz848. https://doi.org/10.1093/bioinformatics/btz848

Chen, Y.-J., Leung, P.M., Wood, J.L., Bay, S.K., Hugenholtz, P., Kessler, A.J., Shelley, G., Waite, D.W., Franks, A.E., Cook, P.L.M., Greening, C., 2021. Metabolic flexibility allows bacterial habitat generalists to become dominant in a frequently disturbed ecosystem. ISME J 1–19. https://doi.org/10.1038/s41396-021-00988-w

Cohan, F.M., 2016. Bacterial Speciation: Genetic Sweeps in Bacterial Species. Current Biology 26, R112–R115. https://doi.org/10.1016/j.cub.2015.10.022

Cohan, F.M., 2005. Periodic Selection and Ecological Diversity in Bacteria, in: Nurminsky, D. (Ed.), Selective Sweep, Molecular Biology Intelligence Unit. Springer US, Boston, MA, pp. 78–93. https://doi.org/10.1007/0-387-27651-3_7

Cohan, F.M., 2001. Bacterial Species and Speciation. Syst Biol 50, 513–524. https://doi.org/10.1080/10635150118398

Cohan, F.M., Perry, E.B., 2007. A Systematics for Discovering the Fundamental Units of Bacterial Diversity. Current Biology 17, R373–R386. https://doi.org/10.1016/j.cub.2007.03.032

Coudert, E., Gehant, S., de Castro, E., Pozzato, M., Baratin, D., Neto, T., Sigrist, C.J.A., Redaschi, N., Bridge, A., The UniProt Consortium, 2023. Annotation of biologically relevant ligands in UniProtKB using ChEBI. Bioinformatics 39, btac793. https://doi.org/10.1093/bioinformatics/btac793

Dini-Andreote, F., de Cássia Pereira e Silva, M., Triadó-Margarit, X., Casamayor, E.O., van Elsas, J.D., Salles, J.F., 2014. Dynamics of bacterial community succession in a salt marsh chronosequence: evidences for temporal niche partitioning. ISME J 8, 1989–2001. https://doi.org/10.1038/ismej.2014.54

Dixon, P., 2003. VEGAN, A Package of R Functions for Community Ecology. Journal of Vegetation Science 14, 927–930.

Douglass, A.P., O'Brien, C.E., Offei, B., Coughlan, A.Y., Ortiz-Merino, R.A., Butler, G., Byrne, K.P., Wolfe, K.H., 2019. Coverage-Versus-Length Plots, a Simple Quality Control Step for de Novo Yeast Genome Sequence Assemblies. G3 (Bethesda) 9, 879–887. https://doi.org/10.1534/g3.118.200745

Friedman, J., Alm, E.J., Shapiro, B.J., 2013. Sympatric Speciation: When Is It Possible in Bacteria? PLoS ONE 8, e53539. https://doi.org/10.1371/journal.pone.0053539

Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., Egozcue, J.J., 2017. Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol 8, 2224. https://doi.org/10.3389/fmicb.2017.02224

Goldenberger, D., Perschil, I., Ritzler, M., Altwegg, M., 1995. A simple "universal" DNA extraction procedure using SDS and proteinase K is compatible with direct PCR amplification. Genome Res. 4, 368–370.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., Tiedje, J.M., 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. International Journal of Systematic and Evolutionary Microbiology 57, 81–91. https://doi.org/10.1099/ijs.0.64483-0

Granato, E.T., Kümmerli, R., 2017. The path to re-evolve cooperation is constrained in Pseudomonas aeruginosa. BMC Evol Biol 17, 214. https://doi.org/10.1186/s12862-017-1060-6

Hamilton, T.L., Koonce, E., Howells, A., Havig, J.R., Jewell, T., de la Torre, J.R., Peters, J.W., Boyd, E.S., 2014. Competition for Ammonia Influences the Structure of Chemotrophic Communities in Geothermal Springs. Applied and Environmental Microbiology 80, 653–661. https://doi.org/10.1128/AEM.02577-13

Hanage, W.P., Fraser, C., Spratt, B.G., 2005. Fuzzy species among recombinogenic bacteria. BMC Biology 3, 6. https://doi.org/10.1186/1741-7007-3-6

Hedlund, B.P., Reysenbach, A.-L., Huang, L., Ong, J.C., Liu, Z., Dodsworth, J.A., Ahmed, R., Williams, A.J., Briggs, B.R., Liu, Y., Hou, W., Dong, H., 2015. Isolation of diverse members of the Aquificales from geothermal springs in Tengchong, China. Frontiers in Microbiology 6.

Hermisson, J., Pennings, P.S., 2017. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods in Ecology and Evolution 8, 700–716. https://doi.org/10.1111/2041-210X.12808

Hetzer, A., McDonald, I.R., Morgan, H.W., 2008. Venenivibrio stagnispumantis gen. nov., sp. nov., a thermophilic hydrogen-oxidizing bacterium isolated from Champagne Pool, Waiotapu, New Zealand. International Journal of Systematic and Evolutionary Microbiology 58, 398–403. https://doi.org/10.1099/ijs.0.64842-0

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., Bork, P., 2017. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Molecular Biology and Evolution 34, 2115–2122. https://doi.org/10.1093/molbev/msx148

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119. https://doi.org/10.1186/1471-2105-11-119

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., Aluru, S., 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9, 5114. https://doi.org/10.1038/s41467-018-07641-9

Kim, H.J., Jang, S., 2017. Optimization of a resazurin-based microplate assay for large-scale compound screenings against Klebsiella pneumoniae. 3 Biotech 8, 3. https://doi.org/10.1007/s13205-017-1034-9

Kolde, R., 2018. pheatmap: Pretty Heatmaps.

Kopac, S., Wang, Z., Wiedenbeck, J., Sherry, J., Wu, M., Frederick M. Cohan, 2014. Genomic Heterogeneity and Ecological Speciation within One Subspecies of Bacillus subtilis. Appl. Environ. Microbiol. 80, 4842–4853. https://doi.org/10.1128/AEM.00576-14

Krassowski, M., 2020. ComplexUpset.

Kupczok, A., Neve, H., Huang, K.D., Hoeppner, M.P., Heller, K.J., Franz, C.M.A.P., Dagan, T., 2018. Rates of Mutation and Recombination in Siphoviridae Phage Genome Evolution over Three Decades. Mol Biol Evol 35, 1147–1159. https://doi.org/10.1093/molbev/msy027

Lasky, J.R., 2019. Eco-evolutionary community turnover following environmental change. Evol Appl 12, 1434–1448. https://doi.org/10.1111/eva.12776

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., Pfister, H., 2014. UpSet: Visualization of Intersecting Sets. IEEE Transactions on Visualization and Computer Graphics 20, 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248

Li, Z., Nair, S.K., 2012. Quorum sensing: How bacteria can coordinate activity and synchronize their response to external signals? Protein Sci 21, 1403–1417. https://doi.org/10.1002/pro.2132

Liu, C., Cui, Y., Li, X., Yao, M., 2021. microeco: an R package for data mining in microbial community ecology. FEMS Microbiology Ecology 97, fiaa255. https://doi.org/10.1093/femsec/fiaa255

Loiacono, S.T., Meyer-Dombard, D.R., Havig, J.R., Poret-Peterson, A.T., Hartnett, H.E., Shock, E.L., 2012. Evidence for high-temperature in situ nifH transcription in an alkaline hot

spring of Lower Geyser Basin, Yellowstone National Park. Environmental Microbiology 14, 1272–1283. https://doi.org/10.1111/j.1462-2920.2012.02710.x

Lopes, L.D., Silva, M. de C.P. e, Weisberg, A.J., Davis, E.W., Yan, Q., Varize, C. de S., Wu, C.-F., Chang, J.H., Loper, J.E., Andreote, F.D., 2018. Genome variations between rhizosphere and bulk soil ecotypes of a Pseudomonas koreensis population. Environmental Microbiology 20, 4401–4414. https://doi.org/10.1111/1462-2920.14363

Majewski, J., Cohan, F.M., 1999. Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. Genetics 152, 1459–1474.

Marttinen, P., Hanage, W.P., 2017. Speciation trajectories in recombining bacterial species. PLOS Computational Biology 13, e1005640. https://doi.org/10.1371/journal.pcbi.1005640

McMurdie, P.J., Holmes, S., 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLOS ONE 8, e61217. https://doi.org/10.1371/journal.pone.0061217

Messer, P.W., Petrov, D.A., 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends in Ecology & Evolution 28, 659–669. https://doi.org/10.1016/j.tree.2013.08.003

Miller, S.R., Carvey, D., 2019. Ecological Divergence with Gene Flow in a Thermophilic Cyanobacterium. Microb Ecol 78, 33–41. https://doi.org/10.1007/s00248-018-1267-0

Morris, J.J., Lenski, R.E., Zinser, E.R., 2012. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. mBio 3, e00036-12. https://doi.org/10.1128/mBio.00036-12

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44, D733-745. https://doi.org/10.1093/nar/gkv1189

Ottesen, E.A., Young, C.R., Eppley, J.M., Ryan, J.P., Chavez, F.P., Scholin, C.A., DeLong, E.F., 2013. Pattern and synchrony of gene expression among sympatric marine microbial populations. Proceedings of the National Academy of Sciences 110, E488–E497. https://doi.org/10.1073/pnas.1222099110

Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., Hugenholtz, P., 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically

consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Research 50, D785–D794. https://doi.org/10.1093/nar/gkab776

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25, 1043–1055. https://doi.org/10.1101/gr.186072.114

Pierce, N.T., Irber, L., Reiter, T., Brooks, P., Brown, C.T., 2019. Large-scale sequence comparisons with sourmash. F1000Res 8, 1006. https://doi.org/10.12688/f1000research.19675.1

Polz, M.F., Alm, E.J., Hanage, W.P., 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends in Genetics 29, 170–175. https://doi.org/10.1016/j.tig.2012.12.006

Pompei, S., Bella, E., Weitz, J.S., Grilli, J., Lagomarsino, M.C., 2023. Metacommunity structure preserves genome diversity in the presence of gene-specific selective sweeps under moderate rates of horizontal gene transfer. PLOS Computational Biology 19, e1011532. https://doi.org/10.1371/journal.pcbi.1011532

Power, J.F., Carere, C.R., Lee, C.K., Wakerley, G.L.J., Evans, D.W., Button, M., White, D., Climo, M.D., Hinze, A.M., Morgan, X.C., McDonald, I.R., Cary, S.C., Stott, M.B., 2018. Microbial biogeography of 925 geothermal springs in New Zealand. Nat Commun 9, 2876. https://doi.org/10.1038/s41467-018-05020-y

Power, J.F., Carere, C.R., Welford, H.E., Hudson, D.T., Lee, K.C., Moreau, J.W., Ettema, T.J.G., Reysenbach, A.-L., Lee, C.K., Colman, D.R., Boyd, E.S., Morgan, X.C., McDonald, I.R., Craig Cary, S., Stott, M.B., 2024. A genus in the bacterial phylum Aquificota appears to be endemic to Aotearoa-New Zealand. Nat Commun 15, 179. https://doi.org/10.1038/s41467-023-43960-2

R Core Team, 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reysenbach, A.-L., Huber, R., Stetter, K.O., Ishii, M., Kawasumi, T., Igarashi, Y., Eder, W., L'Haridon, S., Jeanthon, C., 2001. Phylum BI. Aquificae phy. nov., in: Boone, D.R., Castenholz, R.W., Garrity, G.M. (Eds.), Bergey's Manual® of Systematic Bacteriology: Volume One : The Archaea and the Deeply Branching and Phototrophic Bacteria. Springer New York, New York, NY, pp. 359–367. https://doi.org/10.1007/978-0-387-21609-6_18

Rodriguez-R, L.M., Conrad, R.E., Viver, T., Feistel, D.J., Lindner, B.G., Venter, S.N., Orellana, L.H., Amann, R., Rossello-Mora, R., Konstantinidis, K.T., 2023. An ANI gap within bacterial species that advances the definitions of intra-species units. mBio 15, e02696-23. https://doi.org/10.1128/mbio.02696-23

Sandberg, M.E., Schellmann, D., Brunhofer, G., Erker, T., Busygin, I., Leino, R., Vuorela, P.M., Fallarero, A., 2009. Pros and cons of using resazurin staining for quantification of viable

Staphylococcus aureus biofilms in a screening assay. Journal of Microbiological Methods 78, 104–106. https://doi.org/10.1016/j.mimet.2009.04.014

Scott, B.J., 1994. Cyclic activity in the crater lakes of Waimangu hydrothermal system, New Zealand. Geothermics 23, 555–572. https://doi.org/10.1016/0375-6505(94)90019-1

Shapiro, B.J., 2016. How clonal are bacteria over time? Current Opinion in Microbiology, Environmental microbiology * Special Section: Megaviromes 31, 116–123. https://doi.org/10.1016/j.mib.2016.03.013

Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabó, G., Polz, M.F., Alm, E.J., 2012. Population Genomics of Early Events in the Ecological Differentiation of Bacteria. Science 336, 48–51. https://doi.org/10.1126/science.1218198

Shapiro, B.J., Polz, M.F., 2014. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol 22, 235–247. https://doi.org/10.1016/j.tim.2014.02.006

Smith, N.H., Kremer, K., Inwald, J., Dale, J., Driscoll, J.R., Gordon, S.V., van Soolingen, D., Hewinson, R.G., Smith, J.M., 2006. Ecotypes of the Mycobacterium tuberculosis complex. J Theor Biol 239, 220–225. https://doi.org/10.1016/j.jtbi.2005.08.036

Song, Z.-Q., Wang, L., Liang, F., Zhou, Q., Pei, D., Jiang, H., Li, W.-J., 2022. nifH gene expression and diversity in geothermal springs of Tengchong, China. Frontiers in Microbiology 13.

Steinegger, M., Söding, J., 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35, 1026–1028. https://doi.org/10.1038/nbt.3988

Sun, J., Lu, F., Luo, Y., Bie, L., Xu, L., Wang, Y., 2023. OrthoVenn3: an integrated platform for exploring and visualizing orthologous data across genomes. Nucleic Acids Research 51, W397–W403. https://doi.org/10.1093/nar/gkad313

Takeuchi, N., Cordero, O.X., Koonin, E.V., Kaneko, K., 2015. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. BMC Biology 13, 20. https://doi.org/10.1186/s12915-015-0131-7

Utter, D.R., Cavanaugh, C.M., Borisy, G.G., 2022. Genome-Centric Dynamics Shape the Diversity of Oral Bacterial Populations. mBio 13, e02414-22. https://doi.org/10.1128/mbio.02414-22

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S., DePristo, M.A., 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols in Bioinformatics 43, 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Vandemeulebrouck, J., Hurst, A.W., Scott, B.J., 2008. The effects of hydrothermal eruptions and a tectonic earthquake on a cycling crater lake (Inferno Crater Lake, Waimangu, New Zealand). Journal of Volcanology and Geothermal Research, Volcanic lakes and environmental impacts of volcanic fluids 178, 271–275. https://doi.org/10.1016/j.jvolgeores.2008.06.021

Viver, T., Conrad, R.E., Orellana, L.H., Urdiain, M., González-Pastor, J.E., Janet K. Hatt, Amann, R., Antón, J., Konstantinos T. Konstantinidis, Ramon Rosselló-Móra, 2021. Distinct ecotypes within a natural haloarchaeal population enable adaptation to changing environmental conditions without causing population sweeps. The ISME Journal 15, 1178–1191. https://doi.org/10.1038/s41396-020-00842-5

Viver, T., Conrad, R.E., Rodriguez-R, L.M., Ramírez, A.S., Venter, S.N., Rocha-Cárdenas, J., Llabrés, M., Amann, R., Konstantinidis, K.T., Rossello-Mora, R., 2024. Towards estimating the number of strains that make up a natural bacterial population. Nat Commun 15, 544. https://doi.org/10.1038/s41467-023-44622-z

Vos, M., Didelot, X., 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J 3, 199–208. https://doi.org/10.1038/ismej.2008.93

Wang, X., Zhang, Y., Ren, M., Xia, T., Chu, X., Liu, C., Lin, X., Huang, Y., Chen, Z., Yan, A., Luo, H., 2020. Cryptic speciation of a pelagic Roseobacter population varying at a few thousand nucleotide sites. ISME J 14, 3106–3119. https://doi.org/10.1038/s41396-020-00743-7

Ward, L., Taylor, M.W., Power, J.F., Scott, B.J., McDonald, I.R., Stott, M.B., 2017. Microbial community dynamics in Inferno Crater Lake, a thermally fluctuating geothermal spring. ISME J 11, 1158–1167. https://doi.org/10.1038/ismej.2016.193

Whitaker, R.J., Grogan, D.W., Taylor, J.W., 2005. Recombination shapes the natural population structure of the hyperthermophilic archaeon Sulfolobus islandicus. Mol. Biol. Evol. 22, 2354–2361. https://doi.org/10.1093/molbev/msi233

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the Tidyverse. Journal of Open Source Software 4, 1686. https://doi.org/10.21105/joss.01686

Wiedenbeck, J., Cohan, F.M., 2011. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35, 957–976. https://doi.org/10.1111/j.1574-6976.2011.00292.x

Wilson, B.A., Pennings, P.S., Petrov, D.A., 2017. Soft Selective Sweeps in Evolutionary Rescue. Genetics 205, 1573–1586. https://doi.org/10.1534/genetics.116.191478

CHAPTER 4: Homologous recombination drives horizontal gene flow in *Aquificota* populations

4.1. Abstract

Delineating bacterial populations through discontinuities in gene flow is predicated on the assumption that homologous recombination drives horizontal gene flow. However, other mechanisms of horizontal gene flow, such as mobile genetic elements (MGEs, viral sequences and integrative and conjugative elements), have been attributed to bacteria's genetic promiscuity. One example is *UBA11096* sp., where regulation of gene flow has been linked with prevalent MGEs and self-targeting CRISPR-Cas systems. Therefore, validating the ecological implication and population structure of *UBA11096* sp. requires elucidation of the role of MGEs in the recent horizontal gene flow of recombining *Aquificota* populations. This chapter aims to identify the co-occurrence of MGEs with recombining populations of *UBA11096* sp. Analysis of the abundance of MGEs with recombining populations of *UBA11096* sp. cluster 2 and *Venenivibrio* showed no significant correlation between MGEs and recently recombining populations of *Aquificota*. Furthermore, spatial and functional genomic analyses showed minimal co-localisation of MGEs with identified horizontally swept regions, indicating that recent horizontal gene flow in *UBA11096* sp. cluster 2 was primarily driven by homologous recombination. Lastly, I uncovered CRISPR arrays and Cas proteins shared between recombining populations of *UBA11096* sp. cluster 2, supporting the theory that recombining populations act as genetic reservoirs for each other. This work demonstrated that horizontal gene flow in the *UBA11096* sp. cluster 2 was primarily driven by homologous recombination and, thus, validated its population structure and the ecological implications previously identified through reverse ecology.

4.2. Introduction

Genomic analysis of natural bacterial populations has shown that horizontal gene flow plays a large role in maintaining bacterial cohesion in natural bacterial communities (Arnold et al., 2022; Cordero and Polz, 2014; Fraser et al., 2007; Hanage et al., 2005; Polz et al., 2013; Shapiro et al., 2012). A high degree of horizontal gene flow in bacteria results in gene-specific sweeps in which adaptive genes are preferentially shared between horizontally linked organisms. Selection for the shared adaptive genes leads to the formation of an ecologically cohesive population (Shapiro et al., 2016; Shapiro and Polz, 2015; Takeuchi et al., 2015). Therefore, novel bacterial speciation models have proposed delineating ecologically cohesive populations or species through discontinuities in recent horizontal gene flow (Arevalo et al., 2019a; Bobay et al., 2018; Marttinen and Hanage, 2017). Since population gene flows carry adaptive genes, identifying horizontally swept genes within a population can also aid in describing the selective pressure experienced by the population and can elucidate the population's ecological niche (Arevalo et al., 2019b; VanInsberghe et al., 2020). This method of ecological analysis called reverse ecology, requires no prior notion of species identification and ecological differences (Arevalo et al., 2019b), thus making it suited for studying novel populations. However, the reverse ecology approach and gene flow-dependent population structure delineation are predicated on the assumption that horizontal gene flow is primarily driven by homologous recombination (Arevalo et al., 2019a; Fraser et al., 2007; Shapiro et al., 2012). Homologous recombination through natural transformation is suggested to be more efficient in maintaining population boundaries as they can more frequently share xenologous alleles without the need to incorporate novel genes, which can lower fitness costs (Mell and Redfield, 2014; Novick and Doolittle, 2019; von Wintersdorff et al., 2016). Furthermore, the effect of non-homologous recombination facilitated horizontal gene flow, such as through mobile genetic elements (MGEs), can blur the delineation of boundaries of ecologically cohesive populations such as linking strictly non-recombining population into a single, albeit transient, ecological trait like antibiotic resistance (Arevalo et al., 2019b; VanInsberghe et al., 2020). Long-distance transfer of genes, such as the spread of antibiotic resistance genes, alters the ecological relevance of a recipient organism, leading to its apparent distinction from other closely related organisms (Doolittle and Papke, 2006). Subsequently, mobile genetic elements are considered to be efficient in introducing novel genes and, thus, facilitate the adaptation of novel ecological niches (Novick and Doolittle, 2019).

The prevalence and activity of non-homologous recombining MGEs between species varies significantly, and not all recombining bacterial populations utilise MGEs (Durrant et al.,

2020; von Wintersdorff et al., 2016). Previous studies on mobile genetic elements in ecological populations defined through gene flow discontinuity show contrasting results—ranging from no significantly observed activity in the population structure of *Vibrio* species (Arevalo et al., 2019a) to plasmid-mediated high-frequency transfer in diverging species of *Sulfolobus islandicus* (Sanchez-Nieves et al., 2023), and to tycheposons, novel gene transfer elements containing genes and the necessary integrase, replication, and recombinase components, driving diversity and adaption in *Prochlorococcus* (Hackl et al., 2023). Furthermore, even with the increasing observation of homologous recombination's role in speciation (Arevalo et al., 2019a; Cordero and Polz, 2014; Fraser et al., 2007; González-Torres et al., 2019; Vos and Didelot, 2009a), studies often still attributed ecological and evolutionary significance of a high degree of horizontal gene flow solely to MGEs due to their independent regulation and long-distance dissemination of ecologically important genes (Acman et al., 2022; Durrant et al., 2020; Johansson et al., 2023; Sharma et al., 2022; Sobecky and Hazen, 2009; Vogan and Higgs, 2011). Therefore, delineating recombining populations and understanding their ecological relevance requires careful examination of the role of non-homologous recombination in horizontal gene flow (VanInsberghe et al., 2020).

In Chapters 2 and 3, I examine members of *Aquificota* phyla as a model of horizontal gene flow-driven bacterial speciation in edge-case populations. In Chapter 2, population clustering analysis using horizontal gene flow discontinuities (Arevalo et al., 2019a) categorised *Aquificota* phyla members into recombining and non-recombining populations (Figure 3.4). The longitudinal study in Chapter 3 shows a shifting population structure in the recombining *UBA11096 sp*. cluster 2 Blue and Orange populations that suggests the potential of horizontal gene flow-driven bacterial speciation in recombining *Aquificota* bacteria (Figure 3.5). Additionally, reverse ecology analysis of the recombining populations (Chapter 3) suggested that differences in energy processing and nitrogen fixation functions potentially drive population-defining adaptation. My previous Chapters suggest that horizontal gene flow-driven bacterial speciation mechanisms mediate the formation of ecologically cohesive populations (ecological populations) in *Aquificota*. However, the validity of these results is uncertain as a recent study has observed a high abundance of MGEs correlated with a high degree of horizontal gene flow (Bowers et al., 2022). It is primarily assumed that homologous recombination is the major player in horizontal gene flow-driven bacterial speciation (Arevalo et.al., 2019a, 2019b). Therefore, confirming that homologous recombination has a significant role in the formation of *Aquificota* ecological populations requires the elucidation of the role of non-homologous recombination in horizontal gene flow in *Aquificota* bacteria.

In this work, I utilised previously acquired genomes from naturally co-occurring *Aquificota* populations to examine the contribution of mobile genetic elements to the overall horizontal gene flow and the bacterial speciation in *Aquificota* bacteria. I identified mobile genetic elements (i.e. conjugation- and virus-driven elements) within the various *Aquificota* populations and determined their correlation with the recombining populations. Genome mapping and functional annotation of genes were also carried out to identify spatial and functional overlap between PopCOGenT-identified horizontally swept regions and mobile genetic elements. MGE evolution and regulation are also often linked to bacterial host defence systems (Koonin et al., 2020; Koonin and Makarova, 2017; Rocha and Bikard, 2022). With previous studies showing high abundance and diversity of CRISPR spacer sequences in *UBA11096* sp. (Bowers et al., 2022), I also examine the differences in CRISPR-Cas system's abundance to determine its role in the total horizontal gene flow of *Aquificota* bacteria.

4.3. Methodology

4.3.1. Genome Acquisition

From the previous chapters, Kuirau Park Sites 1 and 2 genomes across five time points were utilised to understand genes and genomic properties that affect horizontal gene flow in recombining populations. From the sequenced genomes in previous chapters, 121 genomes of *Aquificales* (i.e. *UBA11096* sp., *Hydrogenobacter thermophilus*) and 39 *Hydrogenothermales* (*Sulfurihydrigonibium* sp. and *Venenivibrio* sp.) were selected and used for downstream analysis based on the <2% contamination rate detected by CheckM (Parks et al., 2015). The genomes were grouped into taxonomic clusters based on the populations identified previously by PopCOGenT analysis (Arevalo et al., 2019a).

4.3.2. Swept Region Annotation

4.3.2.1. Swept Gene Analysis: Flexible Genome

Previously identified recombining clusters of *UBA11096* sp. cluster 2 using PopCOGenT analysis were used for swept gene analysis. Swept gene analysis has been limited to genomes isolated from July 2019 and July 2020. Swept genes from the flexible genome analysis were done using the modified pipeline of PopCOGenT (Arevalo et al., 2019a). A gene is identified as a part of the flexible genome when it is swept within a population or PopCOGenT subcluster (e.g., *UBA11096* sp. cluster Blue or Orange population) but is not found throughout the other

populations within its cluster. The PopCOGenT pipeline identifies the clusters of genes that swept a population. A custom python script fetches their sequences from a representative genome used for downstream analysis. The sequences were translated into predicted proteins using Prodigal ver. 4.8.4 (Hyatt et al., 2010), which were afterwards clustered per population into clusters of orthologous genes (COGs) using mmseqs2 (ver. 15.6f452) (Steinegger and Söding, 2017). Functional analysis of each cluster was done using the Diamond ver. 2.1.8.162 BLASTp module (Buchfink et al., 2021) against the Uniprot database 2023_1 release (Coudert et al., 2023) and mapped to COG categories using eggNOG mapper ver. 5 (Huerta-Cepas et al., 2017).

4.3.2.2. Swept Gene Analysis: Core Genome

Similar to swept genes identified in the flexible genome, identification of swept genes in the core genome is limited to genomes in the recombining *UBA11096* sp. cluster 2 Blue and Orange populations from July 2019 and July 2020. The swept genes from the core genome were analysed using the modified PopCOGenT Core Genome pipeline in a Docker container (glbcabria/popcogent-revised:1.3.2). The PopCOGenT Core Genome pipeline identifies a swept gene in the core genome through identification of low-diversity loci in the alignment of all members of the population but not in other populations. Open reading frame and protein prediction were identified from swept regions using Prodigal ver. 4.8.4. The predicted genes were annotated using the Diamond BLASTp module against the Uniprot database release 2023_1, and COG categories were identified using the eggNOG mapper (ver. eggNOG5 database).

4.3.3. Mobile Genetic Element Analysis

Viral sequences from all *Aquificota* populations were quantified and identified using the Virsorter ver 2.2.4 (Guo et al., 2021), while integrative and conjugative elements, along with their signature proteins, were predicted using the ICEscreen pipeline ver. 1.3.2 (Lao et al., 2022). Prodigal predicted ORFs and proteins from the viral sequences and mobilisable elements. The predicted proteins from viral sequences were clustered per population using mmseqs2 (ver. 15.6f452). COGs were subsequently annotated using the Diamond ver. 2.1.8.162 BLASTp module against the UniProt Database 2023_1 release and categorised through eggNOG mapper (ver. 5).

The mobileOG-db (ver. Beatrix-1.6 database) was also used to predict mobile genetic elements within the reference genomes (Brown et al., 2022). Alien Hunter (ver. 1.7) was also used to identify putative HGT regions or genomic islands through interpolated variable order

motifs (Vernikos and Parkhill, 2006). The predicted regions were visualised spatially on the reference genome using the ProkSee pipeline (accessed Jan 2024) (Grant et al., 2023).

4.3.4. CRISPR-Cas Analysis

CRISPR spacer arrays and CAS proteins were identified from each *Aquificota* population using the CRISPRCasFinder pipeline ver. 4.2.20 (Abby et al., 2014; Couvin et al., 2018; Grissa et al., 2007). Quantifying CRISPR targets was done using a custom script utilising BLASTn of predicted CRISPR spacer sequences against the *Aquificota* genomes with '--max-target-sequences' and e-value set at 25 and $1x10^{-5}$, respectively. Self-targeting CRISPRs are identified as sequences with spacer query and subjects belonging to the same genome. Meanwhile, a 'within-population' targeting CRISPR spacers are query sequences belonging to the same Aquificota population as the subject genome, while 'between-population' is counted when the spacer matches other populations. Identification of shared CRISPR spacers follows the same logic but uses predicted CRISPR spacers from other genomes as BLASTn subjects.

The CRISPR spacers and CAS proteins were also identified from the horizontally swept regions (core and flexible genome) and the reference genomes using CRISPRCasFinder. Predicted CRISPR motifs and CAS proteins in the reference genomes were also visualised using the ProkSee Pipeline (accessed Jan 2024).

4.3.5. Statistical Analysis and Visualisation

Statistical analysis of the abundance of mobile genetic elements with *Aquificota* populations was computed and visualised using the ggpubr ver. 0.6.0 (Kassambara, 2023), and ggplot2 package ver. 3.4.4 (Wickham, 2016) in R ver. 4.3.1 (R Core Team, 2022). Comparisons of viral sequence and integrative and conjugative signature protein abundances between populations were computed using Wilcoxon ranked sum and signed rank tests. Comparison of differences in COG category distribution was done through ANOVA and Tukey Post Hoc Test using the base R package.

Clustering of unique orthologous groups between the swept regions and mobile genetic elements was performed using the OrthoVenn pipeline ver. 3 (Sun et al., 2023). Non-clustered predicted proteins from the three groups (flexible, core, and MGE-predicted proteins) were clustered using the OrthoMCL algorithm with an e-value of $1x10^{-2}$ and an inflation factor of 1.50. The cluster proteins were visualised using a Venn diagram generated by ComplexUpset ver. 1.3.5 (Krassowski, 2020; Lex et al., 2014) and the ggplot2 package in R ver. 4.3.1.

4.4. Results and Discussion

4.4.1. A highly recombining population does not necessarily have a large number of mobile genetic elements.

The presence of abundant non-homologous recombination in bacteria challenges the primary assumption in studying the ecological relevance and population structure of recombining populations that horizontal gene flow is driven mainly by homologous recombination (Arevalo et al., 2019b; VanInsberghe et al., 2020). Previous studies on *UBA11096* sp. from natural populations have been observed to have a high degree of horizontal gene flow, as shown by linkage disequilibrium studies (Bowers et al., 2022) and inference of recent gene flow through PopCOGenT analysis (Chapters 1 and 2), yet also exhibited a large number of MGEs (Bowers et al., 2022). It raises the question of whether MGEs played a role in forming a recombining population structure in *Aquificota* populations. To examine whether the previously observed abundance of MGEs in *UBA11096* sp. is unique to that population or is a trait of recombining *Aquificota* populations, I analysed the prevalence of MGEs, specifically viral and integrative and conjugative signature proteins, and its correlation with clonal and recombining Aquificota populations previously acquired in Chapters 2 and 3. Viral sequences were recovered using Virsorter (Guo et al., 2021), while the integrative and conjugative signature proteins and mobilisable elements were identified using ICEscreen and the MobileOG database pipelines (Lao et al., 2022). A comparison of the number of viral sequences per genome shows no significant difference between recombining and non-recombining populations (Figure 4.1A). As only one complete mobile genetic element was found across all *Aquificota* genomes studied (belonging to a genome from *UBA11096 sp.* Cluster 2 Orange population), the integrative and conjugative properties between genomes were assessed by comparing integrative and conjugative element signature proteins (ICE protein) abundances between populations. Similar to viral sequences, comparing ICE proteins between recombining and non-recombining populations yields no significant differences (Figure 4.1B), highlighting the lack of correlation between the abundance of MGEs to a population's ability to form recombining populations.

Furthermore, individual comparisons between recombining *Aquificota* populations show significant differences in the prevalence of viral sequences and ICE proteins (Figure 4.1C-D). For example, recombining population *UBA11096* sp. cluster 2 Orange was observed to have the highest amount of both viral sequences and ICE proteins. However, recombining population *UBA11096 sp.* cluster 2 Blue, its closely related taxa, has significantly fewer viral sequences and ICE proteins. Similar observations can be observed within clonal populations. The discrepancy in MGE abundance within the same clonal or recombining population is influenced by a high abundance of genomes carrying multiple viral sequences, such as in *Sulfurihydrogenibium sp.*

and *UBA11096 sp.* cluster 2 Orange (see SUPP. 17 - Figure 8.5). The abundance of MGEs in a few genomes highlights the specificity of MGEs to specific hosts, as they are often subject to host-parasite dynamics (Durrant et al., 2020; Iranzo et al., 2016) and, thus, a unique property of a population.

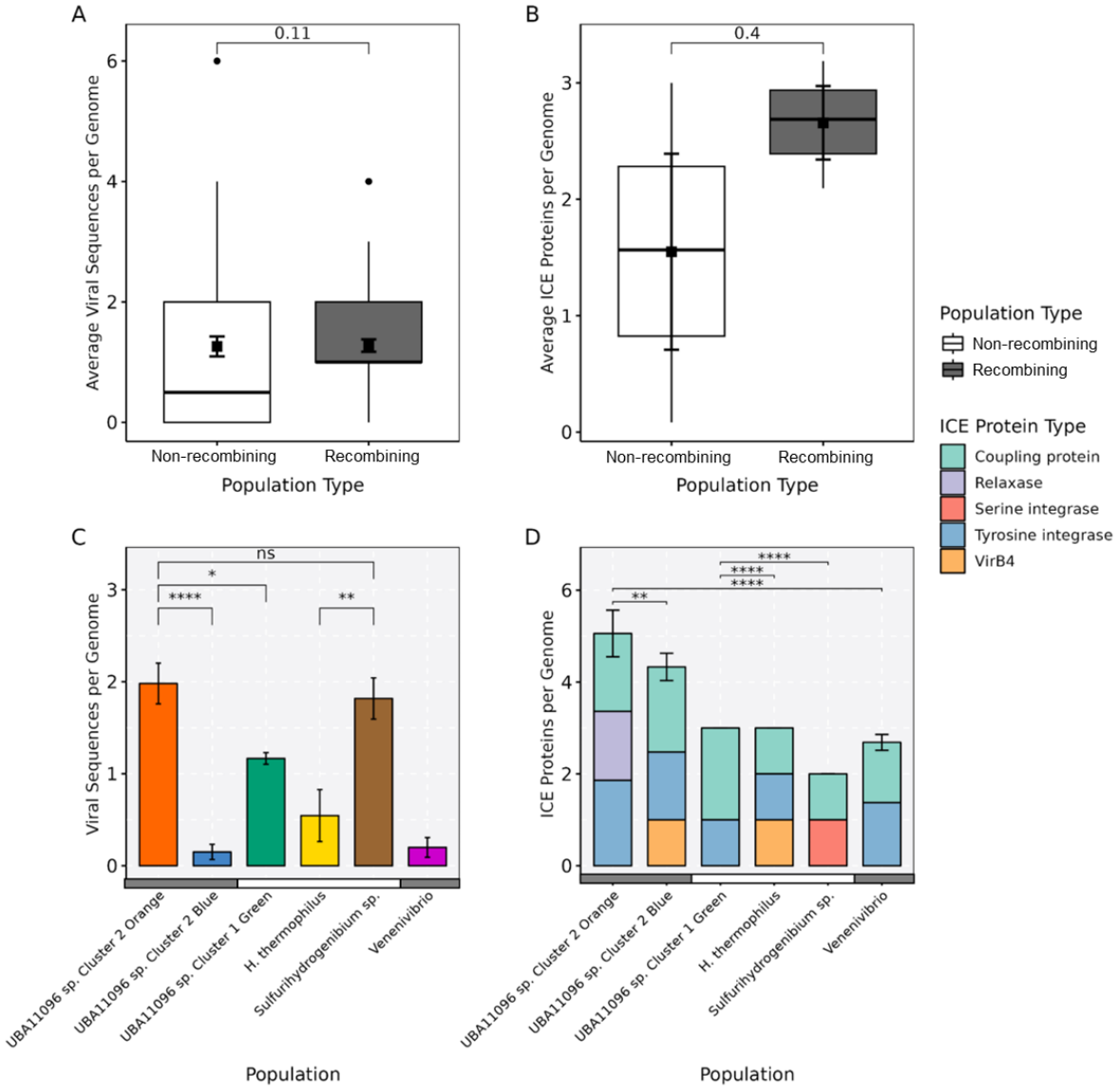

Figure 4.1 Prevalence of viral sequences and integrative and conjugative proteins in different Aquificota Populations. (A)Comparison of the average mobile genetic elements, such as viral sequences (A) and integrative and conjugative signature proteins (B) per genome in the recombining and non-recombining populations, show no significant difference between the population types. Pairwise comparisons of populations show significant differences in the number of viral sequences (C) and predicted integrative and conjugative proteins (D), even within the same population types. Values are means (bar and squares) ± standard error. Statistical analysis by Wilcoxon rank sum tests (a-d), *p<0.05, **p<0.01, ****p<0.0001. Statistical significance is visualised in selected comparisons.

4.4.2. Mobile genetic elements do not spatially or functionally overlap with horizontally swept genes in *Aquificota*

The high abundance of MGEs in the recombining *UBA11096* sp. cluster 2 Orange populations raises the question of whether there is an influence or association between its MGEs and the regions that were swept. To determine the role of MGEs within the recombining populations, I investigated if there is a significant physical overlap between the swept regions in the core and flexible genome identified through PopCOGenT analysis and those of the mobile genetic elements. Unlike its typical definition as used with pangenomes, the core genomes identified by PopCOGenT are defined as local collinear blocks or regions aligned across the whole cluster (i.e. *UBA11096* sp. cluster 2 Blue and Orange population) but with low-diversity areas only for one population. Meanwhile, the flexible genome consists of genes swept within the population but have not swept to other populations within its cluster (i.e. *UBA11096* sp. cluster 2 Blue population but not the Orange population).

The PopCOGenT-identified swept regions in the flexible and core genome of the population were mapped into the reference genome for each population from different time points using the Proksee platform (Figure 4.2). The July 2019 and July 2020 time points were used to compare the overlaps between swept regions and MGEs, as they were previously identified to be where *UBA11096* sp. cluster 2 is highly recombining. The distribution of horizontal swept regions differs for each population per time point since the analysis of the swept regions is limited to co-habiting bacteria within the population per given time point. Within the reference genomes, viral sequences and mobile genetic elements were identified using Virsorter and MobileOG-Db, respectively. The mapped viral sequences and mobile genetic elements have minimal overlap with swept regions identified by PopCOGenT (Figure 4.2), thus supporting the fact that MGEs are not associated with the horizontal gene flow identified by PopCOGenT.

Aside from MGEs, genome islands are seen as horizontally derived regions and are often marked by variations in gene content signatures (e.g., GC, dinucleotides, or codon frequencies) (Carraro et al., 2017). Genomic islands have been associated with pathogenicity islands and mobilisable gene content (Carraro et al., 2017; da Silva Filho et al., 2018). I identified and mapped genomic islands to the reference genomes using interpolated variable order motif prediction software Alien Hunter to examine whether genomic islands also have a role in the horizontal gene flow in Aquificota populations. Interestingly, the identified swept regions also do not fully overlap with the genomic islands predicted by Alien Hunter (Figure 4.2 – first and second inner rings). Many identified swept regions do not appear to be located in genomic islands, suggesting minimal discrepancies between the gene content signatures of donor and recipient genomes of

the identified swept genes, which are assumed to come from closely related and co-habiting organisms (Arevalo et al., 2019b; Bobay and Ochman, 2017; Hanage et al., 2005; Konstantinidis and DeLong, 2008). An additional interesting observation from the mapped genomes is the presence of several CRISPR-Cas system components (Figure 4.2 - outer rings) with the horizontally swept regions of the *UBA11096* sp. cluster 2 populations, indicating a potential for a shared immune defence system within the population.

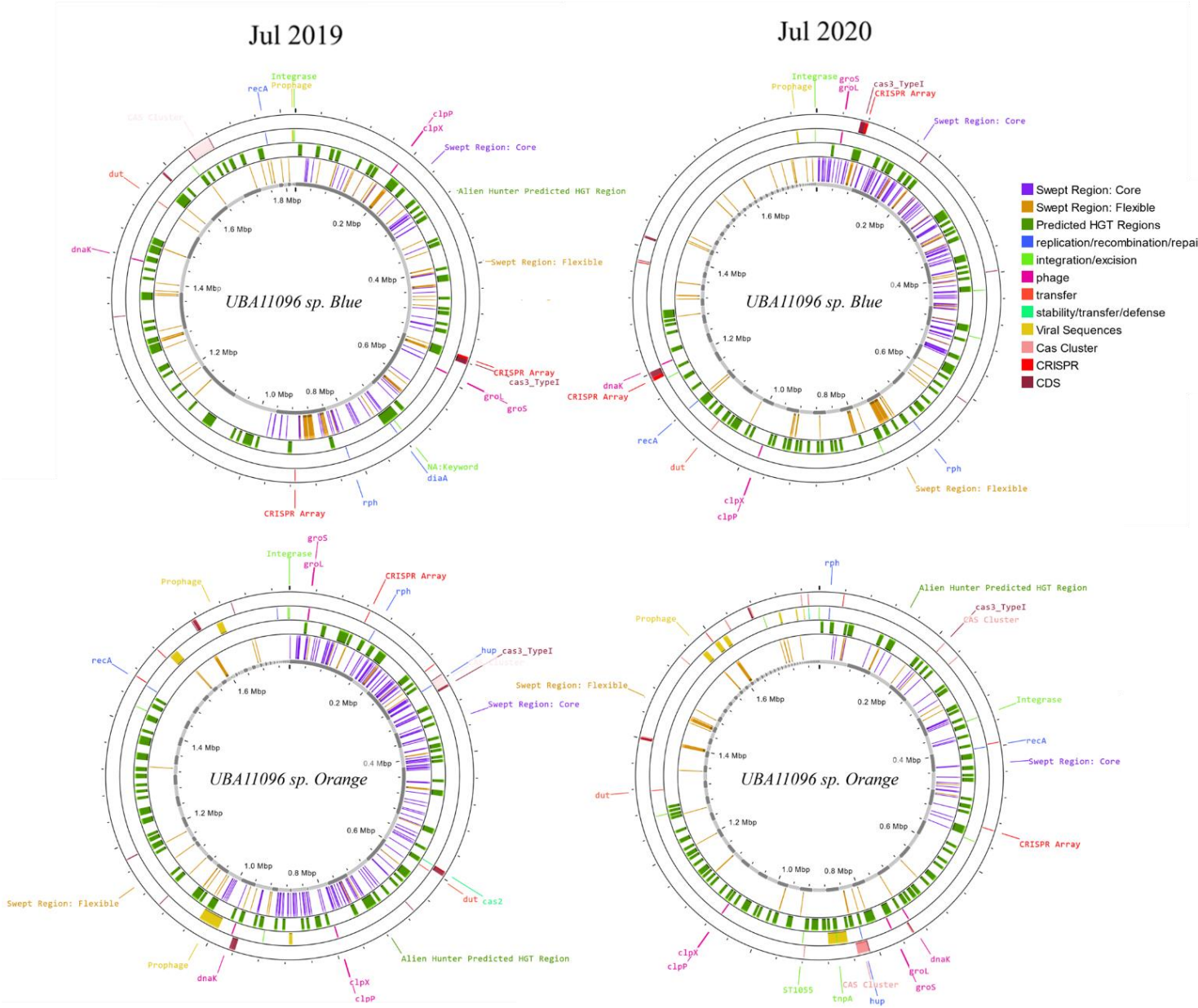

Figure 4.2 Distribution of horizontally swept regions and mobile genetic elements in recombining bacterial genomes. The density of PopCOGenT-identified horizontally swept regions (innermost ring: purple and yellow) in the representative genomes of each recombining population varies between time points. The identified swept regions do not co-localise with the predicted genomic islands by Alien Hunter (second ring: green) and detected mobile genetic elements (third ring). The CRISPR spacer arrays (fourth/outer ring) also do not co-localise with any swept regions, but several Cas3 proteins have spatial overlap with the identified swept regions.

Besides the spatial mapping of mobile genetic elements with the swept regions, I also compare the overlap in function of the predicted genes in the identified horizontally swept regions and the MGEs composed mostly of viral sequences. Clustering of the predicted proteins using the Orthovenn pipeline shows minimal overlap in the clusters of orthologous genes between the viral sequences and the horizontally swept regions (Figure 4.3A). Most COGs between MGEs and horizontally swept regions in the flexible and core genomes are distinct, suggesting that the functions are unique for each set. To validate this, I compared the distribution of the predicted protein across different functional categories (Figure 4.3B) using mmseqs clustering followed by EGGNOG mapping. The functions seen in the horizontally swept regions appear to be distributed differently than those in the viral sequence. The cluster of orthologous genes in the viral sequences was primarily found in replication, recombination and repair categories (L Category)—informational genes needed in the viral life cycle (Mahmoudabadi and Phillips, 2018). Additionally, viral COGs were found to be abundant also in energy production (C Category) and cell wall/membrane function (M Category), with few other COGs distributed across translation (J), transcription (K), and transport and metabolism functions of amino acid (E), nucleotide (F), coenzyme (H). These genes could be part of viral auxiliary metabolic genes that hijack host metabolic machinery or viral competition and survival mechanisms (de la Higuera and Lázaro, 2022; Luo et al., 2022).

Meanwhile, COGs from the identified horizontally swept regions from both core and flexible genomes are dispersed widely across all COG categories. They are also more abundant than the COGs from viral sequences. The COGs from the swept regions are found to be most abundant in the translation functional (J) category and closely followed by functional categories for replication (L), cell wall/membrane (M), motility(N), and protein turnover (O) functions (Figure 4.3B). These functional categories are dominated mainly by COGs found in the swept region in the core genome. COGs from the swept regions in the flexible genome are fewer but abundant in similar functional categories, especially in the replication (L) category. Also, COGs in the energy production (C) and signalling (T) functions are relatively abundant in the swept regions of the flexible genome. Nevertheless, the difference in distribution and abundance of the COGs of the horizontally swept regions and MGEs, coupled with their COG clustering and minimal overlap in genomic mapping, further support the lack of association of MGE with the recombining properties of the *UBA11096* sp. cluster 2 populations. This result implies that the population structure of recombining populations and recent gene flow within *UBA11096 sp.* is primarily driven by homologous recombination.

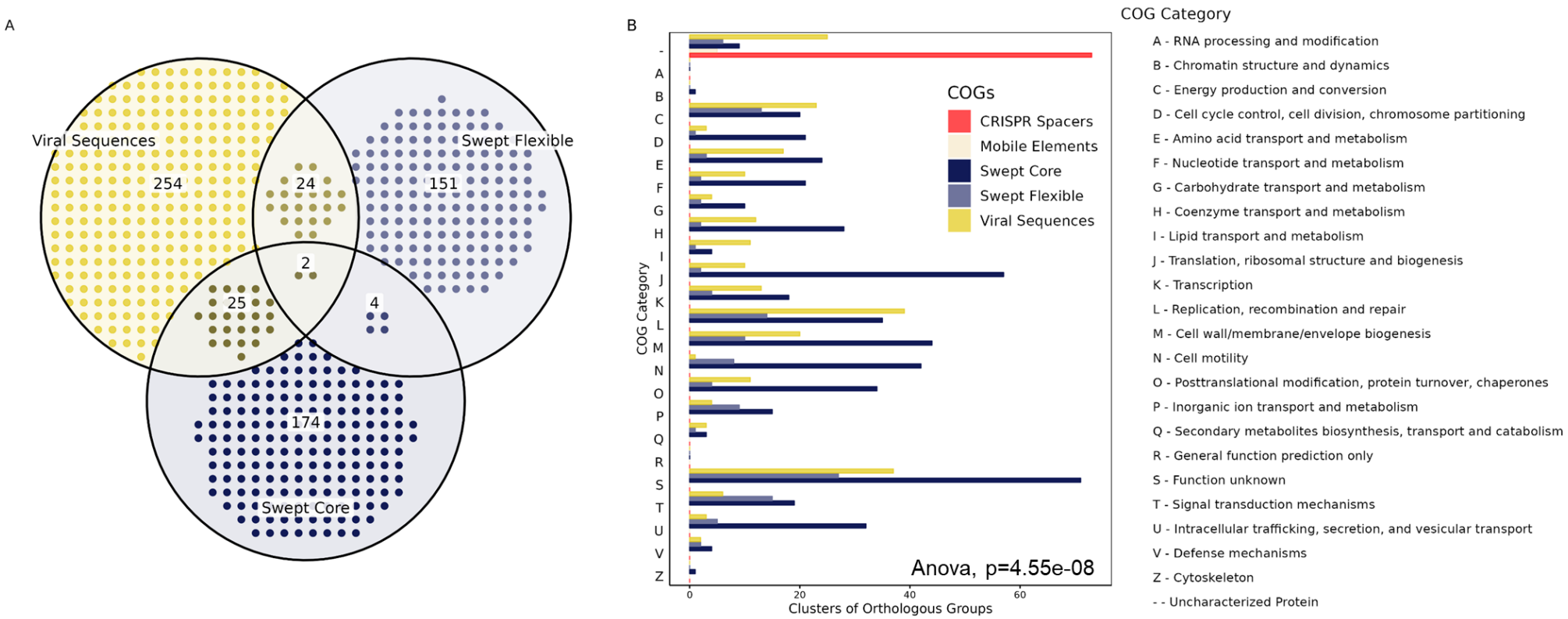

Figure 4.3 Comparison of the functional annotated COGs between the swept regions and mobile genetic elements. Clustering of the predicted proteins from the swept regions and viral sequences of recombining UBA11096 sp. Cluster 2 Blue and Orange Populations (A) shows minimal overlap in the identified clusters of orthologous genes. The identified COGs from the horizontally swept regions (flexible and core) and viral sequences are distributed across different COG categories (B), primarily with translation, post-translational [J], cell wall biogenesis [M], motility [N] and unknown functional categories [S]. The single complete mobile element identified has six predicted proteins with no identified functions. BLASTX-like identification of CRISPR spacer arrays yielded no matches. Statistical analysis through ANOVA shows significant differences in the distribution of COGs between CRISPR, swept regions, mobile elements and viral sequences (see SUPP. 19 Table 8.1). Tukey's Post Hoc Test shows a significant difference between COGs from swept regions against viral sequences, mobile elements and CRISPR spacers (see SUPP. 19 Table 8.2).

4.4.3. CRISPR-Cas systems are shared between recombining populations

Bacterial host defence systems are often entangled in an evolutionary race with mobile genetic elements for survival (Koonin et al., 2020). These defence mechanisms, such as the CRISPR-Cas system, are utilised to control and limit the insertion of MGEs into bacterial genomes (Iranzo et al., 2016; Koonin et al., 2020; Koonin and Makarova, 2017). A previous study shows a high abundance and diversity of CRISPR spacer sequences in *UBA11096 sp.* (identified as *Hydrogenobacter sp.*) (Bowers et al., 2022). These spacer sequences are composed of mostly lineage-specific spacer sequences and, thus, could be unique to each population. Since there are significant differences in the abundance of MGEs between recombining populations of the *UBA11096* sp. cluster 2 populations, I investigated whether there are differences in the CRISPR Cas systems between populations, which might have an effect on the total horizontal gene flow within and between populations. BLASTn analysis of the identified CRISPR spacer sequences against the nucleotide sequences of *Aquificota* genomes' predicted proteins (Figure 4.4A) shows that most genomes have CRISPR spacer sequences that target other genomes from within their own population and genomes from other *Aquificota* populations—thus suggesting these CRISPR spacers provide their host genome protection from MGEs prevalent within the community. Similar to a previous study (Bowers et al., 2022), a majority of *Aquificota* genomes also have CRISPR spacers that target proteins from their own genome. Self-targeting CRISPR sequences are often associated with detrimental autoimmunity (Wimmer and Beisel, 2020) but can also be advantageous in prophage regulation (Edgar and Qimron, 2010; Vercoe et al., 2013) and suggested for other functions such as transcription regulation (Wimmer and Beisel, 2020). Nevertheless, few significant differences exist among the *Aquificota* populations on their CRISPR sequence targets (Figure 4.4A). It should also be noted that none of the nine predicted CRISPR spacers in the *Hydrogenobacter thermophilus* population have been found to target any other *Aquificota* genomes or their own genomes, even with their relatively high MGE count. Lastly, the significantly lower percentile of the genomes in *UBA11096 sp.* cluster 2 Blue population with 'within population' CRISPR spacers targeting could be due to the low abundance of MGEs within the population.

Aside from CRISPR targets, I also compare the CRISPR sequences that are shared between genomes to evaluate whether populations shares adaptive immunity against similar pathogens or act individually. A comparison of shared CRISPR sequences within different *Aquificota* populations (Figure 4.4B) shows no significant differences in the ability to confer adaptive immunity within each population between non-recombining and recombining *Aquificota*. However, the sharing of CRISPR sequences across population boundaries appears to

mainly occur between recombining populations, specifically *UBA11096* sp. Blue and Orange populations (Figure 4.4B, see SUPP. 18 Figure 8.6). The close relationship between the recombining *UBA11096* sp. (ANI ~92% similarity) suggests that the two populations may respond to similar pathogenic threats.

The shared CRISPR sequences between the two recombining populations also support the theory that recombining populations can act as genetic reservoirs to each other (Baumdicker et al., 2012; Kupczok et al., 2018), not only for ecologically functional genes in their swept regions (Chapter 3) but also for their immune or defence systems (Koonin et al., 2020). The idea is that defence against common threats between recombining populations are shared between similar genomes and populations—so that if one or part of a population lost their CRISPR sequences, recombination would allow these sequences to recover their immunity from the rest of the population. The ability to regain adaptive immunity is important as host defence systems also undergo selective pressure as they carry significant fitness costs brought by the disadvantageous effects of autoimmunity (Iranzo et al., 2017; Puigbò et al., 2017; Vale et al., 2015). Defence systems can be easily overwhelmed by high mutation rates in MGEs, resulting in their loss from the host genome (Chevallereau et al., 2019; Koonin et al., 2020) and thought to only persist within a genome through constant mobility within and between genomes through HGT (Rocha and Bikard, 2022), primarily piggybacking with MGEs (Koonin et al., 2020). Previously, bacterial defence systems have been observed to hitchhike directly via the pathogenic viral sequences or mobile genetic elements, allowing defensive elements to spread along with the pathogen (Koonin et al., 2020). However, our BLAST results and the observation of Cas proteins within identified swept regions of *UBA11096* sp. populations (Figure 4.2) suggest that the CRISPR-Cas system can also be shared through homologous recombination.

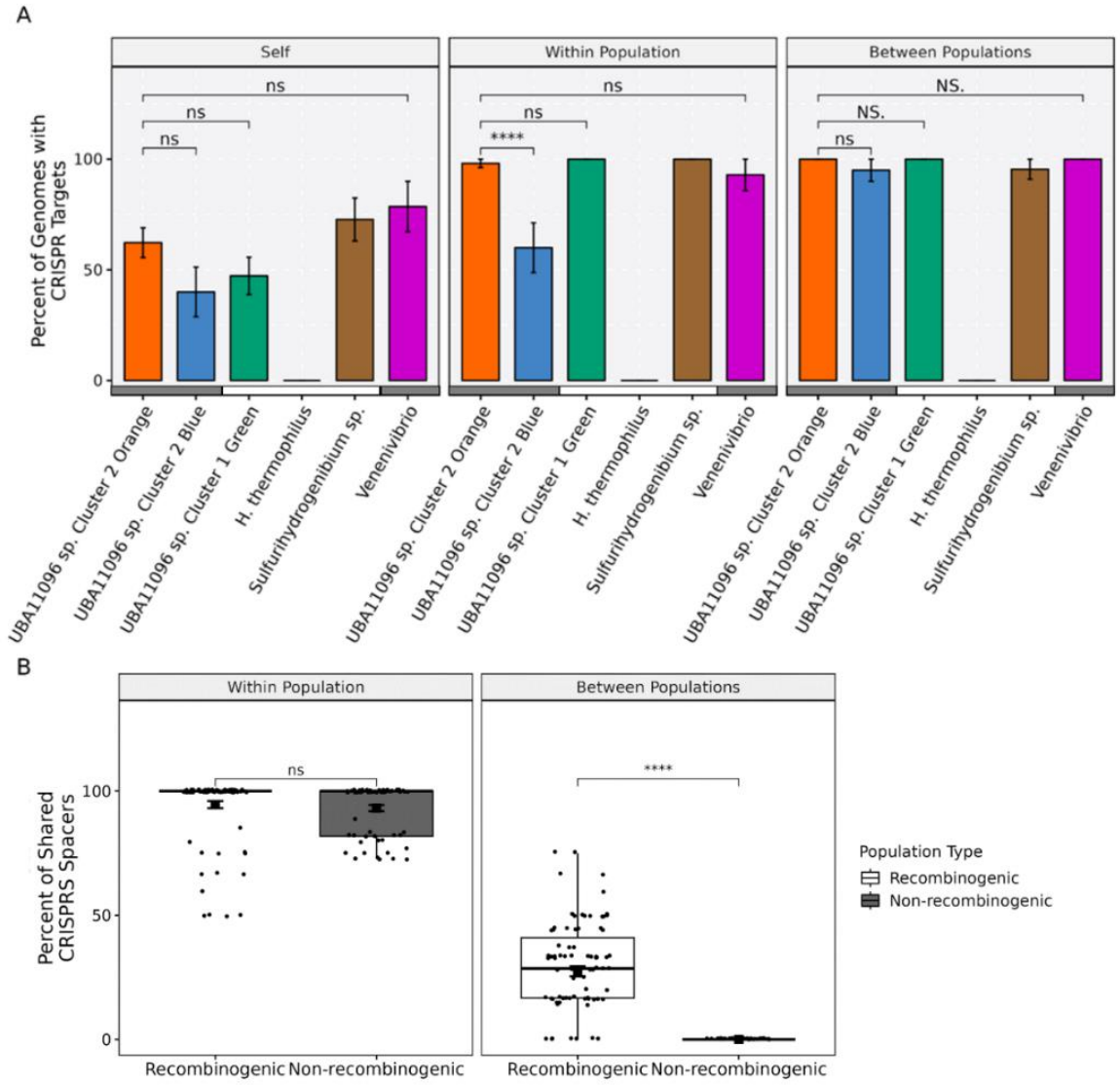

Figure 4.34. Prevalence of CRISPR spacers in *Aquificota* genomes and their targets. No significant differences in the targets of CRISPR spacers, whether self-targeting or targets within and between population (A), were observed between recombining and non-recombining populations. There is no significant difference between population types on the percentile of CRISPR spacers within a genome shared with others within the same populations (B) but, significantly differs with spacers shared between populations. UBA11096 sp. Cluster 2 Orange and Blue share the highest amount of its CRISPR spacers with each other (see SUPP. 18 Figure 8.6). Values are means (bar and squares) ± standard error. Statistical analysis by Wilcoxon rank sum tests (a-d), *p<0.05, **p<0.01, ****p<0.0001.

4.5. Conclusion

This chapter examined MGEs' role in horizontal gene flow-driven bacterial speciation in recombining *Aquificota* populations. Comparison of MGE abundance between recombining and clonal *Aquificota* shows no distinct correlation, suggesting that MGE abundance is unique in each population. Furthermore, genomic mapping of MGEs with identified swept regions and functional analysis of their orthologous genes showed that MGEs do not co-occur with population-differentiating swept regions, and MGEs have a functional profile distinct from ecologically relevant homologously recombined swept regions. Disentangling MGEs' role in the horizontal gene flow in the recombining *Aquificota* population supports the reverse ecology approach (Arevalo et al., 2019b) and confirms that horizontal gene flow-driven bacterial adaptation in recombining *Aquificota* populations is driven by homologous recombination (Arevalo et al., 2019a; Fraser et al., 2007; Shapiro et al., 2012).

Additionally, in contrast to previous studies showing CRISPR-Cas bacterial defence systems as shared between genomes primarily through MGE (Koonin et al., 2020), this study demonstrated that bacterial defence systems could also be potentially shared through homologous recombination as exemplified by shared CRISPR spacer arrays and Cas proteins in recombining *UBA11096 sp.* populations. Transfer of bacterial host defence systems through homologous recombination shows that homologous recombination also plays a role in regulating host-parasite dynamics in bacteria (Croucher et al., 2016). Lastly, the persistence of bacterial immune defence systems in the bacterial community through homologous recombination supports previous studies showing that recombining populations act as reservoirs for genetic material (Baumdicker et al., 2012).

4.6. Acknowledgement

This research was supported by the Marsden Fund, administered by the Royal Society of New Zealand (17-UOW-026). Access to Kuirau Park was granted by the Rotorua City Council.

4.7. References

Abby, S.S., Néron, B., Ménager, H., Touchon, M., Rocha, E.P.C., 2014. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. PLOS ONE 9, e110726. https://doi.org/10.1371/journal.pone.0110726

Acman, M., Wang, R., van Dorp, L., Shaw, L.P., Wang, Q., Luhmann, N., Yin, Y., Sun, S., Chen, H., Wang, H., Balloux, F., 2022. Role of mobile genetic elements in the global dissemination of the carbapenem resistance gene blaNDM. Nat Commun 13, 1131. https://doi.org/10.1038/s41467-022-28819-2

Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J., Polz, M.F., 2019a. A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. Cell 178, 820-834.e14. https://doi.org/10.1016/j.cell.2019.06.033

Arevalo, P., VanInsberghe, D., Polz, M.F., 2019b. A Reverse Ecology Framework for Bacteria and Archaea, in: Polz, M.F., Rajora, O.P. (Eds.), Population Genomics: Microorganisms, Population Genomics. Springer International Publishing, Cham, pp. 77–96. https://doi.org/10.1007/13836_2018_46

Arnold, B.J., Huang, I.-T., Hanage, W.P., 2022. Horizontal gene transfer and adaptive evolution in bacteria. Nat Rev Microbiol 20, 206–218. https://doi.org/10.1038/s41579-021-00650-4

Baumdicker, F., Hess, W.R., Pfaffelhuber, P., 2012. The Infinitely Many Genes Model for the Distributed Genome of Bacteria. Genome Biology and Evolution 4, 443–456. https://doi.org/10.1093/gbe/evs016

Bobay, L.-M., Ellis, B.S.-H., Ochman, H., 2018. ConSpeciFix: classifying prokaryotic species based on gene flow. Bioinformatics 34, 3738–3740. https://doi.org/10.1093/bioinformatics/bty400

Bobay, L.-M., Ochman, H., 2017. Biological species are universal across Life's domains. Genome Biol Evol 9, 491–501. https://doi.org/10.1093/gbe/evx026

Bowers, R.M., Nayfach, S., Schulz, F., Jungbluth, S.P., Ruhl, I.A., Sheremet, A., Lee, J., Goudeau, D., Eloe-Fadrosh, E.A., Stepanauskas, R., Malmstrom, R.R., Kyrpides, N.C., Dunfield, P.F., Woyke, T., 2022. Dissecting the dominant hot spring microbial populations based on community-wide sampling at single-cell genomic resolution. ISME J 16, 1337–1347. https://doi.org/10.1038/s41396-021-01178-4

Brown, C.L., Mullet, J., Hindi, F., Stoll, J.E., Gupta, S., Choi, M., Keenum, I., Vikesland, P., Pruden, A., Zhang, L., 2022. mobileOG-db: a Manually Curated Database of Protein Families Mediating the Life Cycle of Bacterial Mobile Genetic Elements. Appl Environ Microbiol 88, e0099122. https://doi.org/10.1128/aem.00991-22

Buchfink, B., Reuter, K., Drost, H.-G., 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18, 366–368. https://doi.org/10.1038/s41592-021-01101-x

Carraro, N., Rivard, N., Burrus, V., Ceccarelli, D., 2017. Mobilizable genomic islands, different strategies for the dissemination of multidrug resistance and other adaptive traits. Mob Genet Elements 7, 1–6. https://doi.org/10.1080/2159256X.2017.1304193

Chevallereau, A., Meaden, S., van Houte, S., Westra, E.R., Rollie, C., 2019. The effect of bacterial mutation rate on the evolution of CRISPR-Cas adaptive immunity. Philosophical Transactions of the Royal Society B: Biological Sciences 374, 20180094. https://doi.org/10.1098/rstb.2018.0094

Cordero, O.X., Polz, M.F., 2014. Explaining microbial genomic diversity in light of evolutionary ecology. Nature Reviews Microbiology 12, 263–273. https://doi.org/10.1038/nrmicro3218

Coudert, E., Gehant, S., de Castro, E., Pozzato, M., Baratin, D., Neto, T., Sigrist, C.J.A., Redaschi, N., Bridge, A., The UniProt Consortium, 2023. Annotation of biologically relevant ligands in UniProtKB using ChEBI. Bioinformatics 39, btac793. https://doi.org/10.1093/bioinformatics/btac793

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., Pourcel, C., 2018. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Research 46, W246–W251. https://doi.org/10.1093/nar/gky425

Croucher, N.J., Mostowy, R., Wymant, C., Turner, P., Bentley, S.D., Fraser, C., 2016. Horizontal DNA Transfer Mechanisms of Bacteria as Weapons of Intragenomic Conflict. PLOS Biology 14, e1002394. https://doi.org/10.1371/journal.pbio.1002394

da Silva Filho, A.C., Raittz, R.T., Guizelini, D., De Pierri, C.R., Augusto, D.W., dos Santos-Weiss, I.C.R., Marchaukoski, J.N., 2018. Comparative Analysis of Genomic Island Prediction Tools. Front Genet 9, 619. https://doi.org/10.3389/fgene.2018.00619

de la Higuera, I., Lázaro, E., 2022. Viruses in astrobiology. Frontiers in Microbiology 13. https://doi.org/10.3389/fmicb.2022.1032918

Doolittle, W.F., Papke, R.T., 2006. Genomics and the bacterial species problem. Genome Biology 7, 116. https://doi.org/10.1186/gb-2006-7-9-116

Durrant, M.G., Li, M.M., Siranosian, B.A., Montgomery, S.B., Bhatt, A.S., 2020. A Bioinformatic Analysis of Integrative Mobile Genetic Elements Highlights Their Role in Bacterial Adaptation. Cell Host & Microbe 27, 140-153.e9. https://doi.org/10.1016/j.chom.2019.10.022

Edgar, R., Qimron, U., 2010. The Escherichia coli CRISPR System Protects from λ Lysogenization, Lysogens, and Prophage Induction. J Bacteriol 192, 6291–6294. https://doi.org/10.1128/JB.00644-10

Fraser, C., Hanage, W.P., Spratt, B.G., 2007. Recombination and the Nature of Bacterial Speciation. Science 315, 476–480. https://doi.org/10.1126/science.1127573

González-Torres, P., Rodríguez-Mateos, F., Antón, J., Gabaldón, T., 2019. Impact of Homologous Recombination on the Evolution of Prokaryotic Core Genomes. mBio 10, e02494-18. https://doi.org/10.1128/mBio.02494-18

Grant, J.R., Enns, E., Marinier, E., Mandal, A., Herman, E.K., Chen, C.-Y., Graham, M., Van Domselaar, G., Stothard, P., 2023. Proksee: in-depth characterization and visualization of bacterial genomes. Nucleic Acids Res 51, W484–W492. https://doi.org/10.1093/nar/gkad326

Grissa, I., Vergnaud, G., Pourcel, C., 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35, W52-57. https://doi.org/10.1093/nar/gkm360

Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O., Pratama, A.A., Gazitúa, M.C., Vik, D., Sullivan, M.B., Roux, S., 2021. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome 9, 37. https://doi.org/10.1186/s40168-020-00990-y

Hackl, T., Laurenceau, R., Ankenbrand, M.J., Bliem, C., Cariani, Z., Thomas, E., Dooley, K.D., Arellano, A.A., Hogle, S.L., Berube, P., Leventhal, G.E., Luo, E., Eppley, J.M., Zayed, A.A., Beaulaurier, J., Stepanauskas, R., Sullivan, M.B., DeLong, E.F., Biller, S.J., Chisholm, S.W., 2023. Novel integrative elements and genomic plasticity in ocean ecosystems. Cell 186, 47-62.e16. https://doi.org/10.1016/j.cell.2022.12.006

Hanage, W.P., Fraser, C., Spratt, B.G., 2005. Fuzzy species among recombinogenic bacteria. BMC Biology 3, 6. https://doi.org/10.1186/1741-7007-3-6

Hetzer, A., McDonald, I.R., Morgan, H.W., 2008. Venenivibrio stagnispumantis gen. nov., sp. nov., a thermophilic hydrogen-oxidizing bacterium isolated from Champagne Pool, Waiotapu, New Zealand. International Journal of Systematic and Evolutionary Microbiology 58, 398–403. https://doi.org/10.1099/ijs.0.64842-0

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., von Mering, C., Bork, P., 2017. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Molecular Biology and Evolution 34, 2115–2122. https://doi.org/10.1093/molbev/msx148

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119. https://doi.org/10.1186/1471-2105-11-119

Iranzo, J., Cuesta, J.A., Manrubia, S., Katsnelson, M.I., Koonin, E.V., 2017. Disentangling the effects of selection and loss bias on gene dynamics. Proceedings of the National Academy of Sciences 114, E5616–E5624. https://doi.org/10.1073/pnas.1704925114

Iranzo, J., Puigbò, P., Lobkovsky, A.E., Wolf, Y.I., Koonin, E.V., 2016. Inevitability of Genetic Parasites. Genome Biol Evol 8, 2856–2869. https://doi.org/10.1093/gbe/evw193

Johansson, M.H.K., Aarestrup, F.M., Petersen, T.N., 2023. Importance of mobile genetic elements for dissemination of antimicrobial resistance in metagenomic sewage samples across the world. PLoS One 18, e0293169. https://doi.org/10.1371/journal.pone.0293169

Kassambara, A., 2023. ggpubr: "ggplot2" Based Publication Ready Plots.

Konstantinidis, K.T., DeLong, E.F., 2008. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. ISME J 2, 1052–1065. https://doi.org/10.1038/ismej.2008.62

Koonin, E.V., Makarova, K.S., 2017. Mobile Genetic Elements and Evolution of CRISPR-Cas Systems: All the Way There and Back. Genome Biol Evol 9, 2812–2825. https://doi.org/10.1093/gbe/evx192

Koonin, E.V., Makarova, K.S., Wolf, Y.I., Krupovic, M., 2020. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. Nat Rev Genet 21, 119–131. https://doi.org/10.1038/s41576-019-0172-9

Krassowski, M., 2020. ComplexUpset.

Kupczok, A., Neve, H., Huang, K.D., Hoeppner, M.P., Heller, K.J., Franz, C.M.A.P., Dagan, T., 2018. Rates of Mutation and Recombination in Siphoviridae Phage Genome Evolution over Three Decades. Mol Biol Evol 35, 1147–1159. https://doi.org/10.1093/molbev/msy027

Lao, J., Lacroix, T., Guédon, G., Coluzzi, C., Payot, S., Leblond-Bourget, N., Chiapello, H., 2022. ICEscreen: a tool to detect Firmicute ICEs and IMEs, isolated or enclosed in composite structures. NAR Genomics and Bioinformatics 4, lqac079. https://doi.org/10.1093/nargab/lqac079

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., Pfister, H., 2014. UpSet: Visualization of Intersecting Sets. IEEE Transactions on Visualization and Computer Graphics 20, 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248

Luo, X.-Q., Wang, P., Li, J.-L., Ahmad, M., Duan, L., Yin, L.-Z., Deng, Q.-Q., Fang, B.-Z., Li, S.-H., Li, W.-J., 2022. Viral community-wide auxiliary metabolic genes differ by lifestyles, habitats, and hosts. Microbiome 10, 190. https://doi.org/10.1186/s40168-022-01384-y

Mahmoudabadi, G., Phillips, R., 2018. A comprehensive and quantitative exploration of thousands of viral genomes. eLife 7, e31955. https://doi.org/10.7554/eLife.31955

Marttinen, P., Hanage, W.P., 2017. Speciation trajectories in recombining bacterial species. PLOS Computational Biology 13, e1005640. https://doi.org/10.1371/journal.pcbi.1005640

Mell, J.C., Redfield, R.J., 2014. Natural Competence and the Evolution of DNA Uptake Specificity. Journal of Bacteriology 196, 1471–1483. https://doi.org/10.1128/jb.01293-13

Novick, A., Doolittle, W.F., 2019. Horizontal persistence and the complexity hypothesis. Biol Philos 35, 2. https://doi.org/10.1007/s10539-019-9727-6

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25, 1043–1055. https://doi.org/10.1101/gr.186072.114

Polz, M.F., Alm, E.J., Hanage, W.P., 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends in Genetics 29, 170–175. https://doi.org/10.1016/j.tig.2012.12.006

Puigbò, P., Makarova, K.S., Kristensen, D.M., Wolf, Y.I., Koonin, E.V., 2017. Reconstruction of the evolution of microbial defense systems. BMC Evol Biol 17, 94. https://doi.org/10.1186/s12862-017-0942-y

R Core Team, 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rocha, E.P.C., Bikard, D., 2022. Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? PLoS Biol 20, e3001514. https://doi.org/10.1371/journal.pbio.3001514

Sanchez-Nieves, R.L., Zhang, C., Whitaker, R.J., 2023. Integrated conjugative plasmid drives high frequency chromosomal gene transfer in Sulfolobus islandicus. Frontiers in Microbiology 14. https://doi.org/10.3389/fmicb.2023.1114574

Shapiro, B.J., Friedman, J., Cordero, O.X., Preheim, S.P., Timberlake, S.C., Szabó, G., Polz, M.F., Alm, E.J., 2012. Population Genomics of Early Events in the Ecological Differentiation of Bacteria. Science 336, 48–51. https://doi.org/10.1126/science.1218198

Shapiro, B.J., Leducq, J.-B., Mallet, J., 2016. What Is Speciation? PLOS Genetics 12, e1005860. https://doi.org/10.1371/journal.pgen.1005860

Shapiro, B.J., Polz, M.F., 2015. Microbial Speciation. Cold Spring Harb Perspect Biol 7, a018143. https://doi.org/10.1101/cshperspect.a018143

Sharma, P., Johnson, M.A., Mazloom, R., Allen, C., Heath, L.S., Lowe-Power, T.M., Vinatzer, B.A., 2022. Meta-analysis of the Ralstonia solanacearum species complex (RSSC) based on comparative evolutionary genomics and reverse ecology. Microb Genom 8, 000791. https://doi.org/10.1099/mgen.0.000791

Sobecky, P.A., Hazen, T.H., 2009. Horizontal Gene Transfer and Mobile Genetic Elements in Marine Systems, in: Gogarten, M.B., Gogarten, J.P., Olendzenski, L.C. (Eds.), Horizontal

Gene Transfer: Genomes in Flux, Methods in Molecular Biology. Humana Press, Totowa, NJ, pp. 435–453. https://doi.org/10.1007/978-1-60327-853-9_25

Steinegger, M., Söding, J., 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35, 1026–1028. https://doi.org/10.1038/nbt.3988

Sun, J., Lu, F., Luo, Y., Bie, L., Xu, L., Wang, Y., 2023. OrthoVenn3: an integrated platform for exploring and visualizing orthologous data across genomes. Nucleic Acids Research 51, W397–W403. https://doi.org/10.1093/nar/gkad313

Takeuchi, N., Cordero, O.X., Koonin, E.V., Kaneko, K., 2015. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. BMC Biology 13, 20. https://doi.org/10.1186/s12915-015-0131-7

Vale, P.F., Lafforgue, G., Gatchitch, F., Gardan, R., Moineau, S., Gandon, S., 2015. Costs of CRISPR-Cas-mediated resistance in Streptococcus thermophilus. Proceedings of the Royal Society B: Biological Sciences 282, 20151270. https://doi.org/10.1098/rspb.2015.1270

VanInsberghe, D., Arevalo, P., Chien, D., Polz, M.F., 2020. How can microbial population genomics inform community ecology? Philosophical Transactions of the Royal Society B: Biological Sciences 375, 20190253. https://doi.org/10.1098/rstb.2019.0253

Vercoe, R.B., Chang, J.T., Dy, R.L., Taylor, C., Gristwood, T., Clulow, J.S., Richter, C., Przybilski, R., Pitman, A.R., Fineran, P.C., 2013. Cytotoxic Chromosomal Targeting by CRISPR/Cas Systems Can Reshape Bacterial Genomes and Expel or Remodel Pathogenicity Islands. PLoS Genet 9, e1003454. https://doi.org/10.1371/journal.pgen.1003454

Vernikos, G.S., Parkhill, J., 2006. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands. Bioinformatics 22, 2196–2203. https://doi.org/10.1093/bioinformatics/btl369

Vogan, A.A., Higgs, P.G., 2011. The advantages and disadvantages of horizontal gene transfer and the emergence of the first species. Biol Direct 6, 1. https://doi.org/10.1186/1745-6150-6-1

von Wintersdorff, C.J.H., Penders, J., van Niekerk, J.M., Mills, N.D., Majumder, S., van Alphen, L.B., Savelkoul, P.H.M., Wolffs, P.F.G., 2016. Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. Front. Microbiol. 7. https://doi.org/10.3389/fmicb.2016.00173

Vos, M., Didelot, X., 2009. A comparison of homologous recombination rates in bacteria and archaea. ISME J 3, 199–208. https://doi.org/10.1038/ismej.2008.93

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wimmer, F., Beisel, C.L., 2020. CRISPR-Cas Systems and the Paradox of Self-Targeting Spacers. Frontiers in Microbiology 10. https://doi.org/10.3389%2Ffmicb.2019.03078

CHAPTER 5: Summary, Conclusions, and Future Directions

5.1. Thesis Framework and Aim

The "species" is the fundamental unit in biology and has large importance in our current society due to its wide use in medicine (Baron, 1996; Petrof et al., 2013; Pidot et al., 2014), environmental conservation (Pfenninger and Schwenk, 2007; Theissinger et al., 2023), and even law (National Research Council (US) Committee on Scientific Issues in the Endangered Species Act, 1995). Defining a species is one of the main unsolved problems in microbiology (Doolittle and Papke, 2006). The lack of a 'species' concept in bacteria is a problem since many microbiologists still erroneously try to apply bacteria theories developed for sexual eukaryotes, often based on assumptions of panmictic populations with freely circulating alleles (Bobay and Ochman, 2017). As such, a universal species concept that works for both eukaryotes and prokaryotes is needed. A universal species concept does not necessarily mean that a single dictated boundary is applied across all domains; it is the application of a single methodological approach to delineate boundaries between species (Staley, 2013). Allowing species to be delineated at the same metric across all domains of life signifies that species act uniformly across taxa and, therefore, is a fundamental unit for studying ecology and evolution.

Varied methods in defining a species have been proposed (Lagache et al., 2013), but they agree that species are organisms with a common evolutionary fate (De Queiroz, 2007). The problem in defining and delineating a species arises from the lack of consensus on the evolutionary mechanisms that lead to speciation. This problem is more pronounced in bacteria, which can acquire adaptive genes through vertical and horizontal gene flow (Cohan, 2002; Hanage et al., 2005; Shapiro, 2016). The current widely used model of speciation in bacteria, the ecotype theory, assumes that vertical gene flow dominates the mechanisms of acquisition of adaptive genes (Cohan, 2001; Majewski and Cohan, 1999), and speciation occurs when adaptive mutants expand into the population through genome-wide sweeps brought by periodic selection (Cohan, 2001). As such, ecotype theory adds another layer of complexity to its model (e.g., Adapt Globally, Act Locally) when considering the role of horizontal gene flow in acquiring adaptive genes. It dismisses the possibility of acquiring adaptive genes through gene-specific sweep (Cohan, 2016; Majewski and Cohan, 1999). Therefore, the ecological population was proposed, which considers both genome-wide and gene-specific sweeps as possible mechanisms for acquiring adaptive genes (Bobay and Ochman, 2017; Fraser et al., 2007; Shapiro and Polz, 2014). The ecological population theory is a mechanism-agnostic model with the only criteria for speciation being discontinuity in total gene flow. It provides a much simpler, potentially

universal, criteria for bacterial speciation. However, a question remains: does ecological population theory work across all bacteria and archaea?

Therefore, this study aimed to examine the applicability of ecological population theory to metabolically limited *Aquificota* bacteria. With the high potential for deleterious consequences during horizontal gene flow in metabolically limited bacteria, the application of ecological population theory in *Aquificota* populations implies its generalisability as a model of speciation in bacteria. Ultimately, this can provide insight into developing a universal definition of a bacterial species and also provide a stepping stone in developing a universal definition of species across all domains of Life.

5.2. Thesis Summary and Highlights

The work presented in this study examined the potential of ecological population theory and its relevance for natural bacterial communities. By isolating closely related and co-habiting *Aquificota* bacteria, this study demonstrated that intra-species genomic variation is driven by different degrees of horizontal gene flow and identified populations that can be delineated through discontinuities in gene flow, i.e., ecological populations. Examination of the population structure of the isolated *Aquificota* across a time series revealed a shifting population structure in recombining populations with a turnover in the dominant genomic background. Meanwhile, clonal populations remained constant with minimal changes in single nucleotide variations. The difference in the shifts in population diversity and structure between clonal and recombining populations implied that bacterial response to environmental change differs by the capacity of a population to undergo horizontal gene flow; it also confirmed that closely related ecological populations with similar metabolic restrictions can respond independently to sweeps (Bendall et al., 2016). Furthermore, reverse ecology analysis of the recombining population highlighted differences in the functionality of swept regions—supporting distinct ecological niches for each recombining ecological population. Lastly, this work demonstrated that horizontal gene flow in ecological populations of *Aquificota* was primarily driven by homologous recombination—an essential assumption of the ecological population theory.

Evaluating co-habiting and closely related *Aquificota* bacteria was made possible by a novel high-throughput methodology to obtain bacterial isolates with minimal enrichment. This isolation approach provided a cheap, scalable, and high-throughput way of studying intra-species genomic variations since it captured a representative diversity of a population in natural communities. It was designed for hydrogen-oxidising bacteria but can be expanded to any bacteria that can be isolated with some growth restrictions, such as auxotrophs, anaerobes, and microaerophilic populations. Isolating generalist heterotrophs using this method can also be done but might require a higher number of samples/wells used per round to achieve similar

throughput as an autotrophic bacteria. Improvements to the methodology (i.e., increasing well density or using liquid handling robots) may increase its throughput and efficiency for generalist heterotrophs. For this study, this method allowed the isolation of more than 440 bacteria from geothermal springs, of which 180 are *Aquificota* bacteria from Kuirau Park.

Chapter 2 illustrated that *Aquificota* exhibits intra-species genomic variations that can be delineated by discontinuities in gene flow. The intra-species genomic variation revealed three distinct populations composed of very closely related genomes, as ANI (>99%) and phylogenomics showed. Even with the closely related ANI for all populations, closer analysis showed heterogeneity among the isolated genomes. This observed heterogeneity was interpreted as differences in the degree of horizontal gene flow between organisms, with highly clonal genomes exhibiting ANI >99.8% while the more recombining populations at slightly lower 99.3-99.5% ANI. A similar gap in the ANI score has been observed in *Salinibacter ruber*, for which a four-fold reduction of the abundance of genomes can be seen between 99.2-99.8% ANI (Rodriguez-R et al., 2023b; Viver et al., 2024). This gap has been attributed to discrete sequence differences called *genomovars* (ANI ~99.5%), which are thought to be more diverse taxonomic units than clones and strains, which are defined as ANI similarity of >99.99% (Rodriguez-R et al., 2023b; Viver et al., 2024). These genomovars exhibit distinct genotypic and even phenotypic advantages (i.e., an abundance of osmoregulation genes) but are not strong enough for a total purge (hard sweep) in diversity; as such, they were thought of as less-adapted variants (Viver et al., 2023). Like the patterns observed in *Aquificota* (Chapter 2 Figure 2.3), I argue that the intra-species genomic variations observed in *Salinibacter ruber* could be driven by differential horizontal gene flows (Chapter 2 Figure 2.5), with more clonal genomes observed with high ANI similarities while genomes with lower ANI (~99.2%) represent different genomic backgrounds (genomovars) within the populations linked through recombination. Additionally, the lack of a purge of total diversity (Viver et al., 2024) may indicate a gene-specific or soft sweep occurring within the *Salinibacter ruber* population similar to that observed in *UBA11096 sp.* in Chapter 3 (Figure 3.6). The unique genomic backgrounds (genomovars) may act as genetic reservoirs within populations that allow the sharing of adaptive genes, resulting in gene-specific sweeps rather than a total purge of genomic diversity. Observing these gaps in many taxa (Rodriguez-R et al., 2023b) may indicate that ecological populations, delineated through discontinuities in gene flow, are more prevalent in natural bacterial communities than we currently realise. Hence, I would suggest that analysis of recent horizontal gene flow in closely related strains address this knowledge gap.

The observation of the gap in ANI values between strains or genomovars in Canadian *Aquificota* isolates is not evident since the maximum ANI observed is only 99.53% (see SUPP.5 Figure 6.5), which can be due to the potential formation of incomplete and discontiguous

genomes using SAGs (Bowers et al., 2022). Interestingly, there is an observed difference in the intrapopulation ANI of the Canadian *Aquificota* isolates compared to New Zealand ones (Chapter 3). The difference in ANI distribution mirrors the lower observed horizontal gene flow between geographical groups of *UBA11096* sp. cluster 2 Blue populations and could be a product of geographical distance between the Canadian and new Zealand populations since horizontal gene flow is primarily affected by sequence divergence (Arevalo et al., 2019b; Shapiro and Polz, 2014). Still, given that geographical distance affects bacterial dispersal (Greenlon et al., 2019), homologous recombination can be expected to be partially or fully affected by geographical distance since it takes many hosts for a gene to hop into and spread large distances. Therefore, it can be assumed that there could be a natural gap in ANI values and length bias (degree of recombination) between geographical variants of a population akin to the genomovar and species gap previously observed. However, such a drop in ANI is most likely not generalisable and only taxa-specific, as shown by previous studies on microbial dispersal in long distances (Hoetzinger et al., 2021) and should be disentangled from the effect of ecological differentiation (Hoetzinger et al., 2021; Jespersen et al., 2023; Wu et al., 2023).

Chapter 3 demonstrated that the population structure of recombining *Aquificota* populations shifted with time. Unlike the constant SNP diversity in metagenomic data and observation of singular clonal lineages in clonal *Aquificota* populations (i.e., *Sulfurihydrogenibium*, *UBA11096* sp. cluster 1 Green, and *H. thermophilus*), the *UBA11096* sp. cluster 2 recombining populations (Blue and Orange) exhibited variations in SNP diversity and shift in population structure from clusters composed of recombining genomes to mostly clonal and back to recombining clusters. Initially interpreted as a potential genome-wide sweep, subsequent analysis showed no unique genomic background expanded within the population. Rather, it was observed that different unique genomic backgrounds appear at each time point, with a few older unique genomic backgrounds returning with time. These changes in population structure suggest a turnover of unique genomic backgrounds. The recombining nature of *UBA11096* sp. cluster 2 Blue and Orange populations likely allows diverse genomic backgrounds to coexist, resulting in a diverse gene collection within the population. A turnover of unique genomic backgrounds within the population, rather than a hard sweep, may allow the population flexibility to respond to environmental shifts.

Functional analysis of swept genes also posited a possible exchange of genetic information between the swept regions *UBA11096* sp. cluster 2 Orange and Blue populations. These observations highlighted the potential for recombining populations to act as genetic reservoirs (Kupczok et al., 2018; Lasky, 2019). Genetic reservoirs allow the persistence of diverse genetic variants distributed in different unique genomic backgrounds. Genetic reservoirs benefit the population as they may allow genes to be selectively carried by a few unique genomic

variants, thereby spreading specialised functions across the population (the Black Queen hypothesis) and minimising fitness costs for each genome (Morris et al., 2012). Additionally, genetic reservoirs can host easily shared but difficult-to-acquire genes through *de novo* mutation (Granato and Kümmerli, 2017). With functions and metabolic products shared across the population or between populations, this event can lead to cooperation and coexistence for mutual survival (Lee et al., 2022; Özkaya et al., 2018; Zhao et al., 2019). Homologous recombination of similar genes and functions within and between populations also suggests a common response to environmental cues or threats, which could be synchronised through sweeps of similar alleles for sensor, receptor and regulatory genes within and between recombining populations.

Among the swept genes in *UBA11096* sp. cluster 2 populations (Blue and Orange) shown in Chapter 3, genes for energy production and conversion were shown to be preferentially shared within the Blue population (Figure 3.7). An increase in the number of shared genes across time points suggested that these genes were essential for the population's survival. Additionally, the preferential spread of adaptive genes in one population implied the potential role of these genes as population-differentiating adaptive genes. Among these genes were members of the electron transport chain and nitrogen fixation pathways, prime candidates to elucidate phenotypic characteristics to aid in describing novel species. Thus, reverse ecology methods, such as functional analysis of horizontal swept genes, may provide a faster and more sensitive analysis of a species or population's ecological niche than traditional screening of chemotaxonomic characteristics.

Lastly, Chapter 4 reaffirmed an assumption of the *ecological population* theory that horizontal gene flow is primarily driven by homologous recombination—through the lack of overlap in the spatial and functional distribution between mobile genetic elements (i.e., viral sequences and integrative and conjugative elements) and horizontally swept regions. Chapter 4 also showed that components of the CRISPR-Cas bacterial host defence system were shared through homologous recombination within and between *UBA11096* sp. cluster 2 Blue and Orange populations. The ability to share bacterial host defence systems supports the notion that recombining populations act as genetic reservoirs between each other. Synchronisation of the immune system within and between populations minimises the spread of the parasitic genetic element to spread wide within the bacterial community and aids in quarantining infected members, akin to the creation of herd immunity (Payne et al., n.d.; Turkington et al., 2019). Modelling of bacteriophage infection shows that the formation of herd immunity is reliant on population heterogeneity, which is often in the form of spatial segregation or micro-niches (Payne et al., n.d.). However, even with the lack of spatial segregation, I suggest that population heterogeneity can be driven by varying degrees of horizontal gene flow, resulting in the mosaic

distribution of bacterial host defence systems within the population (i.e., CRISPR arrays) (Andersson and Banfield, 2008). A mosaic pattern in CRISPR distribution alters the total fitness cost of the host defence system as it distributes the cost of hosting CRISPR arrays across the population.

Previous studies have shown that the bacterial host defence system is utilised by parasites as both an offensive tool for parasites/MGEs to integrate itself into the host genomes through homologous enzymatic machinery (Koonin et al., 2020) and also as a defensive tool for host and parasites to prevent hyper-infection by other parasites (Koonin and Makarova, 2017). This 'guns-for-hire' paradigm is often associated with CRISPR-Cas and other defence systems, such as restriction-modification nucleases, due to the evolutionary and functional similarity of bacterial host defence systems with MGE-related proteins, such as Cas proteins with transposons (Peters et al., 2017) and its translocase-functions (He et al., 2020). The paradigm is also supported by bacterial host systems often being horizontally dispersed through mobile genetic elements (Iranzo et al., 2017; Koonin and Makarova, 2017). Hijacking of the host defence system by parasites can also be beneficial to host bacteria as it facilitates the spread of immune defence systems, which can often be lost from the genome due to the fitness cost of carrying these systems and autoimmunity (Iranzo et al., 2017). However, as previously mentioned, Cas proteins can be shared solely through homologous recombination along with other immunity genes, such as restriction-modification enzyme ResIII (Saha et al., 1998). I suggest that aside from defence system hijacking by parasites, the recombining *Aquificota* host genomes leverage their own horizontal gene flow mechanisms (i.e., homologous recombination) to gain a proactive defence. Additionally, these host defence systems can be used for other purposes. An example is Cas3, which has helicase and translocase functions. Cas3 aids in the regulatory processing of RNA molecules, such as initiating biofilm formation in *Pseudomonas* (Cady) and spore differentiation and development in *Myxococcus xanthus* (Viswanathan et al., 2007).

5.3. Future works

5.3.1. Determining the role of negative-frequency dependent selection in gene-level sweeps in Aquificota

Analysis of recent horizontal gene flow using PopCOGenT relies on identifying genomic regions with minimal nucleotide variation. Random mutations would decrease the size of the identical genomic regions, and thus, genomes will not be identified as recombining. Recent and frequent horizontal gene flow would appear to have larger, long identical regions (Arevalo et al., 2019a). A problem, however, in using identical regions as a basis for studying horizontal gene flow is that enrichment of identical regions can also happen through negative frequency-

dependent selection (NFDS), which is not currently addressed in the pipeline (Arevalo et al., 2019a; Takeuchi et al., 2015).

Gene-level sweeps can also occur in bacteria through NFDS, such as those populations under the 'kill the winner' and 'black queen' dynamics (Cordero and Polz, 2014; Takeuchi et al., 2015). One example is that populations with loci that attract threats (i.e., viral predation or toxins) need to be kept at low copies in the population to minimise negative consequences. A high recombination rate in these loci could attract more threats and thus lead to lower fitness of the entire population. The formation of ecological populations through frequent recombination and gene-specific sweep in these loci is less likely. However, a modelling study of the population under NFDS shows that gene-specific sweeps can still occur, albeit with strict requirements (Cordero and Polz, 2014; Takeuchi et al., 2015). Gene-specific sweeps can occur when multiple loci are maintained under NFDS dynamics with sufficiently low recombination rates, and a separate but linked locus is under positive selection (Takeuchi et al., 2015). The significantly lower recombination rate in NFDS loci results in permissible recombination of adaptive traits at other loci of the genome. Recombination at this environmentally significant locus leads all genotypes with different variations of NFDS loci to be coherent in a single ecological niche dimension. The selection at the environmentally adaptive loci would result in a gene-specific sweep of the environmentally adapted alleles along other linked loci, including those under NFDS, forming a population with diverse genomic backgrounds. Meanwhile, a higher recombination rate at NFDS loci could lead to the NFDS loci becoming unlinked with the environmental loci. Selection against the environmentally adaptive loci would result in a genome-wide sweep of distinct genomic backgrounds carrying the adapted alleles (Takeuchi et al., 2015).

Determining the role of NFDS in the gene-specific sweep in recombining Aquificota populations can shed light on the observed turnover of genomic background in the longitudinal study and the potential for genetic reservoirs within these recombining populations. Additionally, as negative frequency-dependent selection is assumed to be common in natural bacterial populations (Takeuchi et al., 2015), the presence of NFDS-driven gene-specific sweeps in recombining natural *Aquificota* populations would also provide more evidence of the commonality of gene-specific sweeps and ecological populations in bacteria. Nevertheless, whether it is rapid horizontal gene flow or NFDS-driven gene-level sweep, the presence of gene-level sweeps in chemolithoautotrophic bacteria challenges the premise of ecotype theory that genome-wide sweeps drive the formation of ecologically and genotypically coherent populations in natural bacterial communities (Cohan, 2016).

5.3.2. Reverse ecology analysis of swept genes

Methods that require no *a priori* knowledge of population boundaries and state, such as PopCOGenT, are ideal in performing reverse ecology as no assumption about the bacteria is required. Reverse ecology tries to infer the role of bacteria in the environment through sequence genomic data (Arevalo et al., 2019b; Li et al., 2008). Unfortunately, verification of the swept genes in the recombined regions of *UBA11096* sp. Blue and Orange populations have not yet been done due to time constraints. As such, validation of the phenotype through physiological and chemotaxonomic analysis of the isolated bacteria is needed. Phenotypic characterisation of putative population differentiating traits, such as the energy production conferred by nitrogen fixation genes sweeping *UBA11096* sp. should be prioritised. Additionally, transcriptomic analysis of these genes in both the lab and monitoring *in situ* can confirm whether the identified genes is truly functional and define population boundaries in natural communities.

Aside from phenotypic characterisation for validating swept gene functions, the three distinct and unnamed 'species' of *UBA11096* sp. should be characterised using traditional species-defining metrics. Defining a bacterial 'species' using physiological, chemotaxonomic, phenotypic, and other traditional criteria (Stackebrandt et al., 2002) could further validate the species boundaries observed through genomic sequencing and analysis. Phenotypic characterisation and naming of the UBA11096 species would allow the potential use of these populations as markers for environmental monitoring. Lastly, as far as I have researched, naming and characterising the UBA11096 sp. populations would be the instance of bacterial species defined and named through discontinuities in horizontal gene flow.

5.3.3. Gene-specific sweeps across vast geographical distances

This study has also shown that recombining populations *UBA11096* sp. remain connected via horizontal gene flow despite the vast geographical distance. This observation raises the question of how horizontal gene flow or homologous recombination is maintained across vast distances. Previous declared hypotheses include infrequent large eruptions leading to slow dispersal of the bacteria across continents (Herbold et al., 2014) and the island hopping effect in which intermediate populations allow dispersal of identical genomic regions across vast distances (Greenlon et al., 2019). Testing these hypotheses requires analysis of Aquificota populations in diverse and geographically distant places to identify potential intermediary populations. Given the observed decrease in ANI and the degree of recombined regions between the Canadian and New Zealand populations, it is worthwhile to check if the removal of geographic barriers can result in the reformation of a homogeneous population or would still lead to further diversification. The results of these studies can aid in determining the limits of

identifying population boundaries through the study of the degree of recombined regions (PopCOGenT).

5.4. Impact of the Study

The establishment of a definition of a species as an ecologically and genotypically coherent group of bacteria allows it to be treated as a true fundamental unit of ecology and evolution. Fundamental units are cohesive units that share similar properties and act as building blocks to evaluate scientific phenomena (Jennings, 1936). Understanding species as a fundamental unit would allow a better understanding of the biological processes that guide their interaction with each other (ecology) and their formation and evolution. (Mahner and Bunge, 1997). Others can argue that genes matter more as they contain information and directly experience diversification (Koonin and Wolf, 2009). However, it can be argued that individuals diversify, but their collective action results in the selective advantages to form a cohesive group (Schmitt, 2004). Therefore, changes in the individual do not matter if changes do not affect the whole population, and evolution affects the whole population and not just the individual. Thus, it is highly important to delineate individuals into a properly defined species reflecting their coherent biological properties.

Overall, this study demonstrated that *Aquificota* bacteria can form ecological populations demarcated by discontinuities in horizontal gene flow primarily driven by homologous recombination. The applicability of ecological population theory in metabolically limited *Aquificota* bacteria advocates for the ecological population theory as a generalisable model of speciation for bacteria. Thus, a generalisable ecological population theory defines bacterial species based on a unified criterion across all domains of life and allows the study of all ecological and evolutionary processes under the same framework.

5.5. References

Andersson, A.F., Banfield, J.F., 2008. Virus population dynamics and acquired virus resistance in natural microbial communities. Science 320, 1047–1050. https://doi.org/10.1126/science.1157358

Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J., Polz, M.F., 2019a. A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. Cell 178, 820-834.e14. https://doi.org/10.1016/j.cell.2019.06.033

Arevalo, P., VanInsberghe, D., Polz, M.F., 2019b. A Reverse Ecology Framework for Bacteria and Archaea, in: Polz, M.F., Rajora, O.P. (Eds.), Population Genomics: Microorganisms, Population Genomics. Springer International Publishing, Cham, pp. 77–96. https://doi.org/10.1007/13836_2018_46

Baron, E.J., 1996. Classification, in: Baron, S. (Ed.), Medical Microbiology. University of Texas Medical Branch at Galveston, Galveston (TX).

Bendall, M.L., Stevens, S.L., Chan, L.-K., Malfatti, S., Schwientek, P., Tremblay, J., Schackwitz, W., Martin, J., Pati, A., Bushnell, B., Froula, J., Kang, D., Tringe, S.G., Bertilsson, S., Moran, M.A., Shade, A., Newton, R.J., McMahon, K.D., Malmstrom, R.R., 2016. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. The ISME Journal 10, 1589–1601. https://doi.org/10.1038/ismej.2015.241

Bobay, L.-M., Ochman, H., 2017. Biological species are universal across Life's domains. Genome Biol Evol 9, 491–501. https://doi.org/10.1093/gbe/evx026

Bowers, R.M., Nayfach, S., Schulz, F., Jungbluth, S.P., Ruhl, I.A., Sheremet, A., Lee, J., Goudeau, D., Eloe-Fadrosh, E.A., Stepanauskas, R., Malmstrom, R.R., Kyrpides, N.C., Dunfield, P.F., Woyke, T., 2022. Dissecting the dominant hot spring microbial populations based on community-wide sampling at single-cell genomic resolution. ISME J 16, 1337–1347. https://doi.org/10.1038/s41396-021-01178-4

Cohan, F.M., 2016. Bacterial Speciation: Genetic Sweeps in Bacterial Species. Current Biology 26, R112–R115. https://doi.org/10.1016/j.cub.2015.10.022

Cohan, F.M., 2002. What are Bacterial Species? Annual Review of Microbiology 56, 457–487. https://doi.org/10.1146/annurev.micro.56.012302.160634

Cohan, F.M., 2001. Bacterial Species and Speciation. Syst Biol 50, 513–524. https://doi.org/10.1080/10635150118398

Cordero, O.X., Polz, M.F., 2014. Explaining microbial genomic diversity in light of evolutionary ecology. Nature Reviews Microbiology 12, 263–273. https://doi.org/10.1038/nrmicro3218

De Queiroz, K., 2007. Species Concepts and Species Delimitation. Systematic Biology 56, 879–886. https://doi.org/10.1080/10635150701701083

Doolittle, W.F., Papke, R.T., 2006. Genomics and the bacterial species problem. Genome Biology 7, 116. https://doi.org/10.1186/gb-2006-7-9-116

Fraser, C., Hanage, W.P., Spratt, B.G., 2007. Recombination and the Nature of Bacterial Speciation. Science 315, 476–480. https://doi.org/10.1126/science.1127573

Granato, E.T., Kümmerli, R., 2017. The path to re-evolve cooperation is constrained in Pseudomonas aeruginosa. BMC Evol Biol 17, 214. https://doi.org/10.1186/s12862-017-1060-6

Greenlon, A., Chang, P.L., Damtew, Z.M., Muleta, A., Carrasquilla-Garcia, N., Kim, D., Nguyen, H.P., Suryawanshi, V., Krieg, C.P., Yadav, S.K., Patel, J.S., Mukherjee, A., Udupa, S., Benjelloun, I., Thami-Alami, I., Yasin, M., Patil, B., Singh, S., Sarma, B.K., von Wettberg, E.J.B., Kahraman, A., Bukun, B., Assefa, F., Tesfaye, K., Fikre, A., Cook, D.R., 2019. Global-level population genomics reveals differential effects of geography and phylogeny on horizontal gene transfer in soil bacteria. Proc Natl Acad Sci U S A 116, 15200–15209. https://doi.org/10.1073/pnas.1900056116

Hanage, W.P., Fraser, C., Spratt, B.G., 2005. Fuzzy species among recombinogenic bacteria. BMC Biology 3, 6. https://doi.org/10.1186/1741-7007-3-6

He, L., St. John James, M., Radovcic, M., Ivancic-Bace, I., Bolt, E.L., 2020. Cas3 Protein—A Review of a Multi-Tasking Machine. Genes (Basel) 11, 208. https://doi.org/10.3390/genes11020208

Herbold, C.W., Lee, C.K., McDonald, I.R., Cary, S.C., 2014. Evidence of global-scale aeolian dispersal and endemism in isolated geothermal microbial communities of Antarctica. Nat Commun 5, 3875. https://doi.org/10.1038/ncomms4875

Hetzer, A., McDonald, I.R., Morgan, H.W., 2008. Venenivibrio stagnispumantis gen. nov., sp. nov., a thermophilic hydrogen-oxidizing bacterium isolated from Champagne Pool, Waiotapu, New Zealand. International Journal of Systematic and Evolutionary Microbiology 58, 398–403. https://doi.org/10.1099/ijs.0.64842-0

Hoetzinger, M., Pitt, A., Huemer, A., Hahn, M.W., 2021. Continental-Scale Gene Flow Prevents Allopatric Divergence of Pelagic Freshwater Bacteria. Genome Biol Evol 13, evab019. https://doi.org/10.1093/gbe/evab019

Iranzo, J., Cuesta, J.A., Manrubia, S., Katsnelson, M.I., Koonin, E.V., 2017. Disentangling the effects of selection and loss bias on gene dynamics. Proceedings of the National Academy of Sciences 114, E5616–E5624. https://doi.org/10.1073/pnas.1704925114

Jennings, H.S., 1936. Fundamental Units in Biology. Science 84, 445–450. https://doi.org/10.1126/science.84.2186.445

Jespersen, M.L., Munk, P., Johansen, J., Kaas, R.S., Webel, H., Vigre, H., Nielsen, H.B., Rasmussen, S., Aarestrup, F.M., 2023. Global within-species phylogenetics of sewage microbes

suggest that local adaptation shapes geographical bacterial clustering. Commun Biol 6, 1–9. https://doi.org/10.1038/s42003-023-05083-8

Koonin, E.V., Makarova, K.S., 2017. Mobile Genetic Elements and Evolution of CRISPR-Cas Systems: All the Way There and Back. Genome Biol Evol 9, 2812–2825. https://doi.org/10.1093/gbe/evx192

Koonin, E.V., Makarova, K.S., Wolf, Y.I., Krupovic, M., 2020. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. Nat Rev Genet 21, 119–131. https://doi.org/10.1038/s41576-019-0172-9

Koonin, E.V., Wolf, Y.I., 2009. The fundamental units, processes and patterns of evolution, and the Tree of Life conundrum. Biol Direct 4, 33. https://doi.org/10.1186/1745-6150-4-33

Kupczok, A., Neve, H., Huang, K.D., Hoeppner, M.P., Heller, K.J., Franz, C.M.A.P., Dagan, T., 2018. Rates of Mutation and Recombination in Siphoviridae Phage Genome Evolution over Three Decades. Mol Biol Evol 35, 1147–1159. https://doi.org/10.1093/molbev/msy027

Lagache, L., Leger, J.-B., Daudin, J.-J., Petit, R.J., Vacher, C., 2013. Putting the Biological Species Concept to the Test: Using Mating Networks to Delimit Species. PLOS ONE 8, e68267. https://doi.org/10.1371/journal.pone.0068267

Lasky, J.R., 2019. Eco-evolutionary community turnover following environmental change. Evol Appl 12, 1434–1448. https://doi.org/10.1111/eva.12776

Lee, I.P.A., Eldakar, O.T., Gogarten, J.P., Andam, C.P., 2022. Bacterial cooperation through horizontal gene transfer. Trends in Ecology & Evolution 37, 223–232. https://doi.org/10.1016/j.tree.2021.11.006

Li, Y.F., Costello, J.C., Holloway, A.K., Hahn, M.W., 2008. "Reverse ecology" and the power of population genomics. Evolution 62, 2984–2994. https://doi.org/10.1111/j.1558-5646.2008.00486.x

Mahner, M., Bunge, M., 1997. Ecology, in: Mahner, M., Bunge, M. (Eds.), Foundations of Biophilosophy. Springer, Berlin, Heidelberg, pp. 171–198. https://doi.org/10.1007/978-3-662-03368-5_5

Majewski, J., Cohan, F.M., 1999. Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. Genetics 152, 1459–1474.

Morris, J.J., Lenski, R.E., Zinser, E.R., 2012. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. mBio 3, e00036-12. https://doi.org/10.1128/mBio.00036-12

National Research Council (US) Committee on Scientific Issues in the Endangered Species Act, 1995. Species Definitions and the Endangered Species Act, in: Science and the Endangered Species Act. National Academies Press (US).

Özkaya, Ö., Balbontín, R., Gordo, I., Xavier, K.B., 2018. Cheating on Cheaters Stabilizes Cooperation in Pseudomonas aeruginosa. Current Biology 28, 2070-2080.e6. https://doi.org/10.1016/j.cub.2018.04.093

Payne, P., Geyrhofer, L., Barton, N.H., Bollback, J.P., n.d. CRISPR-based herd immunity can limit phage epidemics in bacterial populations. eLife 7, e32035. https://doi.org/10.7554/eLife.32035

Peters, J.E., Makarova, K.S., Shmakov, S., Koonin, E.V., 2017. Recruitment of CRISPR-Cas systems by Tn7-like transposons. Proceedings of the National Academy of Sciences 114, E7358–E7366. https://doi.org/10.1073/pnas.1709035114

Petrof, E. o., Claud, E. c., Gloor, G. b., Allen-Vercoe, E., 2013. Microbial ecosystems therapeutics: a new paradigm in medicine? Beneficial Microbes 4, 53–65. https://doi.org/10.3920/BM2012.0039

Pfenninger, M., Schwenk, K., 2007. Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. BMC Evolutionary Biology 7, 121. https://doi.org/10.1186/1471-2148-7-121

Pidot, S.J., Coyne, S., Kloss, F., Hertweck, C., 2014. Antibiotics from neglected bacterial sources. International Journal of Medical Microbiology, Special issue: Antiinfectives 304, 14–22. https://doi.org/10.1016/j.ijmm.2013.08.011

Rodriguez-R, L.M., Conrad, R.E., Viver, T., Feistel, D.J., Lindner, B.G., Venter, S.N., Orellana, L.H., Amann, R., Rossello-Mora, R., Konstantinidis, K.T., 2023. An ANI gap within bacterial species that advances the definitions of intra-species units. mBio 15, e02696-23. https://doi.org/10.1128/mbio.02696-23

Saha, S., Ahmad, I., Reddy, Y.V., Krishnamurthy, V., Rao, D.N., 1998. Functional analysis of conserved motifs in type III restriction-modification enzymes. Biol Chem 379, 511–517. https://doi.org/10.1515/bchm.1998.379.4-5.511

Schmitt, M., 2004. The species as a unit of evolution and as an element of order. pp. 79–88.

Shapiro, B.J., 2016. How clonal are bacteria over time? Current Opinion in Microbiology, Environmental microbiology * Special Section: Megaviromes 31, 116–123. https://doi.org/10.1016/j.mib.2016.03.013

Shapiro, B.J., Polz, M.F., 2014. Ordering microbial diversity into ecologically and genetically cohesive units. Trends Microbiol 22, 235–247. https://doi.org/10.1016/j.tim.2014.02.006

Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A.D., Kämpfer, P., Maiden, M.C.J., Nesme, X., Rosselló-Mora, R., Swings, J., Trüper, H.G., Vauterin, L., Ward, A.C., Whitman, W.B., 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. International Journal of Systematic and Evolutionary Microbiology 52, 1043–1047. https://doi.org/10.1099/00207713-52-3-1043

Staley, J.T., 2013. Transitioning Toward a Universal Species Concept for the Classification of all Organisms, in: The Species Problem - Ongoing Issues. IntechOpen. https://doi.org/10.5772/53218

Takeuchi, N., Cordero, O.X., Koonin, E.V., Kaneko, K., 2015. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. BMC Biology 13, 20. https://doi.org/10.1186/s12915-015-0131-7

Theissinger, K., Fernandes, C., Formenti, G., Bista, I., Berg, P.R., Bleidorn, C., Bombarely, A., Crottini, A., Gallo, G.R., Godoy, J.A., Jentoft, S., Malukiewicz, J., Mouton, A., Oomen, R.A., Paez, S., Palsbøll, P.J., Pampoulie, C., Ruiz-López, M.J., Secomandi, S., Svardal, H., Theofanopoulou, C., de Vries, J., Waldvogel, A.-M., Zhang, G., Jarvis, E.D., Bálint, M., Ciofi, C., Waterhouse, R.M., Mazzoni, C.J., Höglund, J., Aghayan, S.A., Alioto, T.S., Almudi, I., Alvarez, N., Alves, P.C., Amorim do Rosario, I.R., Antunes, A., Arribas, P., Baldrian, P., Bertorelle, G., Böhne, A., Bonisoli-Alquati, A., Boštjančić, L.L., Boussau, B., Breton, C.M., Buzan, E., Campos, P.F., Carreras, C., Castro, L.Fi.C., Chueca, L.J., Čiampor, F., Conti, E., Cook-Deegan, R., Croll, D., Cunha, M.V., Delsuc, F., Dennis, A.B., Dimitrov, D., Faria, R., Favre, A., Fedrigo, O.D., Fernández, R., Ficetola, G.F., Flot, J.-F., Gabaldón, T., Agius, D.R., Giani, A.M., Gilbert, M.T.P., Grebenc, T., Guschanski, K., Guyot, R., Hausdorf, B., Hawlitschek, O., Heintzman, P.D., Heinze, B., Hiller, M., Husemann, M., Iannucci, A., Irisarri, I., Jakobsen, K.S., Klinga, P., Kloch, A., Kratochwil, C.F., Kusche, H., Layton, K.K.S., Leonard, J.A., Lerat, E., Liti, G., Manousaki, T., Marques-Bonet, T., Matos-Maraví, P., Matschiner, M., Maumus, F., Mc Cartney, A.M., Meiri, S., Melo-Ferreira, J., Mengual, X., Monaghan, M.T., Montagna, M., Mysłajek, R.W., Neiber, M.T., Nicolas, V., Novo, M., Ozretić, P., Palero, F., Pârvulescu, L., Pascual, M., Paulo, O.S., Pavlek, M., Pegueroles, C., Pellissier, L., Pesole, G., Primmer, C.R., Riesgo, A., Rüber, L., Rubolini, D., Salvi, D., Seehausen, O., Seidel, M., Studer, B., Theodoridis, S., Thines, M., Urban, L., Vasemägi, A., Vella, A., Vella, N., Vernes, S.C., Vernesi, C., Vieites, D.R., Wheat, C.W., Wörheide, G., Wurm, Y., Zammit, G., 2023. How genomics can help biodiversity conservation. Trends in Genetics 39, 545–559. https://doi.org/10.1016/j.tig.2023.01.005

Turkington, C.J.R., Morozov, A., Clokie, M.R.J., Bayliss, C.D., 2019. Phage-Resistant Phase-Variant Sub-populations Mediate Herd Immunity Against Bacteriophage Invasion of Bacterial Meta-Populations. Frontiers in Microbiology 10.

Viswanathan, P., Murphy, K., Julien, B., Garza, A.G., Kroos, L., 2007. Regulation of dev, an Operon That Includes Genes Essential for Myxococcus xanthus Development and CRISPR-Associated Genes and Repeats. J Bacteriol 189, 3738–3750. https://doi.org/10.1128/JB.00187-07

Viver, T., Conrad, R.E., Lucio, M., Harir, M., Urdiain, M., Gago, J.F., Suárez-Suárez, A., Bustos-Caparros, E., Sanchez-Martinez, R., Mayol, E., Fassetta, F., Pang, J., Mădălin Gridan, I., Venter, S., Santos, F., Baxter, B., Llames, M.E., Cristea, A., Banciu, H.L., Hedlund, B.P., Stott, M.B., Kämpfer, P., Amann, R., Schmitt-Kopplin, P., Konstantinidis, K.T., Rossello-Mora, R., 2023. Description of two cultivated and two uncultivated new Salinibacter species, one named following the rules of the bacteriological code: Salinibacter grassmerensis sp. nov.; and three named following the rules of the SeqCode: Salinibacter

pepae sp. nov., Salinibacter abyssi sp. nov., and Salinibacter pampae sp. nov. Systematic and Applied Microbiology 46, 126416. https://doi.org/10.1016/j.syapm.2023.126416

Viver, T., Conrad, R.E., Rodriguez-R, L.M., Ramírez, A.S., Venter, S.N., Rocha-Cárdenas, J., Llabrés, M., Amann, R., Konstantinidis, K.T., Rossello-Mora, R., 2024. Towards estimating the number of strains that make up a natural bacterial population. Nat Commun 15, 544. https://doi.org/10.1038/s41467-023-44622-z

Wu, Y., Hao, C., Chen, T.-W., Xie, Z., Zhang, Y., Guan, P., Wu, D., Scheu, S., 2023. Environmental distances are more important than geographic distances for predicting earthworm gut bacterial community composition. European Journal of Soil Biology 118, 103520. https://doi.org/10.1016/j.ejsobi.2023.103520

Zhao, K., Liu, L., Chen, X., Huang, T., Du, L., Lin, J., Yuan, Y., Zhou, Y., Yue, B., Wei, K., Chu, Y., 2019. Behavioral heterogeneity in quorum sensing can stabilize social cooperation in microbial populations. BMC Biol 17, 20. https://doi.org/10.1186/s12915-019-0639-3

APPENDIX A - CHAPTER 2

6.1. SUPPLEMENTARY 1. *Aquificales*-targeted Media and Cultivation

Table 6.1 Modified *Aquificales*-targeted Media based on Hetzer's Media (Hetzer et al., 2008)

	Molarity (M)	g per 1000 mL
18.2 MΩ · cm dH2O		
NaOH	0.0038	0.15
KCl	0.0067	0.50
MgCl2 · 6H2O	0.0104	1.36
MgSO4 · 7H2O	0.0284	7.00
Na2S2O3 · 5H2O	0.0081	2.00
CaCl2 · 2H2O	0.0027	0.40
NH4Cl	0.0037	0.20
K2HPO4	0.0014	0.25
MES	0.0100	1.95
-or MOPPSO	0.0100	2.25
-or HEPPSO	0.0100	2.68
Mineral Solution	1X	
1000X Mineral Solution		
18.2 MΩ · cm dH2O		1000 mL
Na2EDTA · 2H2O		5.00 g
CoCl2· 6H2O		1.50 g
MnCl2· 4H2O		1.00 g
FeSO4· 7H2O		1.00 g
ZnCl2		1.00 g
AlCl3· 6H2O		0.40 g
Na2WO4· 2H2O		0.30 g
CuCl2· 2H2O		0.20 g
NiSO4· 6H2O		0.20 g
Na2SeO3		0.10 g
H3BO3		0.10 g
Na2MoO4· 2H2O.		0.10 g

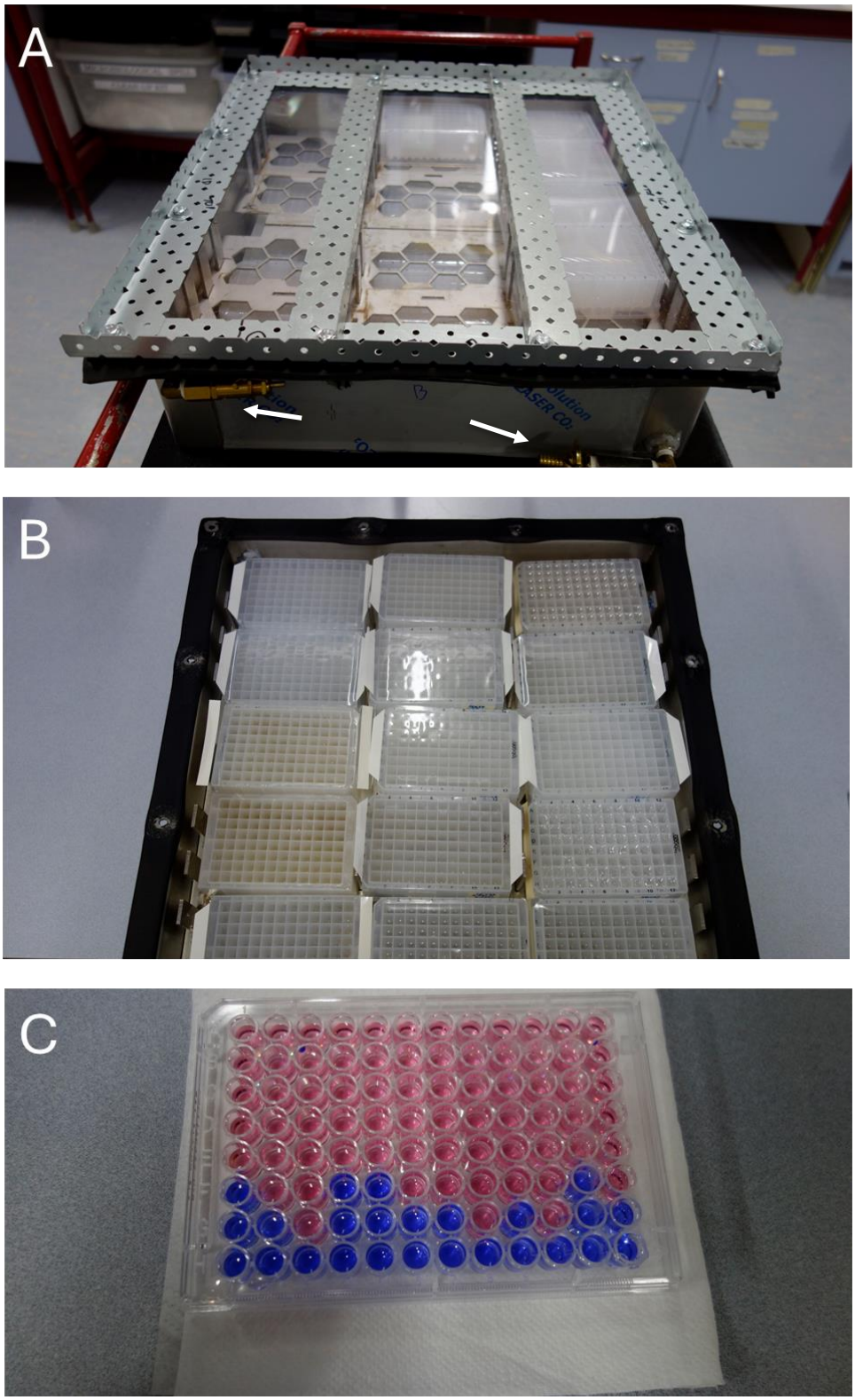

Figure 6.1 Environmental chamber setup. The chamber (A) was made from a steel box that was sealed using neoprene gaskets to minimise gas leakage with inlet or outlet one-way valve couplings to regulate gas flow (A – arrow). The chamber can hold a max of 30 deep-well plates (B) simulateneously and has a theoretical isolation capacity of 2,880 isolates. The growth in each well were tested using the resazurin assay (C).

6.2. SUPPLEMENTARY 2. Genome Completeness and Contamination

Table 6.2 CheckM contamination and completeness of *Aquificota* genomes from Kuirau Park

Genome	Completeness	Contamination	Genome size (bp)	# scaffolds	# contigs	N50 (scaffolds)	Mean scaffold length (bp)	Longest scaffold (bp)	GC	# predicted genes
G01_cluster_3	100	0.81	1726017	61	65	60902	28295	175401	42.6	1893
G02_cluster_3	100	1.63	1763844	76	93	52947	23208	154521	42.5	1949
G04_cluster_2	100	1.22	1772068	24	24	134588	73836	305112	42.5	1929
G05_cluster_2	100	1.22	1757641	61	68	53911	28813	145535	42.6	1938
G06_cluster_2	100	1.22	1757684	51	59	60005	34464	145535	42.6	1930
G07_cluster_2	100	1.22	1759041	57	65	53914	30860	145535	42.6	1936
G08_cluster_1	100	1.63	1821041	26	26	127962	70040	305112	42.5	1988
G10_cluster_1	100	0.41	1723545	19	20	120842	90712	304831	42.5	1860
G11_cluster_3	99.59	1.63	1774480	88	103	49219	20164	127344	42.6	1966
G12_cluster_3	100	0.81	1711966	70	81	56736	24456	167354	42.5	1892
G13_cluster_1	100	1.02	1713166	29	29	111301	59074	204743	42.6	1862
G14_cluster_1	100	2.03	1832342	84	95	51822	21813	172128	42.5	2040
G15_cluster_2	99.19	0.41	1739018	74	86	46164	23500	163328	43.1	1932
G16_cluster_2	99.59	0	1673288	61	71	61534	27430	163089	43.2	1852
G17_cluster_2	99.59	0.41	1721654	47	47	104073	36630	192643	43.1	1894

G18_cluster_2	99.59	0.41	1721633	46	46	120290	37426	223434	43.1	1893
G20_cluster_2	94.85	1.02	1727310	228	312	25690	7575	67915	43.1	2033
G21_cluster_1	99.59	0.41	1727168	52	56	67638	33214	223434	43.1	1906
G22_cluster_2	100	0.41	1731371	85	100	44205	20369	103376	42.6	1910
G23_cluster_1	96.14	1.83	1723382	215	302	32920	8015	92717	42.6	1996
G24_cluster_1	97.15	5.28	1812616	259	334	27056	6998	87265	42.4	2089
G25_cluster_1	100	0.41	1697536	16	17	160555	106096	335334	47.6	1910
G25_cluster_3	100	0.61	1730865	67	79	59951	25833	245730	42.6	1899
G26_cluster_1	100	0.41	1781960	23	24	114728	77476	305086	42.5	1946
G27_cluster_2	100	0.41	1753546	28	28	114682	62626	305078	42.5	1891
G28_cluster_2	100	0.81	1725753	45	48	76032	38350	198917	42.6	1871
G29_cluster_1	99.59	0	1676399	56	57	58141	29935	174085	43.2	1839
G30_cluster_2	100	0.41	1694882	16	18	173962	105930	282270	47.6	1899

6.3. SUPPLEMENTARY 3. Pairwise ANI Score of Aquificota genomes

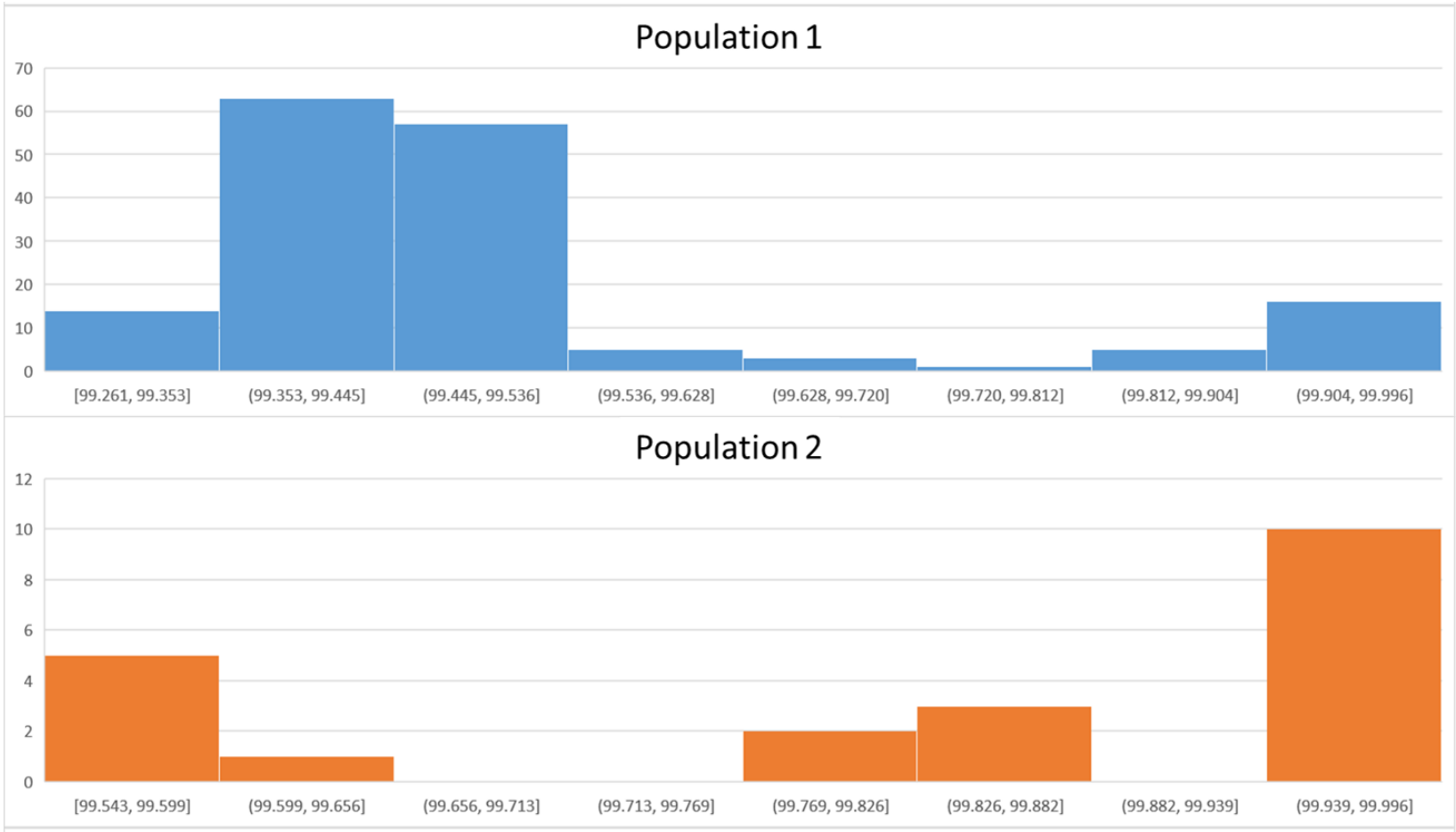

Figure 6.2 Intrapopulation heterogeneity in Population 1 and 2. Even with ANI values for both Populations 1 and 2 at >99%, Population 1 shows higher heterogeneity, with most pairwise genome comparisons having ANI values <99.536% (median 99.451%), while Population 2 has most pairwise ANI values at >99.769% (median 99.857%). Population 1 has a median value of 99.451%, while Population 2 is 99.857%.

6.4. SUPPLEMENTARY 4. Comparison of diversity from different dilution methods

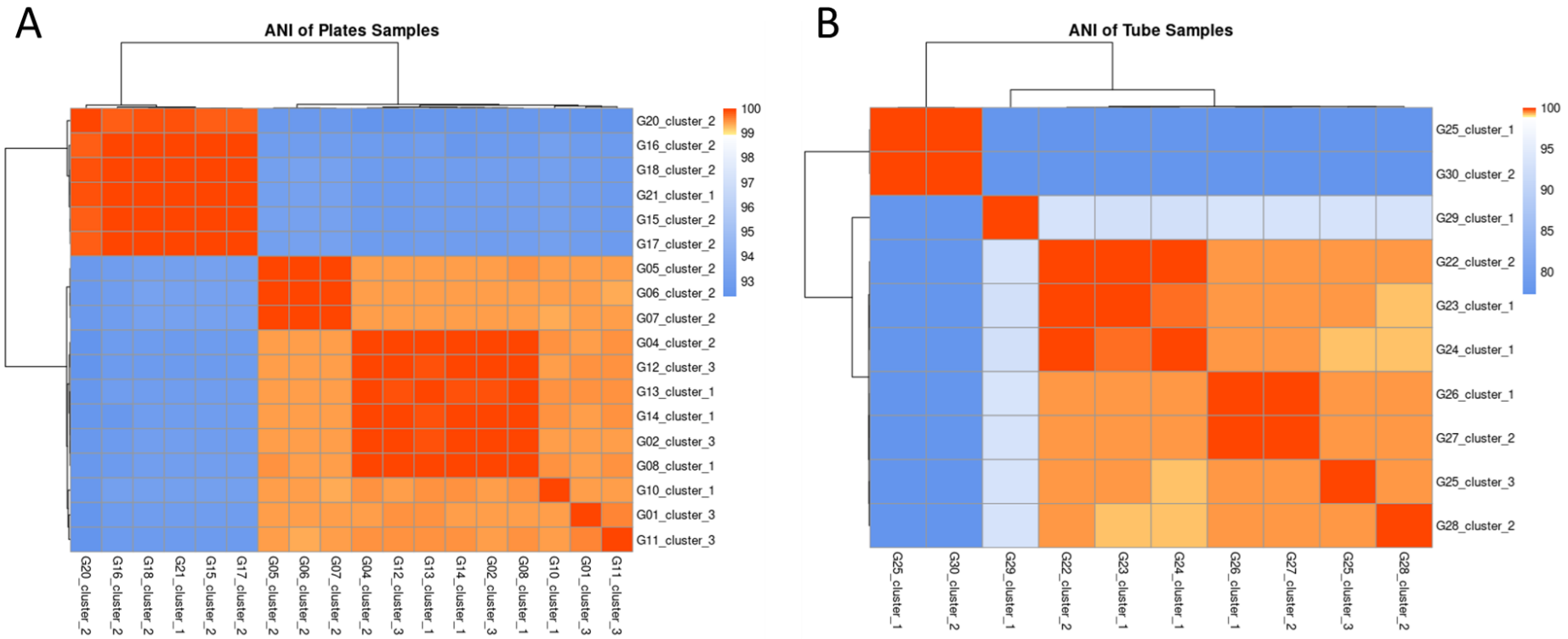

Figure 6.3 Effect of different serial dilution-to-extinction methods in genome diversity. The two serial dilution methods – plates (A) and tubes (B) – vary with the volume used to inoculate and perform the dilutions at 2 mL and 10 mL, respectively. Both techniques capture similar diversity of Population 1 as several distinct clonal complexes can be seen in both methods. Population 3 (G25 cluster 1 and G30 cluster 2) is only captured by tubes which could be a product of the low sampling depth.

6.5. SUPPLEMENTARY 5. Comparison of length bias between different SAG completeness

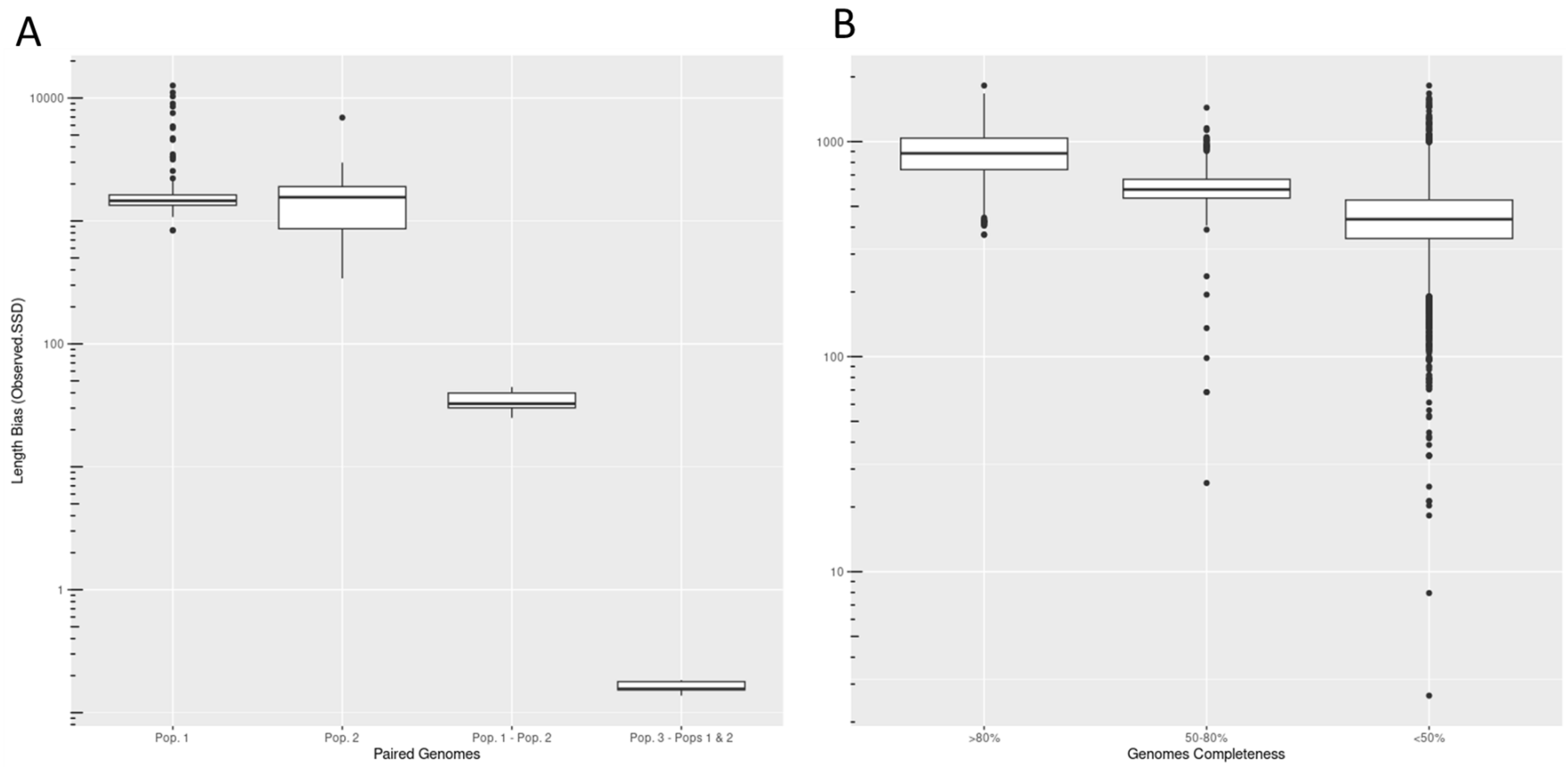

Figure 6.4 Comparison of observed length bias. (A) The observed length bias within populations (Population 1: 1880.46 ± 1693.73; Population 2: 1748.70 ± 1378.24) are significantly higher than those comparisons between populations – intra-cluster (Population 1 &2). The length bias across different completeness of SAGs varies with >80% completeness (901.44 ± 293.27), SAG with 50-80% completeness (613.52 ± 106.74) and SAG with <50% completeness (454.97 ± 171.11)

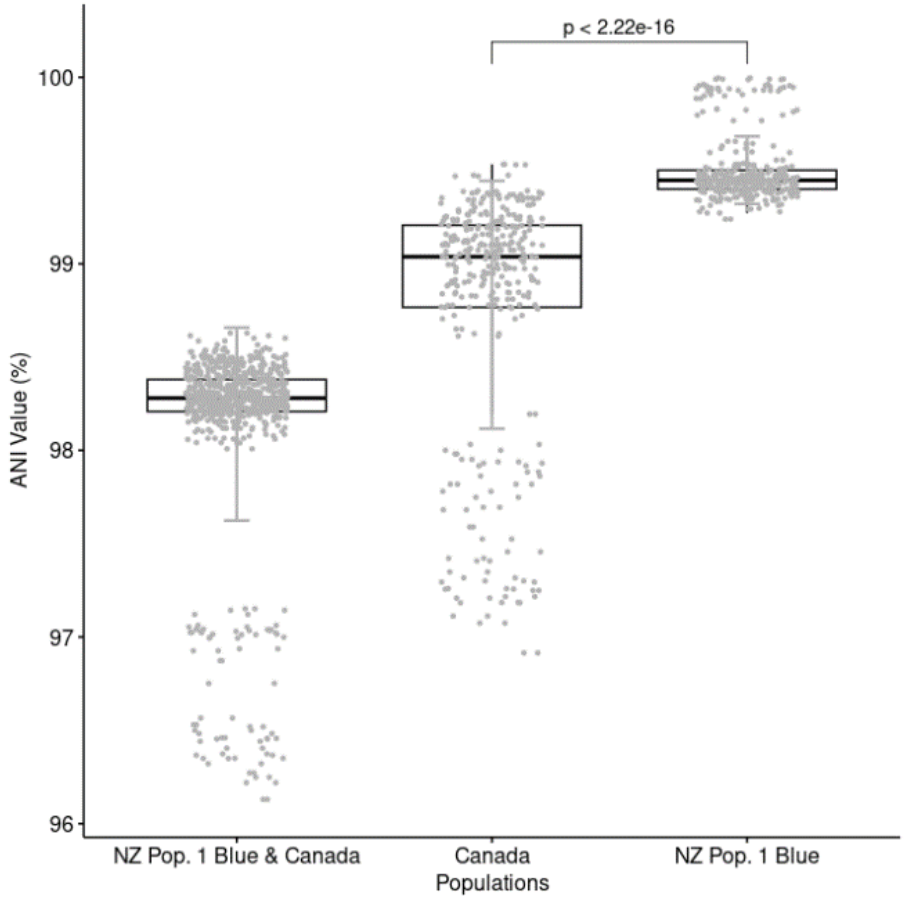

Figure 6.5 Comparison of ANI between Canada and New Zealand isolates. Canadian isolates shows higher diversity between genomes and has ANI values that are more dispersed than New Zealand isolates. There is significant differences in the ANI values computed between the two geographical populations. Statistical analysis done using Wilcoxon rank sum test. Mean values and error (standard deviation) are denoted by square dot and grey bar, respectively.

6.6. SUPPLEMENTARY 6. GTDB Classification

Table 6.3 Pplacer ANI and GTDB Classification of Kuirau Park Isolates

Genome	Classification	Closest Placement Reference	Closest Placement ANI
G01_cluster_3	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.87
G02_cluster_3	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.78
G04_cluster_2	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.79
G05_cluster_2	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.82
G06_cluster_2	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.9
G07_cluster_2	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.87
G08_cluster_1	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.81
G10_cluster_1	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.83
G11_cluster_3	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.75
G12_cluster_3	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.77
G13_cluster_1	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.8
G14_cluster_1	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.72
G15_cluster_2	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp003534055	GCA_003534055.1	99.09

G16_cluster_2	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp003534055	GCA_003534055.1	99.07
G17_cluster_2	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp003534055	GCA_003534055.1	99.11
G18_cluster_2	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp003534055	GCA_003534055.1	99.08
G20_cluster_2	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp003534055	GCA_003534055.1	98.96
G21_cluster_1	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp003534055	GCA_003534055.1	99.1
G22_cluster_2	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.8
G23_cluster_1	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.76
G24_cluster_1	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.73
G25_cluster_1	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__	GCF_011006175.1	79.29
G25_cluster_3	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.79
G26_cluster_1	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.81
G27_cluster_2	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.78
G28_cluster_2	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp011054805	GCA_011054805.1	98.73
G29_cluster_1	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__UBA11096 sp003534055	GCA_003534055.1	98.99
G30_cluster_2	d__Bacteria; p__Aquificota; c__Aquificae; o__Aquificales; f__Aquificaceae; g__UBA11096; s__	GCF_011006175.1	79.25

APPENDIX B - CHAPTER 3

7.1. SUPPLEMENTARY 7. DNA Extraction Buffer

Table 7.1 Reagents for the modified metagenomic extraction buffer for in-filter extraction

Reagents	**Units**	**Concentration**
Lysozyme	mg/mL	50
Proteinase K	mg/mL	20
Sodium dodecyl sulphate	%	20
Sodium Acetate pH 5.2	M	3
DNA Extraction Buffer (45mL)		
Buffer Component	**Units**	**Weight**
Tris-HCl (pH8)	g	0.71
Sodium-EDTA (pH8)	g	1.68
NaH2PO4	g	0.54
Sodium Chloride	g	3.95
1%CTAB	g	0.45

7.2. SUPPLEMENTARY 8. Intrapopulation Mapping of Genomic Reads

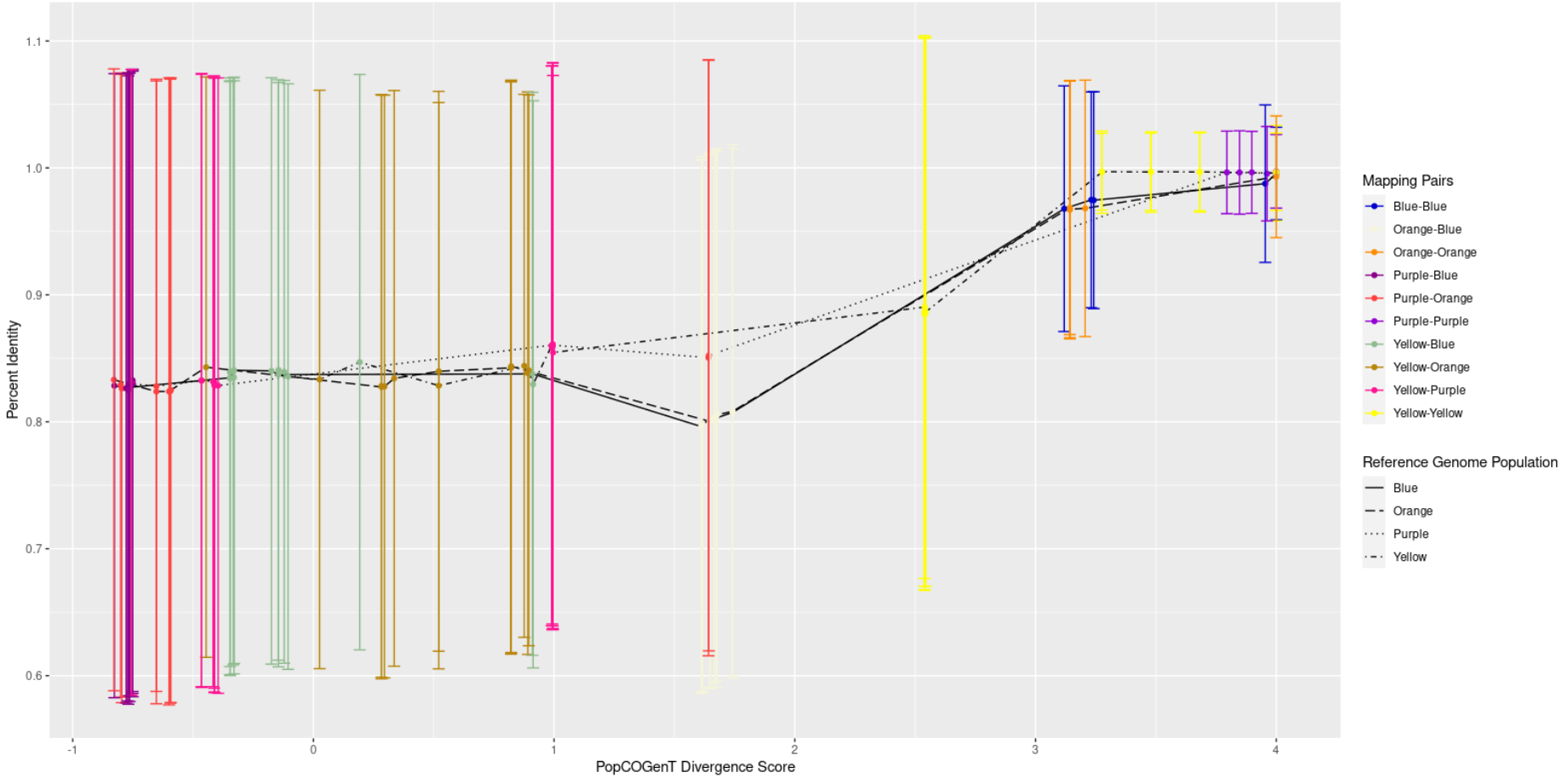

Figure 7.1 Identification of metagenomic reads cutoff for genome mapping. Reads of the assembled genomes were mapped against a reference genome from each population. The mean percentage identity of each mapping was compared with the PopCOGenT divergence score of the read-genome pair. Higher PopCOGenT divergence score implies stronger recombination between paired genomes.

7.3. SUPPLEMENTARY 9. Analysis of the diversity and environmental factors

Table 7.2 Analysis of Variance between temperature, pH and water level

Response Variable: Temperature

	Df	Sum Sq	Mean Sq	F	Value	Pr(>F)
Water.Level	1	48.4	48.401	0.3495	0.5862	
Residuals	4	553.98	138.495			

Response Variable: pH

	Df	Sum Sq	Mean Sq	F	Value	Pr(>F)
Water.Level	1	0.00296	0.002956	0.0254	0.8811	
Residuals	4	0.46549	0.116373			

Response Variable: pH

	Df	Sum Sq	Mean Sq	F	Value	Pr(>F)
Temperature	1	0.00053	0.000529	0.0045	0.9496	
Residuals	4	0.46792	0.116979			

Table 7.3 PerMANOVA test of microbial composition within Site, Time point and Water Level Number of permutations: 999

	Df	SumOfSqs	R2	F	Pr(>F)	
Site	2	247.54	0.35093	2.5964	0.006	**
Time Points	3	241.33	0.34213	1.6875	0.01	**
Water.Level	1	73.5	0.10419	1.5417	0.147	
Residual	3	143.01	0.20274			
Total	9	705.39	1			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 7.4 Pairwise PerMANOVA test of microbial composition within a Site or Time point

Site

	Groups	measure	F	R2	p.value	p.adjusted(fdr)	Significance
1	KRP1 vs KRP2	robust.aitchison	1.023669	0.145746	0.397	0.397	
2	KRP1 vs KRP2B	robust.aitchison	2.483351	0.383035	0.066667	0.1	
3	KRP2 vs KRP2B	robust.aitchison	2.222206	0.357141	0.066667	0.1	

Time point

	Groups	measure	F	R2	p.value	p.adjusted	Significance
1	Jan 20 vs Mar 20	robust.aitchison	1.256441	0.385833	0.333333	0.5	
2	Jan 20 vs Jul 20	robust.aitchison	0.802878	0.211124	0.9	0.9	
3	Jan 20 vs Sep 20	robust.aitchison	2.109469	0.412855	0.1	0.3	
4	Mar 20 vs Jul 20	robust.aitchison	1.057344	0.2606	0.3	0.5	
5	Mar 20 vs Sep 20	robust.aitchison	1.510817	0.334932	0.1	0.3	
6	Jul 20 vs Sep 20	robust.aitchison	1.161572	0.225042	0.5	0.6	

Table 7.5 Mantel's test of the correlation between environmental factors and microbial community composition

Variables	Test	Correlation Method	Correlation Coefficient	p.value	p.adjusted (fdr)	Significance
pH	mantel	pearson	-0.130	0.734	0.798	
Temperature	mantel	pearson	0.005	0.424	0.530	
Chloride	mantel	pearson	0.154	0.114	0.231	
Sulfate	mantel	pearson	0.096	0.352	0.489	
Nitrate.N	mantel	pearson	-0.022	0.525	0.625	
Nitrite.N	mantel	pearson	0.693	0.078	0.231	
Ammonia.as.N	mantel	pearson	0.119	0.199	0.332	
DRP.Phosphorus.	mantel	pearson	0.683	0.060	0.231	
Arsenic	mantel	pearson	0.717	0.071	0.231	
Aluminium	mantel	pearson	0.401	0.119	0.231	
Barium	mantel	pearson	0.707	0.120	0.231	
Beryllium	mantel	pearson	-0.151	0.732	0.798	
Boron	mantel	pearson	0.688	0.080	0.231	
Calcium	mantel	pearson	0.714	0.102	0.231	
Caesium	mantel	pearson	0.737	0.029	0.231	
Copper	mantel	pearson	0.048	0.372	0.489	
Germanium	mantel	pearson	0.526	0.017	0.231	
Potassium	mantel	pearson	0.728	0.068	0.231	
Lithium	mantel	pearson	0.692	0.120	0.231	
Magnesium	mantel	pearson	-0.287	0.817	0.851	
Manganese	mantel	pearson	-0.431	0.966	0.966	
Sodium	mantel	pearson	0.316	0.292	0.429	
Rubidium	mantel	pearson	0.104	0.277	0.429	
Antimony	mantel	pearson	0.730	0.055	0.231	

Strontium	mantel	pearson	0.282	0.143	0.255	

Table 7.6 Mantel's test of the correlation between environmental factors and abundance of members of phylum *Aquificota*

Variables	Test	Correlation method	Correlation coefficient	p.value	p.adjusted (fdr)	Significance
Temperature	mantel	pearson	0.675556	0.023	0.41	
Barium	mantel	pearson	0.277703	0.036	0.41	
Nitrate	mantel	pearson	0.741116	0.057	0.41	
Magnesium	mantel	pearson	0.267838	0.068	0.41	
Chloride	mantel	pearson	0.175201	0.082	0.41	
Sulfate	mantel	pearson	0.302013	0.115	0.479167	
Calcium	mantel	pearson	0.207677	0.205	0.659375	
Strontium	mantel	pearson	0.17005	0.211	0.659375	
Manganese	mantel	pearson	0.053101	0.278	0.698913	
Copper	mantel	pearson	0.099344	0.292	0.698913	
Lithium	mantel	pearson	-0.07293	0.313	0.698913	
Boron	mantel	pearson	-0.08897	0.347	0.698913	
Sodium	mantel	pearson	-0.10723	0.374	0.698913	
Nitrite	mantel	pearson	-0.11885	0.438	0.698913	
Potassium	mantel	pearson	-0.12767	0.457	0.698913	
Arsenic	mantel	pearson	-0.00393	0.485	0.698913	
Aluminium	mantel	pearson	-0.12734	0.502	0.698913	
Germanium	mantel	pearson	-0.1369	0.54	0.698913	
Beryllium	mantel	pearson	-0.14309	0.566	0.698913	
Caesium	mantel	pearson	-0.15342	0.566	0.698913	
Phosphorus	mantel	pearson	-0.16559	0.631	0.698913	
pH	mantel	pearson	-0.08717	0.64	0.698913	
Rubidium	mantel	pearson	-0.16721	0.643	0.698913	
Antimony	mantel	pearson	-0.20174	0.721	0.751042	
Ammonia	mantel	pearson	-0.27375	0.896	0.896	

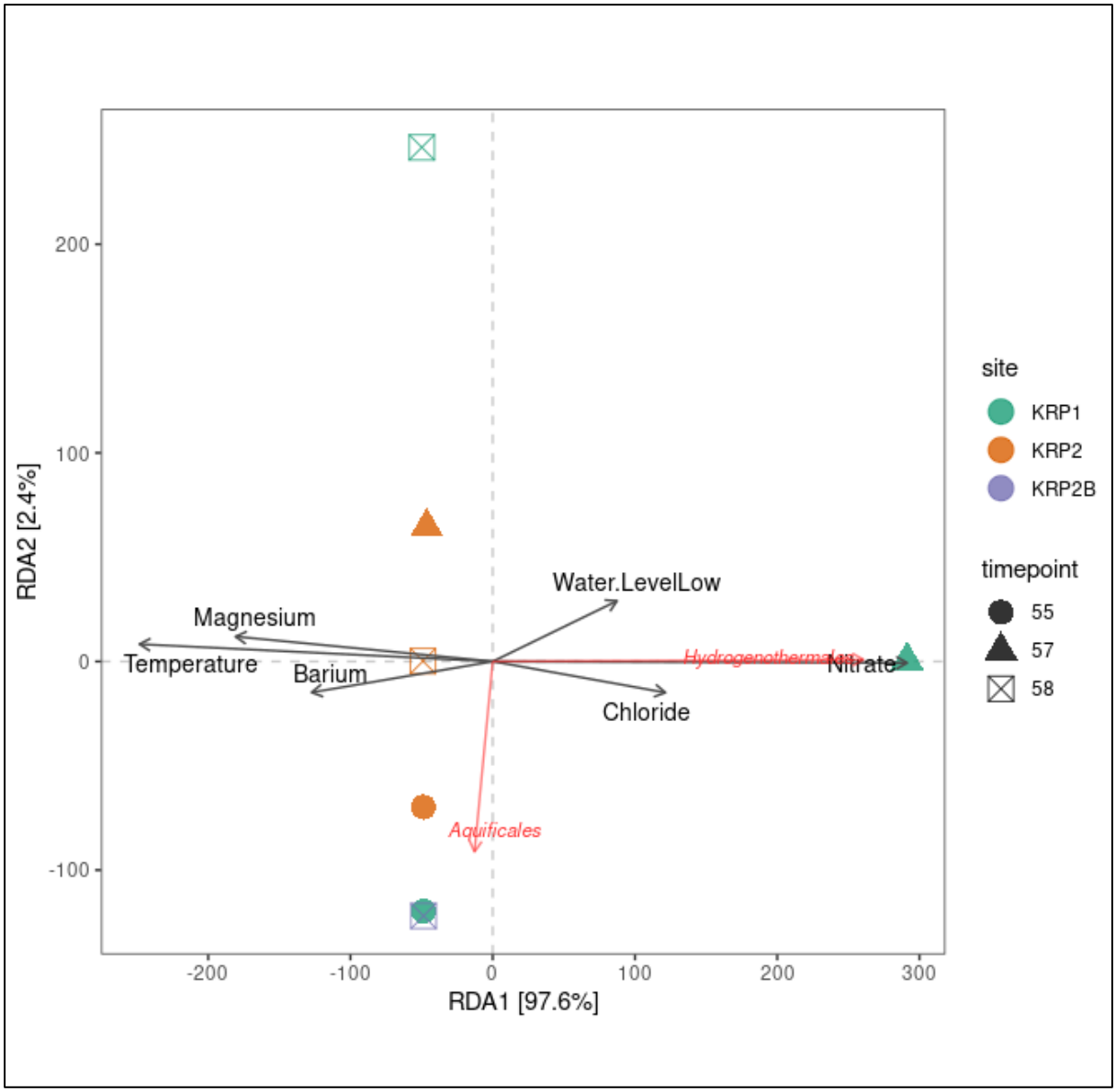

Figure 7.2 Distance-based Redundancy Analysis of the Aquificota members and selected environmental parameters. Distances were measured using the robust Aitchinson of the centre log-ratio transformed read abundance measured from containment queries of Genbank representative genomes. The mantel test without p-value adjustment shows temperature and barium as most associated with the abundance of H*ydrogenothermales*. A potential positive relationship between *Aquificales* and water level was observed

7.4. SUPPLEMENTARY 10. GTDB Classification

Table 7.7 GTDB, ANI, and Cultivation metadata of sequenced *Aquificota* genomes

Genome	Date	Site	Clusters	Media	Population	Clonal Group	*GTDB Classification*	Reference	ANI
SE046_G01_cluster_3	Jul-19	KRP1	SE046_0	MOPSO	Blue	A	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.87
SE046_G02_cluster_3	Jul-19	KRP1	SE046_0	MOPSO	Blue	E	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.78
SE046_G04_cluster_2	Jul-19	KRP1	SE046_0	MOPSO	Blue	E	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.78
SE046_G05_cluster_2	Jul-19	KRP1	SE046_0	MOPSO	Blue	B	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.81
SE046_G06_cluster_2	Jul-19	KRP1	SE046_0	MOPSO	Blue	B	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.88
SE046_G07_cluster_2	Jul-19	KRP1	SE046_0	MOPSO	Blue	B	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.85
SE046_G08_cluster_1	Jul-19	KRP1	SE046_0	MOPSO	Blue	E	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.75
SE046_G10_cluster_1	Jul-19	KRP1	SE046_0	MOPSO	Blue	C	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.83
SE046_G11_cluster_3	Jul-19	KRP1	SE046_0	MOPSO	Blue	D	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.75
SE046_G12_cluster_3	Jul-19	KRP1	SE046_0	MOPSO	Blue	E	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.77
SE046_G13_cluster_1	Jul-19	KRP1	SE046_0	MOPSO	Blue	E	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.79
SE046_G15_cluster_2	Jul-19	KRP2	SE046_0.1	MOPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.46
SE046_G16_cluster_2	Jul-19	KRP2	SE046_0.1	MOPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.51

SE046_G17_cluster_2	Jul-19	KRP2	SE046_0.1	MOPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.49
SE046_G18_cluster_2	Jul-19	KRP2	SE046_0.1	MOPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.5
SE046_G20_cluster_2	Jul-19	KRP2	SE046_0.1	MOPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.42
SE046_G21_cluster_1	Jul-19	KRP2	SE046_0.1	MOPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.52
SE046_G22_cluster_2	Jul-19	KRP1	SE046_0	MOPSO	Blue	F	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.84
SE046_G23_cluster_1	Jul-19	KRP1	SE046_0	MOPSO	Blue	F	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.8
SE046_G25_cluster_1	Jul-19	KRP1	SE046_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.29
SE046_G25_cluster_3	Jul-19	KRP1	SE046_0	MOPSO	Blue	G	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.77
SE046_G26_cluster_1	Jul-19	KRP1	SE046_0	MOPSO	Blue	H	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.81
SE046_G27_cluster_2	Jul-19	KRP1	SE046_0	MOPSO	Blue	H	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.78
SE046_G28_cluster_2	Jul-19	KRP2	SE046_0	MOPSO	Blue	R	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.75
SE046_G29_cluster_1	Jul-19	KRP2	SE046_0.1	MOPSO	Orange	L	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.47
SE046_G30_cluster_2	Jul-19	KRP2	SE046_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.25
SE053_G27_cluster_3	Nov-19	KRP1	SE053_0	MOPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	99.03
SE053_G36_cluster_2	Nov-19	KRP2	SE053_0	MOPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	99.02

SE053_G37_cluster_2	Nov-19	KRP2	SE053_0	MOPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	98.99
SE053_G39_cluster_2	Nov-19	KRP2	SE053_0	MOPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	99.03
SE053_G44_cluster_2	Nov-19	KRP2	SE053_0	MOPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	98.95
SE055_G01_cluster_2_cluster_2	Jan-20	KRP1	SE055_0.1	HEPPSO	Orange	I	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.47
SE055_G02_cluster_1	Jan-20	KRP1	SE055_0.1	HEPPSO	Orange	I	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.41
SE055_G02_cluster_2	Jan-20	KRP1	SE055_0	HEPPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	98.98
SE055_G05_cluster_3	Jan-20	KRP1	SE055_0	HEPPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	98.96
SE055_G05_cluster_4	Jan-20	KRP1	SE055_0.1	HEPPSO	Orange	I	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.45
SE055_G06_cluster_1	Jan-20	KRP1	SE055_0.1	HEPPSO	Orange	I	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.41
SE055_G06_cluster_2	Jan-20	KRP1	SE055_0	HEPPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	98.98
SE055_G08_cluster_3	Jan-20	KRP1	SE055_0.1	HEPPSO	Orange	I	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.38
SE055_G09_cluster_2	Jan-20	KRP1	SE055_0.1	HEPPSO	Orange	I	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.41
SE055_G09_cluster_3	Jan-20	KRP1	SE055_0	HEPPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	99.04

SE055_G10_cluster_1	Jan-20	KRP1	SE055_0.1	HEPPSO	Orange	I	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.44
SE055_G10_cluster_2	Jan-20	KRP1	SE055_1	HEPPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.15
SE055_G12_cluster_1	Jan-20	KRP1	SE055_0.1	HEPPSO	Orange	I	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.34
SE055_G12_cluster_2_cluster_1	Jan-20	KRP1	SE055_1	HEPPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.21
SE055_G12_cluster_2_cluster_3	Jan-20	KRP1	SE055_0	HEPPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	99.04
SE055_G15_cluster_1	Jan-20	KRP2B	SE055_0.1	HEPPSO	Orange	I	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.39
SE055_G15_cluster_3	Jan-20	KRP2B	SE055_1	HEPPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.14
SE055_G17_cluster_1	Jan-20	KRP2B	SE055_0.1	HEPPSO	Orange	I	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.47
SE055_G18_cluster_2	Jan-20	KRP2B	SE055_0.1	HEPPSO	Orange	I	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.37
SE055_G20_cluster_1	Jan-20	KRP2B	SE055_1	HEPPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.23
SE055_G20_cluster_2	Jan-20	KRP2B	SE055_0.1	HEPPSO	Orange	I	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.47
SE055_G22_cluster_1	Jan-20	KRP2B	SE055_0.1	HEPPSO	Orange	I	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.44
SE055_G24_cluster_2	Jan-20	KRP2B	SE055_0.1	HEPPSO	Orange	I	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.42
SE055_G24_cluster_3	Jan-20	KRP2B	SE055_0	HEPPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	99
SE055_G32_cluster_2	Jan-20	KRP2B	SE055_1	HEPPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.2

SE055_G33_cluster_2	Jan-20	KRP2B	SE055_1	HEPPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.29
SE057_G03_cluster_1	Mar-20	KRP1	SE057_0	MOPSO	*Venenivibrio* sp.	N	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__;s__*	N/A	N/A
SE057_G05_cluster_2	Mar-20	KRP1	SE057_0	MOPSO	*Venenivibrio* sp.	N	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__;s__*	N/A	N/A
SE057_G07_cluster_1	Mar-20	KRP1	SE057_0	MOPSO	*Venenivibrio* sp.	N	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__;s__*	N/A	N/A
SE057_G07_cluster_2	Mar-20	KRP1	SE057_0	MOPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	99.04
SE057_G10_cluster_1	Mar-20	KRP1	SE057_0	MOPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	99.02
SE057_G10_cluster_2	Mar-20	KRP1	SE057_0	MOPSO	*Venenivibrio* sp.	N	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__;s__*	N/A	N/A
SE057_G12_cluster_2	Mar-20	KRP1	SE057_0	MOPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	99.03
SE057_G17_cluster_1	Mar-20	KRP1	SE057_0	MOPSO	*Hydrogenobacter thermophilus*	S	*o__Aquificales;f__Aquificaceae;g__Hydrogenobacter;s__Hydrogenobacter thermophilus*	GCF_000010785.1	98.52
SE057_G23_cluster_1	Mar-20	KRP1	SE057_0	MOPSO	*Hydrogenobacter thermophilus*	S	*o__Aquificales;f__Aquificaceae;g__Hydrogenobacter;s__Hydrogenobacter thermophilus*	GCF_000010785.1	98.52
SE057_G27_cluster_2	Mar-20	KRP2	SE057_0	MOPSO	*Hydrogenobacter thermophilus*	S	*o__Aquificales;f__Aquificaceae;g__Hydrogenobacter;s__Hydrogenobacter thermophilus*	GCF_000010785.1	98.5
SE057_G28_cluster_3	Mar-20	KRP2	SE057_0	MOPSO	*Hydrogenobacter thermophilus*	S	*o__Aquificales;f__Aquificaceae;g__Hydrogenobacter;s__Hydrogenobacter thermophilus*	GCF_000010785.1	98.52
SE057_G30_cluster_2	Mar-20	KRP2	SE057_0	MOPSO	*Hydrogenobacter thermophilus*	S	*o__Aquificales;f__Aquificaceae;g__Hydrogenobacter;s__Hydrogenobacter thermophilus*	GCF_000010785.1	98.47

SE057_G31_cluster_2	Mar-20	KRP2	SE057_0	MOPSO	*Hydrogenobacter thermophilus*	S	*o__Aquificales;f__Aquificaceae;g__Hydrogenobacter;s__Hydrogenobacter thermophilus*	GCF_000010785.1	98.49
SE057_G34_cluster_1	Mar-20	KRP2	SE057_0	MOPSO	*Hydrogenobacter thermophilus*	S	*o__Aquificales;f__Aquificaceae;g__Hydrogenobacter;s__Hydrogenobacter thermophilus*	GCF_000010785.1	98.52
SE057_G42_cluster_2	Mar-20	KRP2	SE057_0.1	MOPSO	Orange	T	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.57
SE057_G48_cluster_2	Mar-20	KRP2	SE057_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.27
SE057_G49_cluster_3	Mar-20	KRP2B	SE057_0	MOPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	98.99
SE057_G50_cluster_3	Mar-20	KRP2B	SE057_0	MOPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	99
SE057_G51_cluster_3	Mar-20	KRP2B	SE057_0	MOPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	99.02
SE057_G52_cluster_2	Mar-20	KRP2B	SE057_0	MOPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	98.99
SE057_G53_cluster_2	Mar-20	KRP2B	SE057_0	MOPSO	*Sulfurihydrogenibium* sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	98.95
SE057_G54_cluster_2	Mar-20	KRP2B	SE057_0	MOPSO	Sulfurihydrogenibium sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	99.02
SE057_G56_cluster_3	Mar-20	KRP2B	SE057_0	MOPSO	Sulfurihydrogenibium sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	98.99
SE057_G57_cluster_3	Mar-20	KRP2B	SE057_0	MOPSO	Sulfurihydrogenibium sp.	P	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__Sulfurihydrogenibium;s__Sulfurihydrogenibium azorense*	GCF_000021545.1	98.96

SE057_G71_cluster_2	Mar-20	KRP1	SE057_0	MOPSO_MALT	Venenivibrio sp.	N	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__;s__*	N/A	N/A
SE057_G72_cluster_2	Mar-20	KRP1	SE057_0	MOPSO_MALT	Venenivibrio sp.	N	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__;s__*	N/A	N/A
SE057_G76_cluster_2	Mar-20	KRP1	SE057_0	HEPPSO/ POPSO_YE	Hydrogenobacter thermophilus	S	*o__Aquificales;f__Aquificaceae;g__Hydrogenobacter;s__Hydrogenobacter thermophilus*	GCF_000010785.1	98.45
SE057_G82_cluster_1	Mar-20	KRP1	SE057_0	HEPPSO/ POPSO_MALT	Hydrogenobacter thermophilus	S	*o__Aquificales;f__Aquificaceae;g__Hydrogenobacter;s__Hydrogenobacter thermophilus*	GCF_000010785.1	98.46
SE057_G83_cluster_3	Mar-20	KRP1	SE057_0	HEPPSO/ POPSO_MALT	Hydrogenobacter thermophilus	S	*o__Aquificales;f__Aquificaceae;g__Hydrogenobacter;s__Hydrogenobacter thermophilus*	GCF_000010785.1	98.47
SE058_G100_cluster_1	Jul-20	KRP2B	SE058_1	MOPSO	Venenivibrio sp.	Q	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__;s__*	N/A	N/A
SE058_G101_cluster_2	Jul-20	KRP2B	SE058_1	MOPSO	Venenivibrio sp.	Q	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__;s__*	N/A	N/A
SE058_G102_cluster_3	Jul-20	KRP2B	SE058_1	MOPSO	Venenivibrio sp.	Q	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__;s__*	N/A	N/A
SE058_G103_cluster_2	Jul-20	KRP2B	SE058_1	MOPSO	Venenivibrio sp.	Q	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__;s__*	N/A	N/A
SE058_G103_cluster_3	Jul-20	KRP2B	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.29
SE058_G107_cluster_1	Jul-20	KRP2B	SE058_1	MOPSO	Venenivibrio sp.	Q	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__;s__*	N/A	N/A
SE058_G108_cluster_1	Jul-20	KRP2B	SE058_1	MOPSO	Venenivibrio sp.	Q	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__;s__*	N/A	N/A
SE058_G109_cluster_1	Jul-20	KRP2B	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.18
SE058_G110_cluster_2	Jul-20	KRP2B	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.36
SE058_G111_cluster_1	Jul-20	KRP2B	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.25

SE058_G112_cluster_2	Jul-20	KRP2B	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.14
SE058_G113_cluster_1	Jul-20	KRP2B	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.18
SE058_G114_cluster_2	Jul-20	KRP2B	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.22
SE058_G116_cluster_2	Jul-20	KRP2B	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.17
SE058_G117_cluster_1	Jul-20	KRP2B	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.33
SE058_G118_cluster_1	Jul-20	KRP2B	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.24
SE058_G119_cluster_2	Jul-20	KRP2B	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.41
SE058_G120_cluster_2	Jul-20	KRP2B	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.39
SE058_G121_cluster_1	Jul-20	KRP1	SE058_0.1	POPSO	Orange	J	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.5
SE058_G122_cluster_2	Jul-20	KRP1	SE058_0.1	POPSO	Orange	J	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.49
SE058_G125_cluster_2	Jul-20	KRP1	SE058_0.1	POPSO	Orange	J	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.48
SE058_G127_cluster_1	Jul-20	KRP1	SE058_0.1	POPSO	Orange	J	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.51
SE058_G128_cluster_2	Jul-20	KRP1	SE058_0.1	POPSO	Orange	J	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.49
SE058_G129_cluster_2	Jul-20	KRP1	SE058_0.1	POPSO	Orange	J	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.46
SE058_G130_cluster_1	Jul-20	KRP1	SE058_0.1	POPSO	Orange	J	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.52
SE058_G131_cluster_1	Jul-20	KRP1	SE058_0.1	POPSO	Orange	J	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.47

SE058_G132_cluster_2	Jul-20	KRP1	SE058_0.1	POPSO	Orange	J	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.46
SE058_G133_cluster_2	Jul-20	KRP2	SE058_0.1	POPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.54
SE058_G134_cluster_2	Jul-20	KRP2	SE058_0.1	POPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.5
SE058_G135_cluster_2	Jul-20	KRP2	SE058_0.1	POPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.5
SE058_G136_cluster_2	Jul-20	KRP2	SE058_0.1	POPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.53
SE058_G137_cluster_1	Jul-20	KRP2	SE058_0.1	POPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.54
SE058_G138_cluster_2	Jul-20	KRP2	SE058_0.1	POPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.52
SE058_G140_cluster_2	Jul-20	KRP2	SE058_0.1	POPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.58
SE058_G141_cluster_2	Jul-20	KRP2	SE058_0.1	POPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.46
SE058_G142_cluster_2	Jul-20	KRP2	SE058_0.1	POPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.53
SE058_G143_cluster_2	Jul-20	KRP2	SE058_0.1	POPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.53
SE058_G144_cluster_1	Jul-20	KRP2	SE058_0.1	POPSO	Orange	U	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.53
SE058_G151_cluster_3	Jul-20	KRP2B	SE058_0	POPSO	Blue	R	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	99.06
SE058_G156_cluster_3	Jul-20	KRP2B	SE058_0	POPSO	Blue	R	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.94
SE058_G173_cluster_1	Jul-20	KRP1	SE058_0.1	MOPSO_7.6	Orange	J	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.51
SE058_G173_cluster_2	Jul-20	KRP1	SE058_0	MOPSO_7.6	*Venenivibrio* sp.	K	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__;s__*	N/A	N/A

SE058_G173_cluster_4	Jul-20	KRP1	SE058_1	MOPSO_7.6	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.19
SE058_G174_cluster_2	Jul-20	KRP1	SE058_0.1	MOPSO_7.6	Orange	J	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.44
SE058_G174_cluster_3	Jul-20	KRP1	SE058_1	MOPSO_7.6	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.14
SE058_G175_cluster_1	Jul-20	KRP1	SE058_0.1	MOPSO_7.6	Orange	J	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.44
SE058_G175_cluster_2	Jul-20	KRP1	SE058_1	MOPSO_7.6	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.18
SE058_G176_cluster_2	Jul-20	KRP1	SE058_1	MOPSO_7.6	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.23
SE058_G176_cluster_3	Jul-20	KRP1	SE058_0	MOPSO_7.6	*Venenivibrio* sp.	K	*o__Hydrogenothermales;f__Hydrogenothermaceae;g__;s__*	N/A	N/A
SE058_G176_cluster_4	Jul-20	KRP1	SE058_0.1	MOPSO_7.6	Orange	J	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.43
SE058_G49_cluster_2	Jul-20	KRP1	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.28
SE058_G49_cluster_3	Jul-20	KRP1	SE058_0.1	MOPSO	Orange	L	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.41
SE058_G50_cluster_1	Jul-20	KRP1	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.22
SE058_G50_cluster_3	Jul-20	KRP1	SE058_0.1	MOPSO	Orange	L	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.47
SE058_G51_cluster_1	Jul-20	KRP1	SE058_0.1	MOPSO	Orange	L	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.48
SE058_G51_cluster_2	Jul-20	KRP1	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.2
SE058_G52_cluster_1	Jul-20	KRP1	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.23
SE058_G52_cluster_3	Jul-20	KRP1	SE058_0.1	MOPSO	Orange	L	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.44

SE058_G53_cluster_1	Jul-20	KRP1	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.28
SE058_G54_cluster_1	Jul-20	KRP1	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.18
SE058_G54_cluster_3	Jul-20	KRP1	SE058_0.1	MOPSO	Orange	L	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.46
SE058_G55_cluster_1	Jul-20	KRP1	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.22
SE058_G55_cluster_3	Jul-20	KRP1	SE058_0.1	MOPSO	Orange	L	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.41
SE058_G56_cluster_1	Jul-20	KRP1	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.11
SE058_G57_cluster_1	Jul-20	KRP1	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.18
SE058_G59_cluster_2	Jul-20	KRP1	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.12
SE058_G59_cluster_3	Jul-20	KRP1	SE058_0.1	MOPSO	Orange	L	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp003534055*	GCA_018771605.1	98.41
SE058_G60_cluster_2	Jul-20	KRP1	SE058_1	MOPSO	Green	O	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__*	GCF_011006175.1	79.11
SE058_G67_cluster_2	Jul-20	KRP1	SE058_0	MOPSO	Blue	M	*o__Aquificales;f__Aquificaceae;g__UBA11096;s__UBA11096 sp011054805*	GCA_011054805.1	98.76

7.5. SUPPLEMENTARY 11. Histogram of Intrapopulaiton ANI

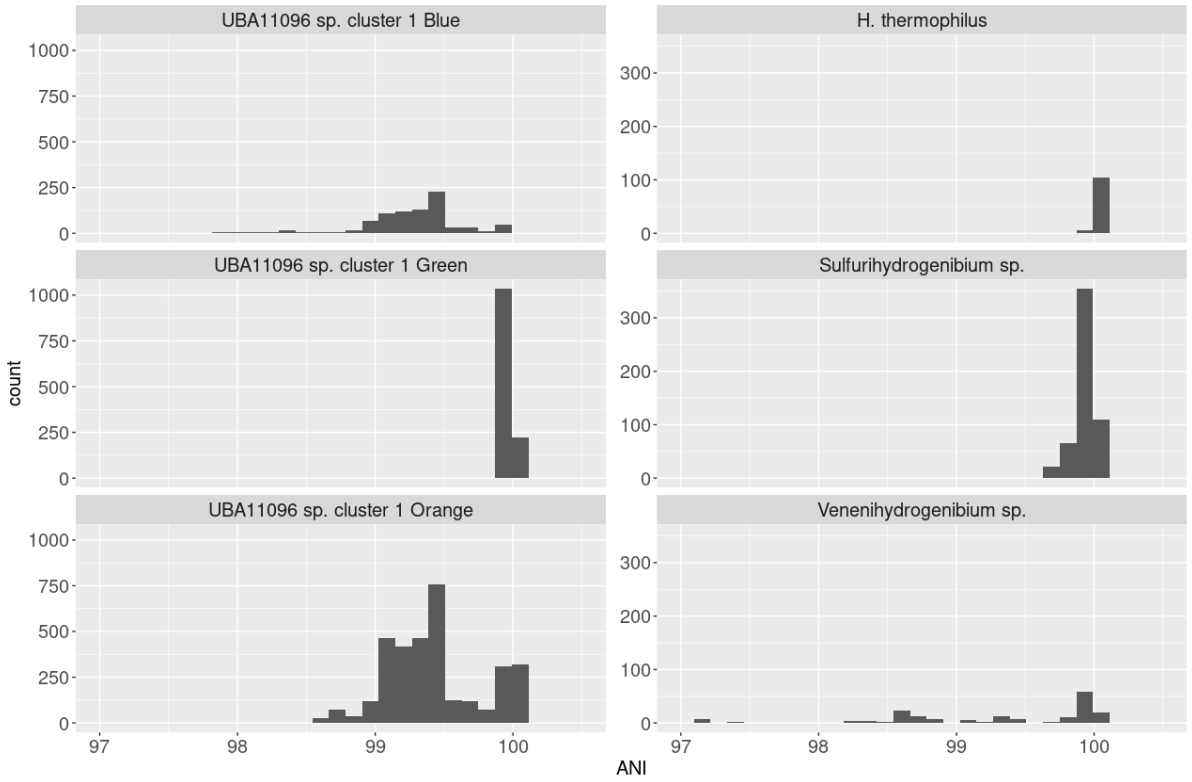

Figure 7.3 Histogram distribution of intrapopulaiton ANI values. Clonal populations *UBA11096* sp. cluster 1 Green, *H. thermophilus,* and *Sulfurhydrogenibium* sp. have median ANI values of 99.98%, 99.99% and 99.96%, respectively. A notable drop in ANI can be observed in *Venenivibrio,* and *UBA11096* sp. cluster 2 Blue and Orange populations at around 99.5% ANI similarity

7.6. SUPPLEMENTARY 12. Comparison of PopCOGenT analysis with varying contamination

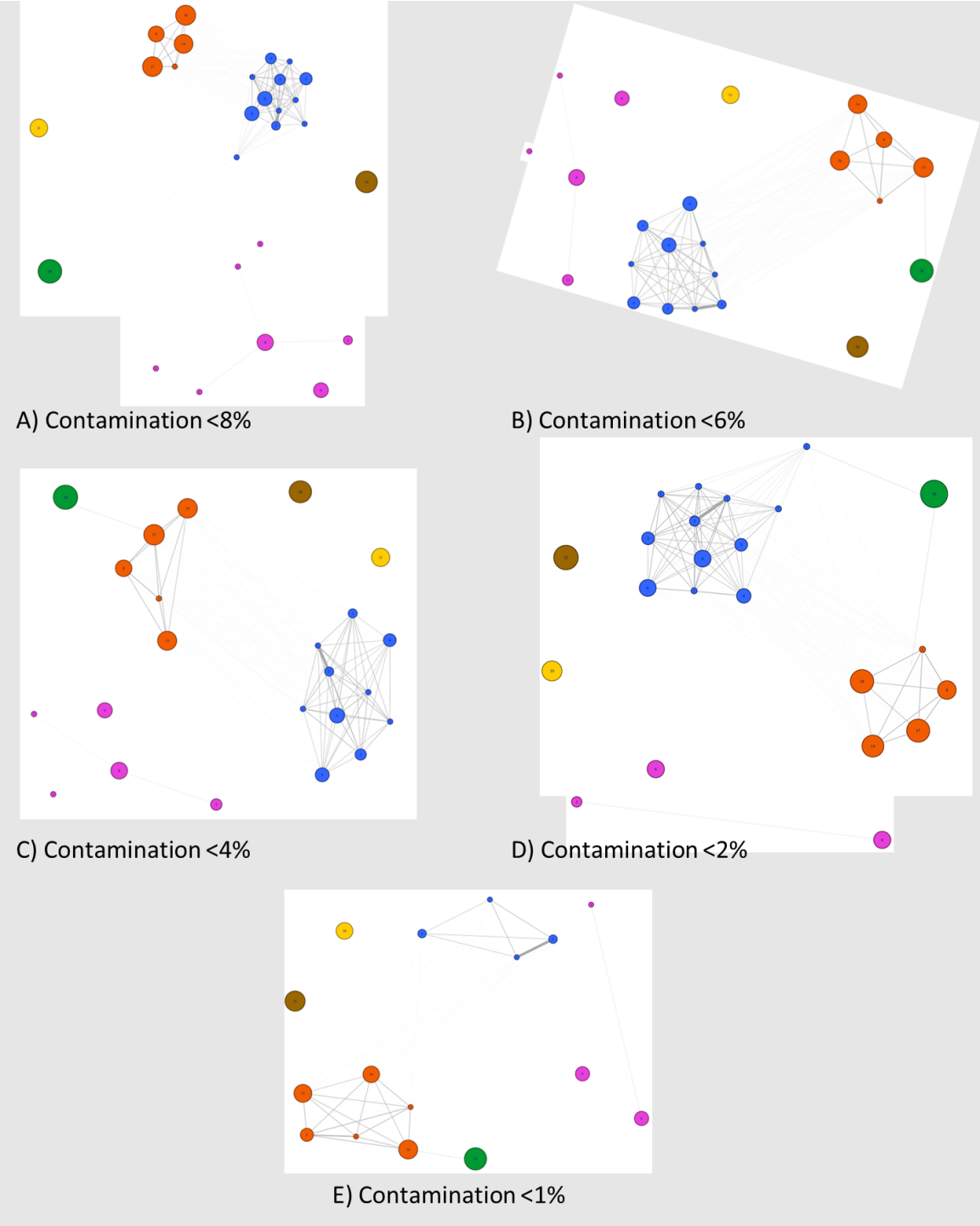

Figure 7.4 PopCOGenT clustering with genomes from different contamination cutoffs. Contamination was measured using CheckM and ranges from 0-8% (A-E). Genomes with minimal contamination of <2% were utilised to minimise the effect of contamination in genome comparisons during horizontal gene flow analyses.

7.7. SUPPLEMENTARY 13. PopCOGenT Metadata

Table 7.8 Summary of clonal group ID of Aquificota genomes from all sites and time

Clonal Group	Genomes per Group	Population	Date isolated	PopCOGenT cluster per time point	Sites isolated
A	1	Blue	Jul-19	SE046_0	KRP1
B	3	Blue	Jul-19	SE046_0	KRP1
C	1	Blue	Jul-19	SE046_0	KRP1
D	1	Blue	Jul-19	SE046_0	KRP1
E	5	Blue	Jul-19	SE046_0	KRP1
F	2	Blue	Jul-19	SE046_0	KRP1
G	1	Blue	Jul-19	SE046_0	KRP1
H	2	Blue	Jul-19	SE046_0	KRP1
M	1	Blue	Jul-20	SE058_0	KRP1
R	3	Blue	Jul-19	SE046_0	KRP2,KRP2B
R	3	Blue	Jul-20	SE058_0	KRP2,KRP2B
O	36	Green	Jul-19	SE046_1	KRP1,KRP2,KRP2B
O	36	Green	Jan-20	SE055_1	KRP1,KRP2,KRP2B
O	36	Green	Mar-20	SE057_1	KRP1,KRP2,KRP2B
O	36	Green	Jul-20	SE058_1	KRP1,KRP2,KRP2B
S	10	*Hydrogenobacter thermophilus*	Mar-20	SE057_0	KRP1,KRP2
I	14	Orange	Jan-20	SE055_0.1	KRP1,KRP2B
J	13	Orange	Jul-20	SE058_0.1	KRP1
L	8	Orange	Jul-19	SE046_0.1	KRP2,KRP1
L	8	Orange	Jul-20	SE058_0.1	KRP2,KRP1
T	1	Orange	Mar-20	SE057_0.1	KRP2
U	17	Orange	Jul-19	SE046_0.1	KRP2
U	17	Orange	Jul-20	SE058_0.1	KRP2
P	22	*Sulfurihydrogenibium sp.*	Nov-19	SE053_0	KRP1,KRP2,KRP2B
P	22	*Sulfurihydrogenibium sp.*	Jan-20	SE055_0	KRP1,KRP2,KRP2B
P	22	*Sulfurihydrogenibium sp.*	Mar-20	SE057_0	KRP1,KRP2,KRP2B
K	2	*Venenivibrio sp.*	Jul-20	SE058_0	KRP1
N	6	*Venenivibrio sp.*	Mar-20	SE057_0	KRP1
Q	6	*Venenivibrio sp.*	Jul-20	SE058_1	KRP2B

7.8. SUPPLEMENTARY 14. SNP and read abundance data per site

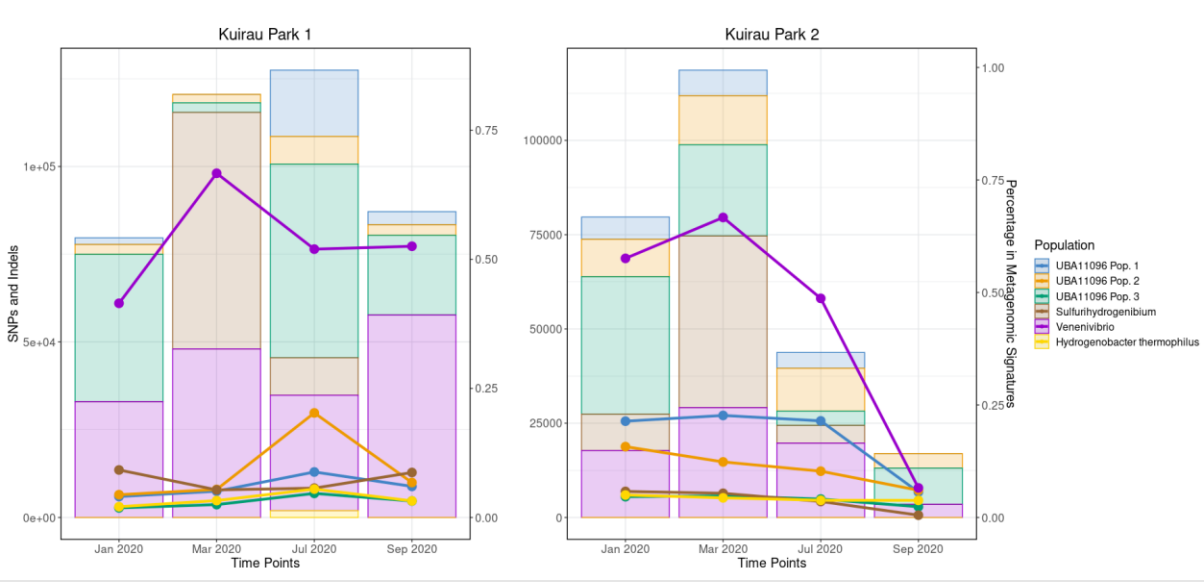

Figure 7.5 Site-specific data for read and SNP abundance. Low SNP variation (line) across time for clonal populations was observed for both sites. Intrapopulation diversity of the recombining populations differs between sites but exhibits a similar pattern that mirrors read abundance patterns (Bar).

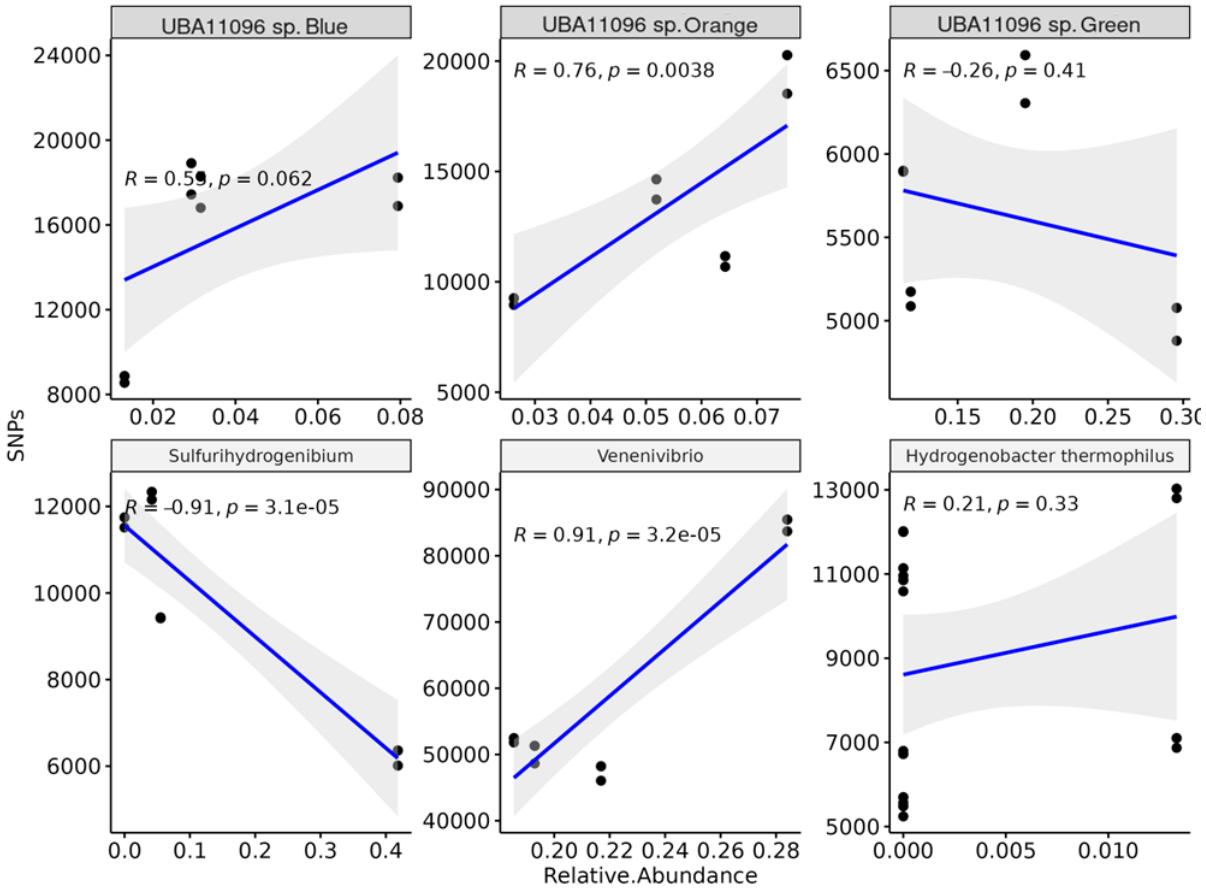

Figure 7.6 Pearson correlation between SNP counts and relative abundance of Aquificota populations. A positive correlation was observed in *Venenivibrio* sp., *UBA11096* sp. Blue and Orange populations. Meanwhile, a negative correlation was observed in *UBA11096* sp. Green and *Sulfurihydrogenibium* sp., while a weak correlation was observed in *Hydrogenobacter thermophilus*.

7.9. SUPPLEMENTARY 15. Distribution of COGs from swept regions in different COG Categories

Table 7.9 Distribution of swept genes in different COG Categories

Category	Detail	Flexible				Core			
		Blue		Orange		Blue		Orange	
		Jul-19	Jul-20	Jul-19	Jul-20	Jul-19	Jul-20	Jul-19	Jul-20
A	RNA processing and modification	0	0	0	0	0	0	0	0
B	Chromatin structure and dynamics	0	0	0	0	0	0	1	0
C	Energy production and conversion	11	36	1	1	1	4	15	0
D	Cell cycle control, cell division, chromosome partitioning	0	0	0	1	7	7	10	3
E	Amino acid transport and metabolism	3	9	0	0	3	7	17	1
F	Nucleotide transport and metabolism	1	3	0	1	2	10	11	2
G	Carbohydrate transport and metabolism	1	3	1	1	2	6	2	0
H	Coenzyme transport and metabolism	1	3	1	0	5	11	13	3
I	Lipid transport and metabolism	1	3	0	0	1	1	2	0
J	Translation, ribosomal structure and biogenesis	1	3	0	1	7	23	37	3
K	Transcription	2	6	1	0	2	10	9	1
L	Replication, recombination and repair	4	15	8	8	7	12	19	3
M	Cell wall/membrane/envelope biogenesis	6	15	0	3	8	15	26	1
N	Cell motility	7	6	3	3	10	14	21	4
O	Posttranslational modification, protein turnover, chaperones	4	6	0	0	9	10	23	0
P	Inorganic ion transport and metabolism	5	12	2	4	4	4	9	2
Q	Secondary metabolites biosynthesis, transport and catabolism	1	3	0	0	0	1	2	0
R	General function prediction only	0	0	0	0	0	0	0	0
S	Function unknown	16	57	4	4	14	25	41	13
T	Signal transduction mechanisms	11	12	3	2	2	11	9	3
U	Intracellular trafficking, secretion, and vesicular transport	2	3	3	5	9	10	19	5
V	Defense mechanisms	0	3	1	1	1	3	1	0
Z	Cytoskeleton	0	0	0	0	0	0	1	0
-	Uncharacterized Protein	4	9	3	3	2	2	4	3

Table 7.10 Wilcoxon Signed-Rank test to compare the similarity of COG categories between populations

Genome	Response Variable	group1	group2	p	p.adj	p.format	p.signif	method
Flexible	Count	Blue	Orange	4.91E-06	9.8E-06	4.90E-06	****	Wilcoxon
Core	Count	Blue	Orange	0.619	0.62	0.62	ns	Wilcoxon

Table 7.11 Wilcoxon Signed-Rank test to compare the similarity of COG categories between Flexible and Core Genomes

Population	Response Variable	group1	group2	p	p.adj	p.format	p.signif	method
Blue	Count	Flexible	Core	0.354	0.35	3.50E-01	ns	Wilcoxon
Orange	Count	Flexible	Core	3.52E-05	0.00007	3.50E-05	****	Wilcoxon

7.10. SUPPLEMENTARY 16. Annotations of swept genes in recombining *UBA11096* sp. cluster 2 populations

Table 7.12 Swept genes common in the core genome of *UBA11096* sp. Cluster 2 Blue populations between July 2019 and July 2020 time points only

Cluster ID	Gene Count	UniProt ID	Functional Annotation
cluster96	2	O67287	GO:0045910; P:negative regulation of DNA recombination; IEA:InterPro
cluster100	2	N/A	N/A
cluster101	2	O67124	GO:0006281; P:DNA repair; IEA:UniProtKB-KW
cluster108	2	N/A	N/A
cluster18	4	O59179	GO:0008236; F:serine-type peptidase activity; IEA:UniProtKB-KW
cluster41	3	P14203	N/A
cluster46	3	N/A	N/A
cluster47	3	O67900	N/A
cluster1	6	N/A	N/A
cluster2	4	O67907	GO:0006220; P:pyrimidine nucleotide metabolic process; IEA:UniProtKB-UniRule
cluster3	4	N/A	N/A
cluster4	4	P28264	GO:0030435; P:sporulation resulting in formation of a cellular spore; IEA:UniProtKB-KW
cluster5	4	Q9A5I5	GO:0000160; P:phosphorelay signal transduction system; IEA:UniProtKB-KW
cluster6	4	O67699	N/A
cluster7	4	P31772	GO:0009306; P:protein secretion; IEA:InterPro
cluster8	4	N/A	N/A
cluster9	4	N/A	N/A
cluster10	4	N/A	N/A
cluster11	4	N/A	N/A
cluster12	4	N/A	N/A
cluster13	4	O67115	GO:0006436; P:tryptophanyl-tRNA aminoacylation; IBA:GO_Central
cluster14	4	Q1DB04	GO:0032880; P:regulation of protein localization; IMP:CACAO
cluster15	4	N/A	N/A
cluster16	4	O66470	N/A
cluster17	4	P47203	GO:0043093; P:FtsZ-dependent cytokinesis; IEA:UniProtKB-UniRule
cluster30	3	O67891	GO:0033317; P:pantothenate biosynthetic process from valine; IBA:GO_Central
cluster33	3	O66529	GO:0009231; P:riboflavin biosynthetic process; IBA:GO_Central

cluster35	3	N/A	N/A
cluster38	3	N/A	N/A
cluster45	3	Q9KQ65	GO:0071973; P:bacterial-type flagellum-dependent cell motility; IBA:GO_Central

Table 7.13 Set intersection of swept genes common in the flexible genome of *UBA11096* sp. Orange populations from all time points

*Set of [Orange Jul-19 ∩ Orange Jul-20]

Cluster ID	Gene Count	UniProt ID	Functional Annotation
cluster8	38	N/A	N/A
cluster9	38	N/A	N/A
cluster10	38	A4W4S5	GO:0006813; P:potassium ion transport; IEA:InterPro
cluster11	38	N/A	N/A
cluster12	38	P94361	GO:0005975; P:carbohydrate metabolic process; IEA:InterPro
cluster13	38	N/A	N/A
cluster14	38	N/A	N/A
cluster15	38	N/A	N/A
cluster16	38	N/A	N/A
cluster17	38	Q06530	GO:0016491; F:oxidoreductase activity; IEA:UniProtKB-KW
cluster18	38	N/A	N/A
cluster19	38	N/A	N/A
cluster20	38	N/A	N/A
cluster21	38	P71366	GO:0009307; P:DNA restriction-modification system; IEA:UniProtKB-KW
cluster22	38	N/A	N/A
cluster23	38	N/A	N/A
cluster24	32	N/A	N/A
cluster49	25	N/A	N/A
cluster51	21	Q06065	GO:0010967; P:regulation of polyamine biosynthetic process; IDA:EcoCyc
cluster52	21	B4F6I3	GO:0043065; P:positive regulation of apoptotic process; IBA:GO_Central
cluster104	19	N/A	N/A
cluster112	17	B4F6I3	GO:0043065; P:positive regulation of apoptotic process; IBA:GO_Central
cluster113	17	Q06065	GO:0010967; P:regulation of polyamine biosynthetic process; IDA:EcoCyc
cluster114	17	P32701	GO:0071111; F:cyclic-guanylate-specific phosphodiesterase activity; IGI:EcoCyc
cluster152	8	P32701	GO:0071111; F:cyclic-guanylate-specific phosphodiesterase activity; IGI:EcoCyc

Table 7.14 Set intersection of swept genes common in the flexible genome of *UBA11096* sp. Blue populations from all time points

*Set of [Blue Jul-19 ∩ Blue Jul-20]

Cluster ID	Gene Count	UniProt ID	Functional Annotation
cluster53	20	N/A	N/A
cluster54	20	Q43883	GO:0009399; P:nitrogen fixation; IEA:UniProtKB-KW
cluster55	20	P00464	GO:0009399; P:nitrogen fixation; IEA:UniProtKB-KW
cluster56	20	N/A	N/A
cluster57	20	N/A	N/A
cluster58	20	P00468	GO:0009399; P:nitrogen fixation; IEA:UniProtKB-KW
cluster59	20	Q44145	GO:0065003; P:protein-containing complex assembly; IEA:InterPro
cluster60	20	A0QTK3	GO:0000155; F:phosphorelay sensor kinase activity; IEA:InterPro
cluster61	20	Q01674	GO:0015891; P:siderophore transport; IEA:InterPro
cluster62	20	N/A	N/A
cluster63	20	P46042	GO:0009399; P:nitrogen fixation; IEA:UniProtKB-KW
cluster64	20	B0TBL5	GO:0022904; P:respiratory electron transport chain; IEA:InterPro
cluster65	20	P32701	GO:0071111; F:cyclic-guanylate-specific phosphodiesterase activity; IGI:EcoCyc
cluster66	20	N/A	N/A
cluster67	20	N/A	N/A
cluster68	20	N/A	N/A
cluster69	20	N/A	N/A
cluster70	20	N/A	N/A
cluster71	20	A7FWJ9	GO:0006520; P:cellular amino acid metabolic process; IEA:InterPro
cluster72	20	N/A	N/A
cluster73	20	Q8RBR1	N/A
cluster74	20	N/A	N/A
cluster75	20	N/A	N/A
cluster76	20	Q1DB04	GO:0032880; P:regulation of protein localization; IMP:CACAO
cluster77	20	N/A	N/A
cluster78	20	N/A	N/A
cluster79	20	Q58157	GO:0005344; F:oxygen carrier activity; IEA:UniProtKB-UniRule
cluster80	20	P26506	GO:0065003; P:protein-containing complex assembly; IEA:InterPro
cluster81	20	N/A	N/A
cluster82	20	N/A	N/A
cluster83	20	Q51739	GO:0046872; F:metal ion binding; IEA:UniProtKB-KW

cluster84	20	P00206	GO:0055114; P:oxidation-reduction process; IEA:UniProtKB-KW
cluster85	20	N/A	N/A
cluster86	20	N/A	N/A
cluster87	20	P0CW46	GO:0006808; P:regulation of nitrogen utilization; IEA:InterPro
cluster88	20	N/A	N/A
cluster89	20	N/A	N/A
cluster90	20	N/A	N/A
cluster91	20	Q58529	N/A
cluster92	20	P98052	GO:0004129; F:cytochrome-c oxidase activity; IEA:UniProtKB-EC
cluster93	20	N/A	N/A
cluster94	20	N/A	N/A
cluster95	20	N/A	N/A
cluster96	20	N/A	N/A
cluster97	20	N/A	N/A
cluster98	20	N/A	N/A
cluster99	19	Q3M644	GO:0009399; P:nitrogen fixation; IEA:UniProtKB-UniRule
cluster100	19	O07643	GO:0009399; P:nitrogen fixation; IEA:UniProtKB-KW
cluster101	19	N/A	N/A
cluster102	19	N/A	N/A
cluster103	19	N/A	N/A
cluster105	18	N/A	N/A
cluster106	18	N/A	N/A
cluster107	17	Q44484	GO:0009399; P:nitrogen fixation; IEA:UniProtKB-UniRule
cluster108	17	N/A	N/A
cluster109	17	N/A	N/A
cluster110	17	N/A	N/A
cluster111	17	O67367	GO:0006364; P:rRNA processing; IEA:UniProtKB-UniRule
cluster139	17	P19487	GO:0030245; P:cellulose catabolic process; IEA:UniProtKB-KW
cluster140	16	Q56318	GO:0016491; F:oxidoreductase activity; IEA:UniProtKB-KW
cluster141	15	N/A	N/A
cluster142	15	N/A	N/A
cluster145	13	N/A	N/A
cluster146	13	Q8RCF9	GO:0009098; P:leucine biosynthetic process; IEA:UniProtKB-UniPathway
cluster147	12	O64046	GO:0098003; P:viral tail assembly; IEA:UniProtKB-KW
cluster148	12	N/A	N/A
cluster149	10	Q0VTI9	GO:0007165; P:signal transduction; IEA:UniProtKB-KW

cluster150	8	N/A	N/A
cluster151	8	O64046	GO:0098003; P:viral tail assembly; IEA:UniProtKB-KW
cluster169	4	N/A	N/A
cluster187	3	P19487	GO:0030245; P:cellulose catabolic process; IEA:UniProtKB-KW
cluster194	3	Q8RCF9	GO:0009098; P:leucine biosynthetic process; IEA:UniProtKB-UniPathway
cluster201	2	P54576	GO:0007165; P:signal transduction; IEA:UniProtKB-KW

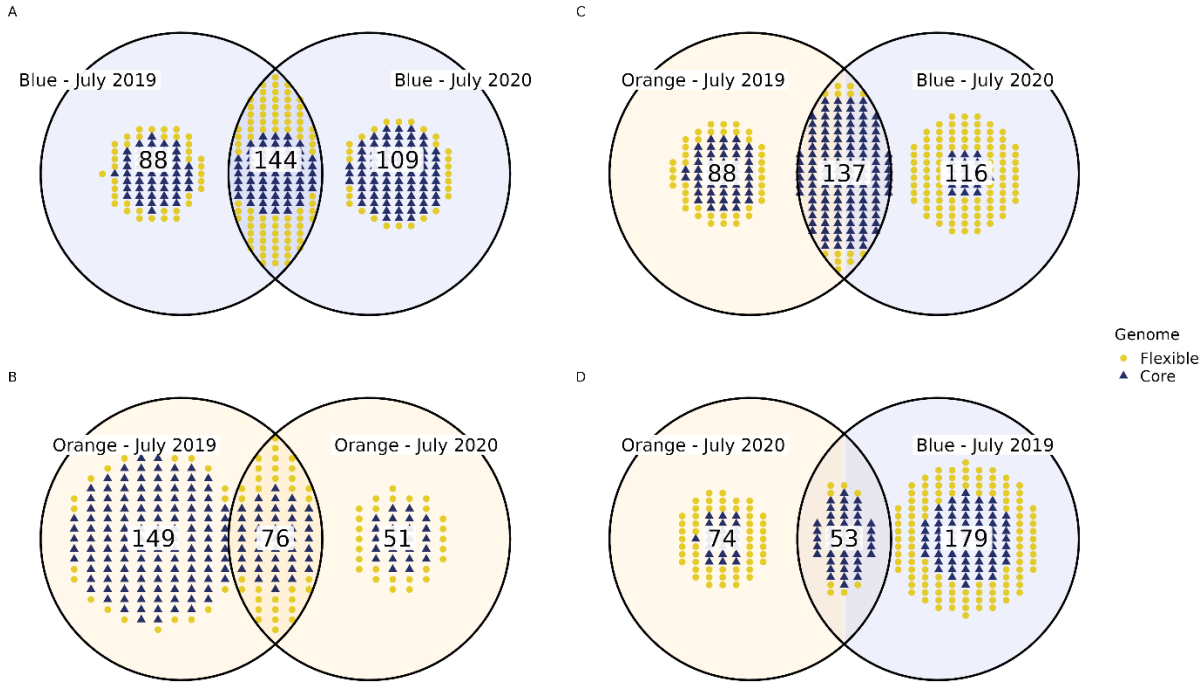

Figure 7.7 Venn diagram of shared COGs between recombining *UBA11096* sp. populations at different time points. (A,B) The same populations of *UBA11096* sp. at different time point share a large number of swept genes indicating that they respond to environmental change with the same processes. (C) Interestingly, *UBA11096* sp. Orange from July 2019 and UBA11096 sp. Blue from July 2020 share also large amount of swept genes indicating that the two closely related responds to similar environmental pressure. (D) This is in contrast to *UBA11096* sp. Orange from July 2020 and July 2019 , which have fewer shared swept genes .

APPENDIX C - CHAPTER 4

8.1. SUPPLEMENTARY 17. Mobile genetic element density in a population

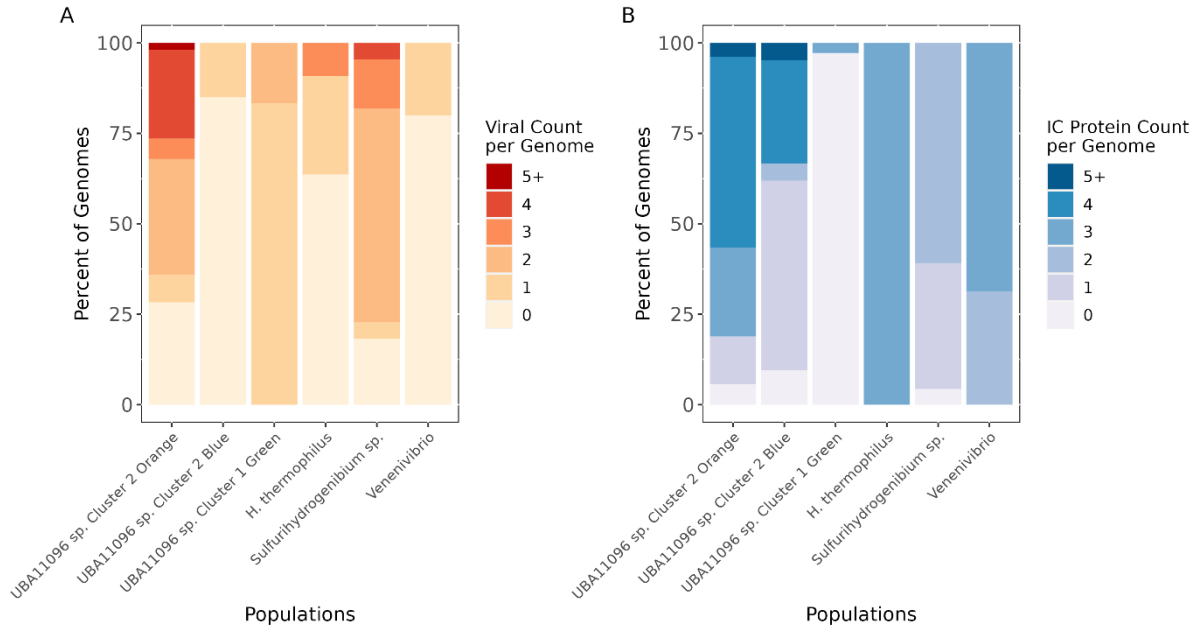

Figure 8.1 Density mapped of viral sequences and IC signature proteins for each population. The density of viral sequences per genome (A) shows that the high amount of viral sequences previously observed in *UBA11096 sp.* cluster 2 Orange and *Sulfurihydrogenibium* is brought by a significant amount of genomes with multiple viral sequences within them. A similar trend can be observed in the IC proteins (B) in which *UBA11096* sp. cluster 2 Orange and Blue populations are dominated by genomes with a large number of IC proteins within them.

8.2. SUPPLEMENTARY 18. Shared CRISPR Spacers

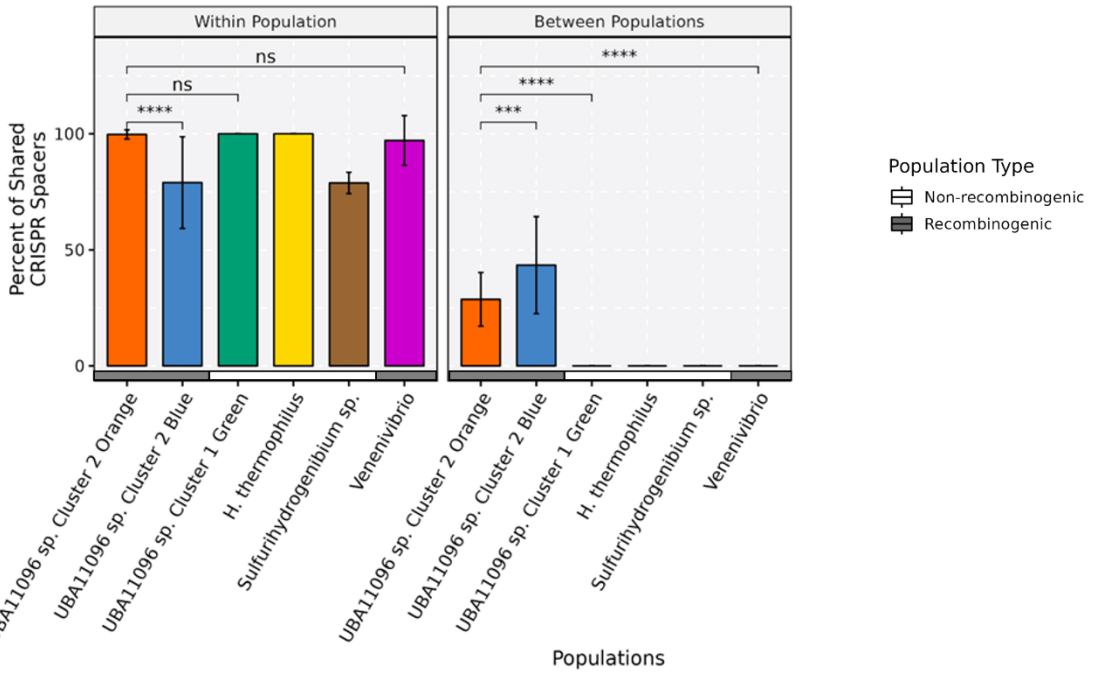

Figure 88.2 Comparison of shared CRISPR spacer arrays between individual populations. No significant differences have been observed between the percentage of shared spacers within populations when comparing recombining populations (*UBA11096 sp.* cluster 2 Blue and Orange, Venenivibrio) with those from non-recombining populations. Besides the *UBA11096 sp.* cluster 2 Blue and Orange populations, CRISPR spacer arrays are not observed to be shared between populations.

8.3. SUPPLEMENTARY 19. Statistical comparison of COGs in swept regions and MGEs

Table 8.1 Analysis of Variance of COG counts

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)	p.Signif
Origin	3	10269	3423	15.1	4.55E-08	***
Residuals	92	20860	227			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 8.2 Tukey multiple comparisons of means between the origin of COGs at 95% confidence interval

Group 1	Group 2	Mean difference	Lower endpoint	Upper endpoint	p.adjusted	p.Signif
Mobile Elements	CRISPR Spacers	-2.83333	-14.2073	8.540669	0.914661	
Swept Region	CRISPR Spacers	23.79167	12.41766	35.16567	2.3E-06	***
Viral Sequences	CRISPR Spacers	7.541667	-3.83234	18.91567	0.311702	
Swept Region	Mobile Elements	26.625	15.251	37.999	1E-07	***
Viral Sequences	Mobile Elements	10.375	-0.999	21.749	0.086788	.
Viral Sequences	Swept Region	-16.25	-27.624	-4.876	0.001799	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1