7-26-2016

Haloarchaeal Species Genesis: Diversity, Recombination, and Evolution

Nikhil Ram Mohan
University of Connecticut - Storrs, r_nikhil85@yahoo.com

Haloarchaeal Species Genesis:

Diversity, Recombination, and Evolution

Nikhil Ram Mohan

University of Connecticut, 2016

ABSTRACT

The processes key to the generation of new species in eukaryotes is well understood but the same is not true for the prokaryotic world. Allopatric speciation is widely accepted as the driving force in eukaryotic species genesis. However, little is known about prokaryotic speciation primarily due to the abundance of prokaryotes in this world. They inhabit every fathomable niche, present high cell densities, complex community structures, and are easily dispersed between large distances. For example, a gram of soil or a milliliter of seawater house millions and millions of a vast variety of microorganisms. This complexity in the prokaryotic world makes inferring an overall evolutionary processes extremely difficult.

To reduce the complexity, the system studied needs to be narrowed down. This thesis focusses on one environment that is so extreme that it is conducive for the growth of only certain groups of organisms. The system is characterized by high salinity, much greater than that of seawater, which is a prerequisite for the survival of an entire class of Archaea call the Halobacteria (Haloarchaea). Studying these environments reduces the microbial complexity in comparison to some of the more easily habitable areas like soil or water. Haloarchaea are the dominant group of

organisms in hypersaline environments. The dissertation describes a multiscale approach studying haloarchaeal communities to determine the evolutionary forces influencing species genesis in the haloarchaea.

First, a spatial distribution analysis of the haloarchaeal communities using both PCR amplified environmental gene marker as well as metagenomic comparisons reveal unique haloarchaeal communities in geographically distant hypersaline environments. Similar to eukaryotes, endemism is apparent in haloarchaea. However, unlike the eukaryotes, there is neither a distance-decay relationship nor is there a similarity in haloarchaeal communities based on whether they are from lakes or salterns. Second, temporal analysis on one haloarchaeal community reveals stability in the community at the genus level. Stable environmental conditions provided by the hypersaline environments ensure stability in the community with only fluctuations in the relative abundances with respect to salinity. Third, comparing individuals within a community showed widespread genomic variations between isolates. Comparing a multi gene concatenated phylogeny and whole genome fingerprinting exposed that even isolates that were identical at each locus tested had varying genome patterns. This happens at a rate much greater than the accumulation of third codon substitutions. Finally, assaying two highly conserved genes from 109 haloarchaeal genomes evidenced the existence of extensive recombination at predicted rates far greater than the rate of mutation in haloarchaea.

This dissertation discusses a working model for species genesis in haloarchaea based on the obtained evidence. It involves frequent recombination within the community members at each geographically distant site, homogenizing them, and maintaining endemic populations while often forming stable chimaeras that could become new species. These findings in haloarchaea offer a peek into the mysteries of the prokaryotic world.

Haloarchaeal Species Genesis:

Diversity, Recombination, and Evolution

Nikhil Ram Mohan

B.Tech, Dr.M.G.R. Educational and Research Institute, 2007

India

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2016

APPROVAL PAGE

Doctor of Philosophy Dissertation

Haloarchaeal Species Genesis:

Diversity, Recombination, and Evolution

Presented by

Nikhil Ram Mohan, B.Tech.

Major Advisor

_____

R. Thane Papke

Associate Advisor

_____

Johann Peter Gogarten

Associate Advisor

_____

Spencer Nyholm

Associate Advisor

_____

Daniel Gage

Associate Advisor

_____

Kent Holsinger

University of Connecticut

2016

ACKNOWLEDGEMENTS

I take this opportunity to express my deepest gratitude to my advisor, Dr. R. Thane Papke, without whose support and invaluable guidance I would not be where I am this day. We have had many exciting discussions through the years that fueled my curiosity and helped develop not only my skills but also mold me into the scientist I am. Thank you for being a great advisor. I would also like to thank my advisory committee, Drs. Peter Gogarten, Spencer Nyholm, Kent Holsinger, and Dan Gage for taking time away from their busy schedules to always help me out with my problems and answer my queries.

Having been in the Papke lab as long as I have, I was lucky to know some very nice people as they made their way through the lab. I was and am fortunate to have had a lab full of friendly people who made the lab a stimulating place to work at. I would also like to thank all of the friends I have made outside of the lab, graduate school would not have been the same without all of you.

None of this would have been possible if not for the everlasting support of my family. My parents, uncle, aunt, brother, and sister have been the foundation that I stand on. I am deeply indebted to them. Though infrequently, visiting my brother and his family would be my only consolation for spending all this time away from my family in India. Finally, I would like to thank my wife, Kunica, for she is my rock, my biggest cheerleader, my best friend. The past five years have been the best of my life and I cannot imagine having gotten this far without you.

The future holds a different field in science for me but the haloarchaea will always have a special place in my heart since this small group of microorganisms have taught me all I know!

Table of Contents

List of Figures

List of Tables

Chapter 1 : Introduction

1.1 Modes of Speciation

Mechanisms of species genesis are well described for macro-organisms. Speciation is typically driven by the formation of reproductive barriers within populations in one of four different ways. These include sympatric, peripatric, parapatric, and allopatric means of species genesis. Sympatric speciation involves the rise of new species within a co-existing ancestral population. A typical example of this mode are the apple maggot flies that developed variation by broadening their host range of apples [3]. Greatest evidence for speciation in the macro world is through the rise of geographic isolation of populations [1]. This is defined as allopatric speciation. The two other modes of speciation – peri and parapatric are similar to allopatric speciation in many ways. In peripatric speciation, new species rise in the periphery of a population since the members in the periphery are isolated from the rest of the population. Parapatric speciation, on the other hand, occurs in a continuous population due to selective mating only with immediate neighbors.

In allopatric speciation, the two basic models describing the rise of geographic separation are further defined as vicariance and dispersal [4] (figure 1-1). Environmental or geological events like the rise of mountain ranges, changes in river direction, or continental breakup drive vicariance, whereas active and passive movements from centers of origin designate dispersal. In either case, barriers to homogenizing forces (e.g., random mating) arise thereby resulting in subpopulations through isolation, eventually inducing divergence leading to speciation.

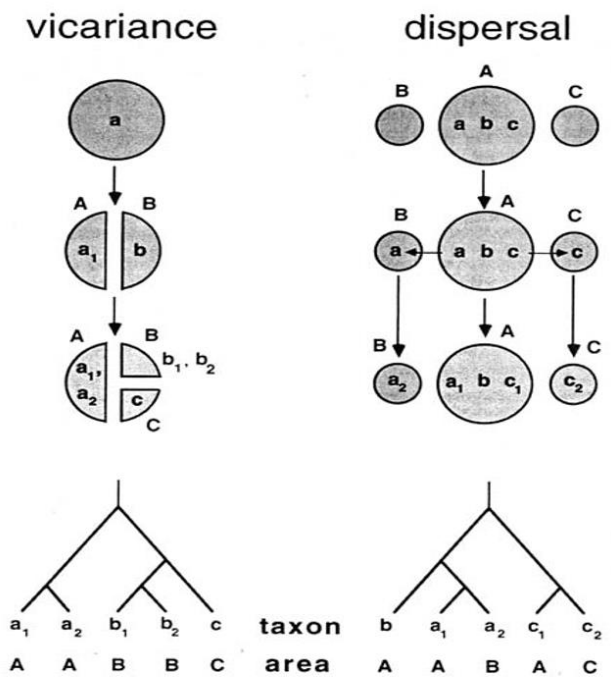


Figure 1-1. Rise of geographic isolation and hence speciation due to vicariance and dispersal as described by Avise [1].

Studying the evolutionary processes leading up to the formation new species is often facilitated by focusing on islands. This field, broadly coined Island Biogeography, has evolved through the centuries owing to the work of giants concentrating on different island and archipelago systems like the Galapagos island by Darwin [5] and Malay Archipelago by Wallace [6] that stemmed from the writings of Buffon (1761) and Linnaeus (1781). The modern take on the concepts of island biogeography is based on the isolation of the islands in question that act as a barrier of dispersal of species thereby diverging the species on different islands through natural drift thereby giving rise to allopatric speciation. This is particularly important since the classical interpretation of the model of island biogeography put forth by McArthur and Wilson (1963) did

not account for speciation. Combining phylogenies to infer species history along with the earth's geological history aided the discrimination between the effect of vicariance or dispersal in the observed pattern. Adopting the models of island biogeography and hence inferring allopatric speciation, caused by leaky barriers to dispersal and invasiveness is a well understood process in plants and animals [7], but its role in microbial evolutionary theory has had restricted impact due to complexities in deriving overarching conclusions for the vast microbial world.

1.2 Challenges in the Prokaryotic World

A few prokaryote biogeographic studies have ventured to tackle placing the microbial world on the map [8-25]. One of the biggest challenges posed to understanding microbial biogeography is that microorganisms have historically been thought as unconstrained by dispersal limitations because of their small size, the formation of resistant stages and their ubiquitous distributions. In fact, a large number of viable microbial cells are transferred annually between continents through the atmosphere, estimated in the order of 10^{18} [26] suggesting that dispersal may in fact homogenize populations distributed globally. Any new species thus evolved must be a result of adaptation to a new niche in the same location and sympatric speciation would be the driving force if geographic isolation is rare [27, 28]. Comparative genome analysis of *Vibrio* [29] and *Sulfolobus* [30] populations corroborate this. However, with the increase in biogeographic analyses of microbial communities from different niches, a mixed bag of patterns is observed. Currently, there is evidence for the existence of pandemic populations completely negating the effect of geographic isolation; unique community assemblages from geographic isolation; and a combination of the two where the dominant members are widespread whereas the low abundant

rare members are site specific (e.g., [12, 17, 31-36]) suggesting that different microorganisms have varying capacities for dispersal and invasion.

Fate of a dispersing microbial species is reliant upon its ability to invade and colonize its destination. This, however, is extremely difficult since the stability of a community inhabiting a niche enforces resistance to the invasion and colonization of the dispersed microbial species [37-39]. The ability of an entering species to survive in a new community is negatively correlated to the species richness of the existing community [40, 41]. Mature and stable communities can facilitate a collective territorial defense against invaders with the help of a suite of antimicrobial agents [42], and an organized and sophisticated colonization resistance pathway has been identified within a host-associated microbial community [43]. Defense against invasion involves segregating species into unique social structures, each group playing a different role in the resistance to colonization. In certain cases, maintenance of high diversity establishes stability in the genetic heterogeneity. The 'kill-the-winner' hypothesis, which describes phage predation of the fittest, most numerous cells indicates single clones or possibly species cannot rise to dominate within populations or communities, and diversity is maintained. In effect, these mechanisms keep the community stable against environmental volatility with respect to invading species [44-46].

Taxonomic stability in microbial communities is also impacted by environmental influences and various patterns are seen in different niches. For instance, microbial communities inhabiting pineland soils exhibited changes in diversity on a seasonal basis, but exhibit no differences in the microbial biomass [47]. Environmental factors have been shown to shape the microbial communities inhabiting different river ecosystems [48-51]. Changes in water temperature and conductivity, for example, were identified to influence temporal partitioning in bacterial communities of a subtropical river. Similar studies on prokaryotic communities from

different lakes revealed seasonal variation in species abundances and functional capacity of the communities [52, 53]. The bacterioplankton community within the Salton Sea was shown to undergo seasonal variations with minimal overlap in the detected community composition between the sampled seasons presumably in response to environmental instability [54]. In contrast to these studies, community stability was observed throughout the seasons for cyanobacterial populations defined by temperature that inhabit a microbial mat from Octopus Spring in Yellowstone National Park, most likely due to a constant abiotic environment not prone to disturbances [55]. While it is clear that the environment has a huge impact on communities, changes in structure due to perturbations however may result only in abundance fluctuations of indigenous populations, rather than opening new niches for the invasion of non-native species, as seen in Florida beach sands before, during and after oil contamination due to the Deep Water Horizon spill [56]. Therefore each environment, species or community requires a case-by-case examination to understand how microorganisms inhabiting different niches and locations are distributed and evolving.

1.3 Hypersaline environments – the ideal 'Island' for Prokaryotic Biogeography

Hypersaline environments are characterized by salt concentrations higher than seawater. These are divided into two major categories based on the ionic composition: thalassohaline and athalassohaline environments [57]. Thalassohaline environments typically result from the evaporation of seawater, which concentrates salts and ion ratios similar to that of its origins, until specific salts (e.g., calcium sulfate) reach saturation and precipitate [57]. Sabkhas, salt marshes, and sea salt production facilities are examples of thalassohaline environments. Many hypersaline lakes are examples of athalassohaline environments, and can be formed when a water body is landlocked and terminal (e.g., Dead Sea). Salinity increases as minerals are transported into the

lake from the surroundings and evaporation occurs. This process dictates that the ionic composition is unique in each athalassohaline environment. However, some inland lakes are thalassohaline, e.g., Tuz Lake, Turkey [58] and Lake Tyrrell, Australia [59].

Physicochemical studies on hypersaline environments, both thalassohaline and athalassohaline, have shown that owing to the high salinity, these environments are subject to low solubility of gases, diffusion rates and very low water activity [60]. These hypersaline environments also vary in pH [61] making them too extreme for most organisms. Given that saturated brines require of microorganisms specific adaptive characteristics to survive in this unique habitat, and that these habitats typically exist in hot, dry climates where environmental conditions are generally stable, it is predicted that many of them will maintain a steady taxon community structure like that seen in some hot springs [55]. Global distribution, geographic isolation, and stringent physicochemical properties supporting the growth of very few organisms, of hypersaline environments make them ideal 'island' like entities to study the biogeography and stability of the inhabitants providing an insight into the prokaryotic spatio-temporal distribution.

1.4 The Haloarchaea

Members of the class Halobacteria (Domain: Archaea; Phylum: Euryarchaeota), usually called haloarchaea to distinguish them from halophilic bacteria, typically thrive in the hypersaline environments, in moderate (15% - 20% NaCl) and saturated brines (~35% NaCl). Crystallizer ponds are a typical example of saturated brines, where NaCl precipitates at ~32-37% and is then harvested for commercial purposes, and is dominated by the haloarchaea [62-73]. To overcome many obstacles posed by hypersaline environments, Haloarchaea can generate ATP from light energy [74] and have gas vesicles to buoyantly lift themselves to the surface [75]. Osmotic

survival in these brines is managed by maintaining a cytosolic salinity in equilibrium with that of the environment, a feat that requires solubilized proteins under those conditions, and solved with a proteome enriched in acidic and depleted of basic amino acids [61]. Several cultivation and molecular based studies on crystallizer ponds have led to some general conclusions about haloarchaeal communities. Individual communities tend to be comprised by a small number of dominating genera, with the square archaeon, *Haloquadratum walsbyi*, often reported as having the largest population sizes [14, 62, 65-67, 76, 77] sometimes comprising >60% of all the archaea [62]. Dominance by the genus *Haloquadratum* is observed in some hypersaline lakes also [58, 63]. This however, does not appear to be the case in every hypersaline environment. Snapshot analyses identified *Halorubrum* related phylotypes as the most frequently retrieved genus from a saltern in Slovenia [78] and *Haloquadratum* 16S rRNA gene sequences were not recovered from a saltern in San Diego [79]. Additionally, in some of these studies, frequently observed halobacterial clones retrieved did not cluster with any cultivated and described halobacterial species.

1.4.1 Frequent gene transfer and recombination in Haloarchaea

Haloarchaea have a well-documented capacity for generating enormous amounts of genetic variation through horizontal gene transfer (HGT) [80-86]. From the very first genome sequence analysis of *Halobacterium* strain NRC-1, evidence was provided for the acquisition of aerobic respiration genes via HGT from Bacteria [87]. Since then, several studies on specific genes of interest [e.g., rhodopsins [88], ribosomal RNAs [89] and tRNA synthetases [90]] have further demonstrated gene transfer into and among the Haloarchaea. A recent report suggested that this process of generating diversity has been ongoing since before the group's last universal common ancestor and that HGT played a huge role in changing their physiology from an autotrophic

anaerobe to a heterotrophic aerobe [91]. Population genetics analysis on strains from the genus *Halorubrum* using multilocus sequence analysis (MLSA) demonstrated that alleles at different loci are unlinked indicating that homologous recombination (HR) is frequent enough within phylogenetically defined groups to randomize traits among individuals [83, 84], an observation once considered unique to sexually reproducing eukaryotes. Analysis of 20 haloarchaeal genomes showed that there are no absolute barriers to HR, which occurs regularly and proportionally to genetic distance throughout the Haloarchaea [85]. Community analyses using metagenomics revealed that genes are coming and going quickly within *Haloquadratum walsbyi* populations, suggesting there may be very few identical genomes within the species [77, 80]. Perhaps most striking is their ability to exchange large swaths of genetic information. Also, genomes of highly divergent strains (e.g., <75% average nucleotide identity) isolated from Deep Lake, Antarctica were shown to share many ~100% identical DNA sequences in fragments up to 35Kb in length [81].

The haloarchaea are recognized to undergo a fourth mode of gene transfer called mating, apart from the three well described modes in prokaryotes namely transformation, conjugation, and transduction [92]. This process is different from conjugation in that unlike conjugation, gene exchange occurs in a bidirectional manner. Though the exact mechanism involved in mating is unknown, a model proposed is described in figure 1-2 [2]. Briefly, when two haloarchaeal cells of the same or related species are in close proximity, they form intercellular cytosolic bridges that expand and eventually the two cells fuse forming a heterodiploid cell (see figure 1-2 A, B, C, D, and E). From here, the genetic material could either segregate to regain the original state (figure 1-2 F) or recombination could occur there by forming new hybrid cells (figure 1-2 G). Mating experiments between *Haloferax volcanii* and *Haloferax mediterranei* demonstrated between ~10

and 18% (~300-500kb) of their chromosome could be transferred in a single fragment and the frequency of recombination in *Haloferax volcanii* is limited by the frequency of mating. [82].

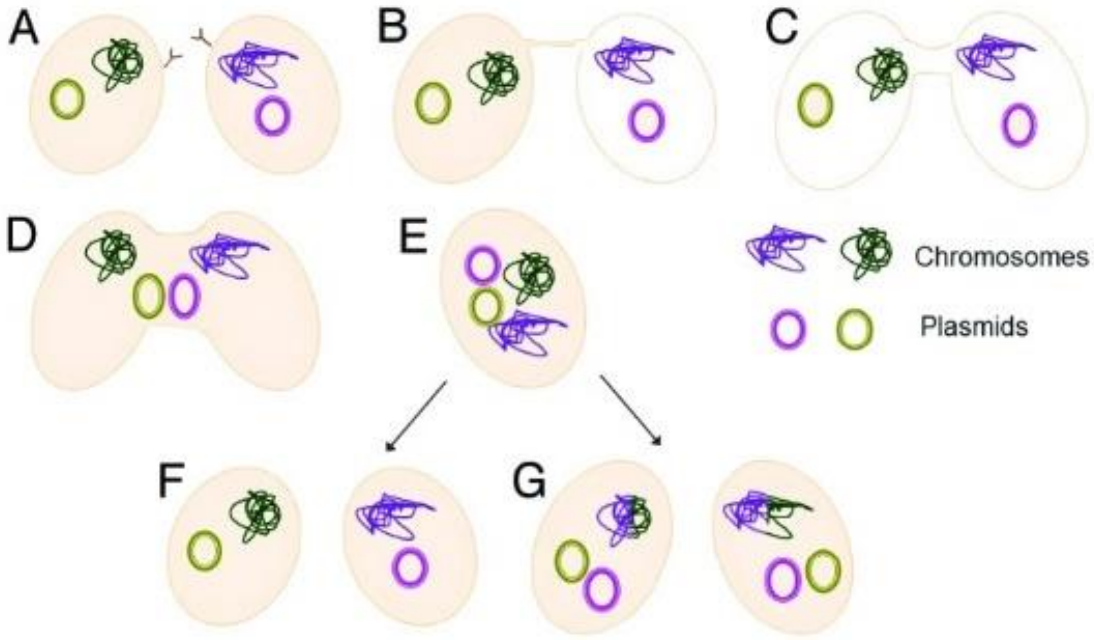


Figure 1-2. Proposed model for haloarchaeal mating [2].

MLSA has often been used as a technique for classifying microorganisms [93], including halophiles [31, 94], but it is also used to estimate population variation and gene flow [95]. Assumptions using MLSA regarding how representative multiple genes are for capturing individual variation, and thus the appearance of clonality, can lead to erroneous conclusions. For instance, two strains may have identical sequences across multiple loci, but unexamined genomic variation might be high and belie the interpretation of little or no recombination. Indeed, studies are demonstrating that there are vast amounts of variation within bacterial species/populations. Environmental isolates with identical HSP-60 genes from a natural coastal *Vibrio* sp. population demonstrated that the overwhelming majority of individual strains were unique as determined by

chromosome pulse field gel electrophoresis, with some strains differing by up to a megabase in genome size [96]. This variation in genome size and the existence of "open" (i.e., infinite) pan-genomes like that of *Prochlorococcus marinus* and others [97, 98] suggest that HGT is so frequent that for at least some species every cell may be genetically distinct. Given that the haloarchaea are highly recombinogenic, it brings to question whether these salt loving members of the archaeal domain are in fact individuals in a population rather than clonal masses.

To summarize, recombination through horizontal gene transfer (HGT) is rampant in the entire class of Halobacteria [80-86, 99-101]. There is evidence for recombination occurring often, even between distantly related taxa [81], transferring not only small but large fragments of DNA [81, 82] at great rates [83, 84], faster even than the rate of accumulation of third codon substitutions [100]. This homogenizing force is unhindered by the presence of the CRISPR-Cas system [101] and is mired only by sequence divergence [82, 85], yet there is recombination between lineages that are ~50% divergent from one another [85].

1.5 Choice of molecular marker

1.5.1 16S rRNA and *rpoB* genes

Molecular markers have been extensively used in the past couple of decades to study the diversity and phylogeny of microorganisms in their natural environment. Two of the most commonly adopted markers for this purpose are the 16S rRNA and ß subunit of the RNA polymerase (*rpoB*) genes both of which are ubiquitous in the prokaryotes. They also have conserved regions constrained by slow rates of evolution [102, 103] making it easy to design primers to amplify a wide range of taxa from an environmental sample. These advantageous

characteristics of the 16S rRNA go hand in hand with the gene being present in multiple copies in the genome. In comparison, rpoB is a single-copy gene. A survey of 111 bacterial genomes resulted in the identification of four hundred and sixty copies of the 16S rRNA and 111 copies of rpoB. Each genome, on average, had 4.2 copies of the 16S rRNA [103]. In a more recent study covering a larger number of available prokaryotic genomes, 425 species of prokaryotes were estimated to have anywhere between 2 and 15 copies of the 16S rRNA in the genome [104]. In a different study looking at 1690 bacterial genomes [105], 7,081 16S rRNA copies were identified with the same average copy number per genome as reported earlier [103]. Only ~15% of the 1690 genomes analyzed had a single copy of the 16S rRNA. Most genomes had between 2 and 7 copies and a few rare genomes had greater copy numbers [105]. The multiple copies within a genome are seldom identical. In most genomes, including *Pseudomonas* [106], at least two copies of the 16S rRNA differ from each other [103, 107]. In certain cases in genomes with multiple copies, each copy is different from the other. *Bacillus subtilis* and *Clostridium perfringens* are examples, each with ten copies of the 16S rRNA and heterogeneity in each copy [108]. In general, increase in copy number of the gene correlates with the heterogeneity. Any genome with six or more copies has at least two variants of the 16S rRNA [105]. The diversity between the heterogeneous copies of the 16S rRNA ranges between 0.06% and 20.38%, with many of the species having ~1.3% divergence while *Thermoanaerobacter tengcongensis* has copies that are ~6.7% divergent and *Borrelia afzelii* has a pseudogene that is ~20.38% divergent [104].

The presence of multiple divergent copies of the 16S rRNA in the genome is not unique to bacteria. There are many examples of Halobacteria that have multiple copies of the 16S rRNA. The archaea of the class Halobacteria have an obligate requirement for high salinity for their survival and are the dominant organisms in hypersaline environments [62, 63]. The Halobacteria

have considerable variation in the 16S rRNA copy numbers [109]. Three copies were identified in *Halobacterium halobium* NCMB 777, and two possible copies in *Haloferax volcanii*. In fact, it was identified that the 16S rRNA from *Halobacterium salinarium* CCM 2148 [109] was the same as that from *Halobacterium halobium* strain R1 [110] and *Halobacterium cultirubrum* [111]. Sequence heterogeneity in duplicate copies of the 16S rRNA were also observed in *Haloarcula marismortui* as well [112, 113] and the second copy is more divergent from *Hfx. volcanii*, *Hbt. cultirubrum*, and *Halococcus morrhuae* than the first copy and the two copies are 5% divergent [112]. A third copy, almost identical to the first one, was found when the genome was sequenced [114]. This phenomenon is pervasive in the Halobacteria with instances in other genera/species as well – *Halosimplex carlsbadense* was identified to have three gene copies, one of which is ~7% divergent from the other two [115], but was later determined that only two copies existed and the third was a PCR induced chimaera [89]; *Natrinema* sp. strain XA3-1 has four copies where one is 5% divergent from the others [89]; *Halomicrobium mukohataei* JCM 9738 has gene copies that are 9% divergent [116].

The heterogeneity in the copies of 16S rRNA of *Natrinema* sp. strain XA3-1 was determined to be localized at hotspots that resulted from recombination events across large taxonomic distances [89]. Despite the extensive proof for HGT and recombination in the Haloarchaea, its effect on the widely assumed 'gold standard' molecular markers is poorly understood.

Many studies in the past on various systems have surveyed the 16S rRNA and *rpoB* genes alongside to compare their efficacies in determining microbial diversity [103, 117-120], and identifying and grouping isolates with better phylogenetic resolution [121-128]. In the Haloarchaea, *rpoB* was used to refine the 16S rRNA gene phylogeny and was deemed to be

successful as a supplementary tool in the taxonomic classification of new isolates [129]. Given that the Haloarchaea have multiple copies of the 16S rRNA while maintaining *rpoB* as a single copy gene and factoring in the frequent gene transfer, it is important to decipher the rate and effect of recombination on the evolutionary histories of these to widely used genes.

1.5.2. The bacteriorhodopsin gene

Bacteriorhodopsin (*bop*), a member of the rhodopsin protein family, is a seven pass transmembrane protein that acts as a light-driven proton-pump (figure 1-2) and generates an electrochemical gradient for the production of ATP [130-132]. Bacteriorhodopsins are present in significant quantities in solar salterns [133, 134]. It is also Halobacteria specific and circumvents the issues of small evolutionary distance resolution of 16S rRNA genes, and the observation that several halobacterial genera have multiple divergent rRNA operons [89, 112, 115]. PCR amplification of the *bop* gene as previously used [13, 135, 136]. *bop* is shown to recover the

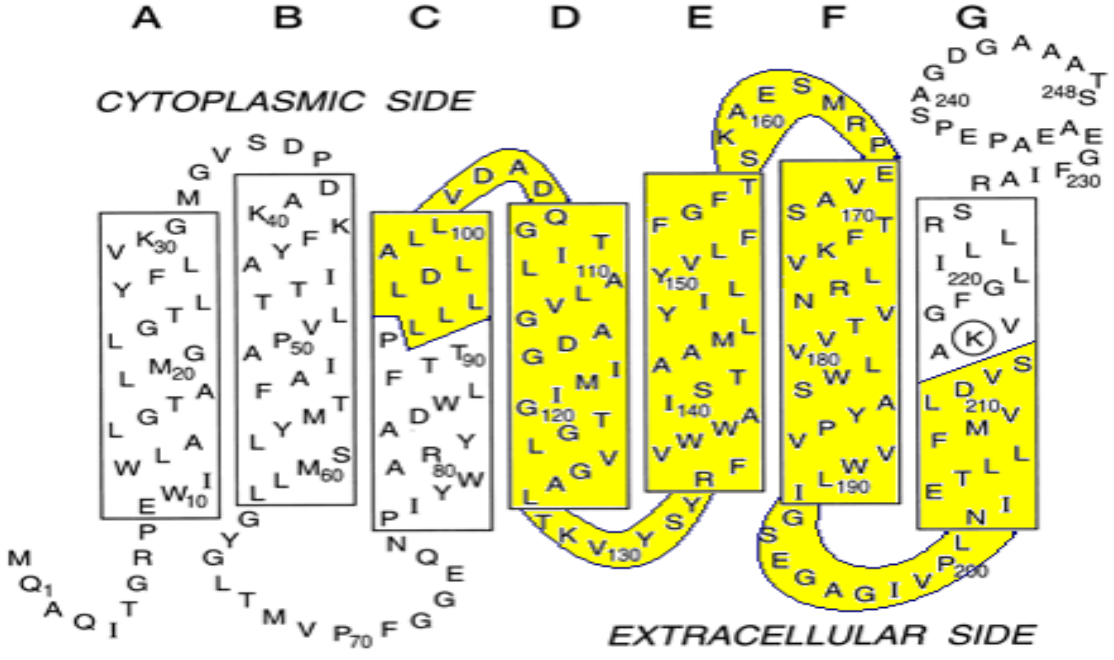


Figure 1-3. Schematic of the bacteriorhodopsin protein Modified from Marti, 1998 by Jani. The region highlighted in yellow represents the area amplified by the PCR primers in the studies listed.

familiar genera and diversity in halophilic environments when compared to the 16S rRNA gene [13, 78] and provides excellent support for binning haplotypes [13]. Given these advantages of *bop* over the 16S rRNA gene, it was adopted as the molecular marker of choice for most analyses described in this thesis.

1.6 Overarching goals of this thesis.

This thesis aims to identify the processes involved in the formation on new haloarchaeal species in the face of frequent recombination acting as a homogenizing force thereby giving an insight into the enigmatic prokaryotic speciation. This is achieved by applying the concepts of biogeography into the world of the haloarchaea and teasing apart the effect of geographic separation of hypersaline environments, impact on the community of inhabitants due to relative stability in these environments in comparison to other complex niches like soil or water, and understanding the dynamics between individuals in nature. In order to do this, this thesis is divided into the following overarching goals and the chapters that follow discuss the methodologies adopted, results obtained and their interpretations.

1.6.1 Biogeography of the Haloarchaea

Aim to place the haloarchaea on the world map and infer the effect of geographic separation on the community composition and structure in various distant hypersaline environments, both manmade salterns and naturally occurring lakes. Based on what is known in the eukaryotes and on previous prokaryotic biogeographic studies (discussed in sections 1.1 and 1.2), predictions can be made as to what patterns might be observed, be it endemism, cosmopolitan distribution, or a

combination of both (see figure 1-3). Allopatric speciation can be inferred if the haloarchaeal communities are endemic and the effect of geographic separation can be negated if the communities exhibit cosmopolitan distribution. If niche isolation does in fact play a prominent role in the observed patterns, the question arises to whether an increase in the distance between two niches impacts the dispersal of haloarchaea from one to another inversely. Or if the similarity in ecotype, whether saltern or lake, plays a role. Here we examine the biogeographic distribution

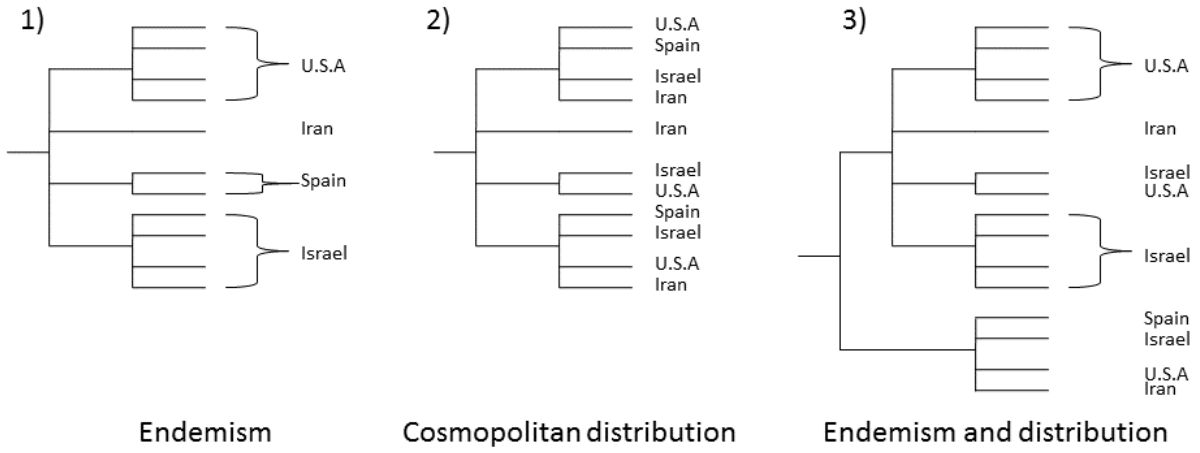


Figure 1-4. Predictions of observed patterns in haloarchaeal communities. Comparing communities from USA, Iran, Spain, and Israel, the three predicted patterns based on phylogenetic clustering are depicted.

and diversity of haloarchaea from multiple saturated brines located around the world. Findings for these questions are described in chapters 2 and 3.

1.6.2 Temporal analysis of one Haloarchaeal community

Aim to determine if hypersaline environments house communities that are stable through time. Stability in prokaryotic communities is known to resist invasion and yet not all prokaryotic

communities are stable through time (as discussed in section 1.2). Studying a community through time will provide insight into its stability. This aim focuses on the structure of the halobacterial community inhabiting the sea salt crystallizer ponds in Eilat, Israel through time. The Israel Salt Industries Ltd. in Eilat, operational since 1977, produces about 170,000 tons of salt a year and, harvesting is year round. Ponds 301 – 304 [137] and, more recently, ponds 305 – 307 are the salt crystallizing ponds with salt concentrations of 300 gL^{-1} and above. Findings for this aim are described in chapter 4.

1.6.3 Dynamics of Individual Haloarchaea in a population

Aim to determine if the haloarchaea are clonal in nature and if not, what could possibly be causing the observed variation. Though clonality is assumed in the haloarchaea just like in all prokaryotes, some studies on other prokaryotes have shown that closely related species vary in their genome content (discussed in section 1.4.1). To get a better understanding for the genomic variation within closely related haloarchaeal strains, naturally co-occurring environmental strains from the genera *Halorubrum* and *Haloarcula* isolated from the Aran-Bidgol salt lake in Iran were examined. MLSA was used to identify closely related strains, and a PCR genome fingerprinting technique that randomly primed amplification sites along the chromosome to generate a gel electrophoresis pattern to compare genomic variation of the isolates. Chapter 5 lists and discusses findings for this aim.

1.6.4 Assay the extent of recombination in the Haloarchaea

Aim to estimate the extent of recombination in the haloarchaeal 16S rRNA and *rpoB* genes. Recombination is rampant in the haloarchaea and there is previous evidence for recombination in the 16S rRNA gene (discussed in section 1.4.1). Yet, the effect of such recombination on the phylogeny of haloarchaea derived by both the genes is poorly documented (discussed in section 1.5). The genes were compared to determine the extent of recombination in each of these across 109 available Haloarchaeal genomes and its role in distorting the true taxonomic delineation in this entire class. Chapter 6 discusses the results from this aim.

Chapter 2 : Cell sorting analysis of geographically separated hypersaline environments [12]

In an effort to analyze the spatial distribution of Haloarchaea, a paper with the above title was published in Extremophiles in 2013 studying three hypersaline environments using cell sorting and single cell amplification of genomic content. This work was in collaboration with Drs. Zhaxybayeva and Stepanauskas. My contribution to this work was the direct comparison of communities to estimate overlapping based on Operational Taxonomic Units (OTUs) defined at 99%, 97%, and 95% sequence similarity of the 16S rRNA gene.

Extremophiles (2013) 17:265–275
DOI 10.1007/s00792-013-0514-z

ORIGINAL PAPER

Cell sorting analysis of geographically separated hypersaline environments

Olga Zhaxybayeva · Ramunas Stepanauskas · Nikhil Ram Mohan · R. Thane Papke

Received: 22 October 2012 / Accepted: 4 January 2013 / Published online: 29 January 2013

Abstract Biogeography of microbial populations remains to be poorly understood, and a novel technique of single cell sorting promises a new level of resolution for microbial diversity studies. Using single cell sorting, we compared saturated NaCl brine environments (32–35 %) of the South Bay Salt Works in Chula Vista in California (USA) and Santa Pola saltern near Alicante (Spain). Although some overlap in community composition was detected, both samples were significantly different and included previously undiscovered 16S rRNA sequences. The community from Chula Vista saltern had a large bacterial fraction, which consisted of diverse Bacteroidetes and Proteobacteria. In contrast, Archaea dominated Santa Pola's community and its bacterial fraction consisted of the previously known *Salinibacter* lineages. The recently reported group of halophilic Archaea, Nanohaloarchaea, was detected at both sites. We demonstrate that cell sorting is a useful technique for analysis of halophilic microbial communities, and is capable of identifying yet unknown or divergent lineages. Furthermore, we argue that observed differences in community composition reflect restricted dispersal between sites, a likely mechanism for diversification of halophilic microorganisms.

Keywords Haloarchaea · Cell sorting · Genome amplification · Biogeography · Nanohaloarchaea · Prokaryotic speciation

Communicated by L. Huang.

Electronic supplementary material The online version of this article (doi:10.1007/s00792-013-0514-z) contains supplementary material, which is available to authorized users.

O. Zhaxybayeva
Department of Biological Sciences, Dartmouth College, Hanover, NH 03755, USA

R. Stepanauskas
Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA

N. R. Mohan · R. T. Papke (✉)
Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269, USA
e-mail: thane@uconn.edu

Introduction

Solar salterns are industrial sea-salt manufacturing facilities found in coastal regions typically with hot and dry climates. Seawater is pumped into a series of shallow ponds, and increased salinity concentrations are achieved through evaporation. Salts, such as calcium carbonate and calcium sulfate, precipitate leaving mainly sodium and magnesium chlorides in the brine solution. When these industrial waters approach NaCl saturation (i.e., become 'crystallizer ponds'), it creates a niche for specialized microbial communities of halophiles (Oren 1994).

Decades-long examinations of saturated NaCl brines, especially from solar salterns located in Spain, Australia, and Israel (e.g., Rodriguez-Valera et al. 1981; Oren et al. 1995; Burns et al. 2004), led to the conclusion that these microbial communities are 'simple'. PCR amplification of 16S rRNA genes and visualization of communities by the fingerprinting method (denaturing gradient gel electrophoresis) showed that crystallizer ponds in Santa Pola (Spain) had limited microbial diversity compared to lower salinity ponds (Casamayor et al. 2002). Quantification of genotypes from the same saltern using florescence in situ hybridization analysis demonstrated that two species, the archaeon *Haloquadratum walsbyi* and the bacterium

Salinibacter ruber, accounted for approximately 75 % of the total cells (Anton et al. 1999, 2002). Remaining archaeal representatives span the class Halobacteria (Anton et al. 1999) and the recently proposed class Nanohaloarchaea (Ghai et al. 2011; Narasingarao et al. 2011). Representatives from the Eukaryotes and Bacteria were also found in these systems, albeit at much lower frequencies: Among the most studied are photosynthetic protists (Oren 2005), fungi (Gunde-Cimerman et al. 2009), and the bacterial groups of Bacteroidetes/Chlorobi (Anton et al. 2002) and gamma-proteobacteria (de la Haba et al. 2011). Furthermore, hypersaline environments, and saturated brines specifically, have clear boundaries from their surrounding environments and are patchily distributed across the globe. Therefore, analogous to places like the Galapagos Islands, hypersaline settings can be viewed as microbial islands of limited biodiversity where restricted dispersal fosters the evolution of unique lineages and communities.

Studies of microbially dominated island-like environments provide evidence for limited dispersal. For example, thermophilic cyanobacterial mats from Zerka Ma'in hot springs in Jordan are dominated by the *Synechococcus* C1 types and lack *Synechococcus* A/B types (Ionescu et al. 2010), while in Yellowstone they are dominated by the A/B types (Ferris et al. 1996). *Synechococcus* spp. genotyping in hot springs from around the globe demonstrates the presence of distinct site-specific cyanobacterial communities that were likely assembled by chance (e.g., Papke et al. 2003; Hongmei et al. 2005; Finsinger et al. 2008; Lau et al. 2009; Ionescu et al. 2010).

The evidence for endemic microbial populations in isolated environments suggests that allopatric speciation might be common there. For example, distribution of *Sulfolobus islandicus* strains from acidic hot springs around the world was correlated with the geographic locations from where they were cultivated, indicating that geographic isolation was driving the divergence of these populations (Whitaker et al. 2003). However, a recent study (Oh et al. 2010) suggested that hypersaline environments might be different: an archaeon *Haloquadratum walsbyi* was found in Australia, Israel, Peru, Spain, Tunisia, and Turkey.

To improve our understanding of microbial diversity in hypersaline environments, we extracted genomic content of individual cells from two geographically separated brine communities, the South Bay Salt Works saltern in Chula Vista (California, USA) and Santa Pola saltern near Alicante (Spain). Individual cells were separated by high-throughput fluorescence-activated cell sorting and deposited into microtiter plate wells (Raghunathan et al. 2005; Zhang et al. 2006; Stepanauskas and Sieracki 2007; Swan et al. 2011). For each well, whole genome multiple displacement amplification (Dean et al. 2002) and 16S rRNA gene PCR and sequencing were performed. 16S rRNA amplification after cell sorting has been previously demonstrated as a useful technique for analyzing hypersaline environments (Trigui et al. 2011). Our additional step of whole genome amplification enables sequencing of multiple genes or genomes from individual, uncultured cells (Woyke et al. 2009; Swan et al. 2011). Indeed, the genome of a nanohaloarchaeaon identified in this study was successfully sequenced (Ghai et al. 2011). In this study, we present analysis of 16S rRNA genes from 207 individual cells, which reveal that two environments harbor significantly different communities of bacteria and archaea.

Materials and methods

Sampling

Approximately 50 mL of saturated thalassohaline brines (32–35 % NaCl) were collected from the South Bay Salt Works in Chula Vista (CV) near San Diego, CA, USA (32°36′, 117°6′) and from the Santa Pola Saltern (SP), near Alicante, Spain (38°11′, 0°33′) in mid September 2009. Samples were shipped overnight directly to the Bigelow Laboratory in West Boothbay Harbor, Maine for the cell sorting, genome amplification and 16S rRNA sequencing.

Cell sorting and molecular analyses

Prokaryotic cell abundances were estimated at 10^8 cells per mL. Water sample was incubated for 10–60 min with SYTO-9 (5 µM final concentration; Invitrogen) and high nucleic acid content prokaryote cells were sorted with a MoFlo™ (Beckman Coulter) flow cytometer using a 488-nm argon laser for excitation, a 70-µm nozzle orifice, and a CyClone™ robotic arm for droplet deposition into microplates. High nucleic acid content was utilized because Halobacteria are known to be polyploid (Breuert et al. 2006) and theoretically would allow us to bias our analysis toward them. The cytometer was triggered on side scatter (see Supplemental Material Fig. S1). The "single 1 drop" mode was used for maximal sort purity, which insured the absence of non-target particles within the target cell drop and the adjacent drops. Under these sorting conditions, sorted drops contain a few tens of pL of sample surrounding the target cell (Sieracki et al. 2005), resulting in low or absent non-target DNA. The accuracy of 10 µm fluorescent bead deposition into the 384-well plates was verified by microscopically examining the presence of beads in the plate wells. Of the 2–3 plates examined on each sort day, <2 % wells were found not to contain a bead and only <0.5 % wells were found to contain more than one bead, indicating very high purity of single cells. In

addition, we verified the lack of DNA contamination in the sheath fluid and in sheath fluid lines by performing real-time multiple displacement amplification with the processed sheath fluid as the template.

Single bacterial cells were deposited into 384-well plates containing 0.6 µL per well of TE buffer. Plates were stored at −80 °C until further processing. Of the 384 wells, 315 were dedicated for single cells, 66 were used as negative controls (no droplet deposition) and 3 received 10 cells each (positive controls). The cells were lysed and their DNA was denatured using cold KOH (Raghunathan et al. 2005). Genomic DNA from the lysed cells was amplified using multiple displacement amplification (MDA) (Dean et al. 2002; Raghunathan et al. 2005) in 10 µL final volume. The MDA reactions contained 2 U/uL Repliphi polymerase (Epicentre), 1× reaction buffer (Epicentre), 0.4 mM each dNTP (Epicentre), 2 mM DTT (Epicentre), 50 mM phosphorylated random hexamers (IDT) and 1 µM SYTO-9 (Invitrogen) (all final concentration). The MDA reactions were run at 30 °C for 12–16 h, and then inactivated by 15 min incubation at 65 °C. The amplified genomic DNA was stored at −80 °C until further processing. We refer to the MDA products originating from individual cells as single amplified genomes (SAGs).

The instruments and the reagents were decontaminated for DNA prior to sorting and MDA setup, as previously described (Stepanauskas and Sieracki 2007). High molecular weight DNA contaminants in all MDA reagents were cross linked by a UV treatment in a Stratalinker (Stratagene). An empirical optimization of the UV exposure was performed to remove all detectable contaminants without inactivating the reaction. Cell sorting and MDA setup were performed in a HEPA-filtered environment. As a quality control, all MDA reaction kinetics were monitored by measuring the SYTO-9 fluorescence using FLUOstar Omega (BMG). The critical point (Cp) was determined for each MDA reaction as the time required to produce half of the maximal fluorescence. The Cp is inversely correlated to the amount of DNA template (Zhang et al. 2006). The Cp values were significantly lower in 1-cell wells compared to 0-cell wells ($p < 0.05$; Wilcoxon Two-Sample Test) in each microplate. The MDA products were diluted 50-fold in sterile TE buffer. Then 0.5 µL aliquots of the dilute MDA products served as templates in 5 µL real-time PCR screens targeting the SSU rRNA gene using bacterial primers 27F (Page et al. 2004) and 907R (Casamayor et al. 2000), archaeal primers Arch_344 and Arch_915R (Casamayor et al. 2000) and prokaryote-wide primers Prok_340F and Prok_806R (Martinez-Garcia et al. 2011). Forward (5′-GTAAAACGACGGCCAGT-3′) or reverse (5′-CAGGAAACAGCTATGACC-3′) M13 sequencing primer was appended to the 5′ end of each PCR primer to aid direct sequencing of the PCR products. All PCRs were performed using LightCycler 480 SYBR Green I Master mix (Roche) in a LightCycler® 480 II real-time thermal cycler (Roche) following previously described cycling conditions (Martinez-Garcia et al. 2011). The real-time PCR kinetics and the amplicon melting curves served as proxies detecting successful SAG target gene amplification. New, 20 µL PCR reactions were set up for the PCR-positive SAGs and the amplicons were sequenced from both ends using M13 targets and Sanger technology by Beckman Coulter Genomics. Single cell sorting, whole genome amplification and PCR were performed at the Bigelow Laboratory Single Cell Genomics Center (http://www.bigelow.org/scgc). Previous studies and recent publications using this single cell sequencing technique demonstrate the reliability of the Center's methodology with high purity of single cell MDA products (Woyke et al. 2009; Fleming et al. 2011; Heywood et al. 2011; Martinez-Garcia et al. 2011; Swan et al. 2011; Yoon et al. 2011). Cloned sequences were edited using Geneious bioinformatics and alignment software (http://www.geneious.com/). 16S rRNA sequences are available in GenBank under accession numbers JN839733-JN839939.

Ribosomal RNA classification

128 sequences from Chula Vista (CV) sample site and 79 16S rRNA sequences from Santa Pola (SP) sample site were placed taxonomically using Ribosomal Database Project's (RDP) Naïve Bayesian Classifier (Wang et al. 2007) with 80 % confidence threshold and visualized as heat maps in RDP Taxomatic (http://rdp.cme.msu.edu/taxomatic). Sequences on heatmaps' axes were arranged according to the RDP classification scheme. RDP database Release 10, Update 26 was used in these analyses (Cole et al. 2009).

Archaeal alignments

16 archaeal sequences from a 'high-intracellular DNA' fraction of the solar saltern sample from Sfax, Tunisia (SFX) (Trigui et al. 2011) were downloaded from GenBank. 76 SP and 55 CV sequences classified by RDP classifier as archaeal, as well as the 16 SFX sequences, were uploaded to myRDP and aligned to the RDP rRNA alignment. These alignments were further used in community composition and phylogenetic analysis (see below).

Community composition analysis of archaeal and bacterial sequences

Sequences within SP, CV and SFX samples were grouped into operational taxonomic units (OTUs) in MOTHUR v.1.20.0 (Schloss et al. 2009) at distance cutoffs of 0.01,

0.03 and 0.05, using average neighbor distance method on uncorrected pairwise distances. Beta diversity between the three samples was calculated with these OTU assignments. Communities were compared using *P* test (Martin 2002), as implemented in the Unifrac program (Lozupone et al. 2006).

Phylogenetic analysis of archaeal sequences

Using RDP Sequence Match tool, similar sequences were added to the dataset of CV and SP Archaeal 16S rRNA sequences using the following criteria: for each CV and SP sequence, a maximum of two type strain matches, one isolate match and one uncultured match were extracted. The alignment of selected sequences was downloaded from RDP and a phylogenetic tree was reconstructed in the FastTree program version 2.1.3 (Price et al. 2009), with 100 bootstrap samples.

In a second sequence data set, all available matches to the putative archaeal class of Nanohaloarchaea, as well as one type strain representative per genus from Halobacteria class and one type strain representative per family in the rest of Euryarchaea, were retrieved from RDP. A crenarchaeote *Aeropyrum pernix* was added as an outgroup. The alignment of the selected sequences was downloaded from RDP and phylogenetic tree was reconstructed in RAxML version 7.0.4 (Stamatakis 2006) under GTR+Gamma model, with 100 bootstrap samples.

Archaeal class-level divergence analysis

For eight euryarchaeal classes represented in RDP database by at least some 16S rRNA isolate sequences $\geq$1200 nt long, pairwise Jukes–Cantor distances of all sequences within each class were downloaded from the RDP. Distances were also obtained for 39 sequences within a novel uncultured clade of Nanohaloarchaea, and for broader groups of 'Halobacteria + Nanohaloarchaea' and 'Halobacteria + Methanomicrobia'. The class Methanomicrobia was chosen because, after Halobacteria, it was the euryarchaeal class best represented in the RDP.

Analyses of Bacteroidetes sequences

Using RDP Sequence Match tool, similar sequences were added to the dataset of 42 CV and SP 16S rRNA sequences classified as Bacteroidetes using the following criteria: for each CV and SP sequence a maximum of two type strain or isolate matches, and three uncultured matches were extracted. The alignment of selected sequences was downloaded from the RDP, and a phylogenetic tree was reconstructed with RAxML version 7.0.4 (Stamatakis 2006) under GTR+Gamma model, with 100 bootstrap samples.

Analyses of proteobacterial sequences

Using RDP Sequence Match tool, 30 CV and SP 16S rRNA sequences classified as proteobacteria were complemented by similar sequences using the following criteria: for each sequence, a maximum of one type strain match, one isolate match, and three uncultured matches were extracted. The dataset was enhanced by 44 proteobacterial 16S rRNA sequences previously observed in salterns by Maturrano et al. (2006) and Jiang et al. (2006). The alignment was downloaded from RDP, and a phylogenetic tree was reconstructed in the FastTree program version 2.1.3 (Price et al. 2009), with 100 bootstrap samples. Based on this initial phylogenetic reconstruction, the dataset was pruned from redundant sequences. The phylogenetic tree for the reduced dataset was reconstructed in RAxML version 7.0.4 (Stamatakis 2006) under GTR+Gamma model, with 100 bootstrap samples.

Results

Taxonomic distribution of the amplified 16S rRNA sequences from Chula Vista (CV) and Santa Pola (SP) saltern samples indicates that representatives of the archaeal class Halobacteria (in particular, *Natromonas*, *Halorubrum*, *Haloquadratum* and *Haloarcula* genera) and of the bacterial phylum Bacteroidetes (in particular, genus *Salinibacter*) dominate both sites, as expected for halophilic environments (Fig. 1). A notable fraction of sorted cells from each sample belongs to unclassified *Halobacteriaceae* (see below for the within-order differences). Additionally, both sites contain a diverse cluster of sequences from the recently discovered putative class of Nanohaloarchaea (Narasingarao et al. 2011) (this class is not yet formally recognized due to lack of cultivated isolates). However, the two samples vary dramatically in the rest of their community composition. The CV sample contains a large proportion (~50 %) of Bacteria: a diverse representation within the genus *Salinibacter*, a cluster of unclassified Bacteroidetes with low divergence (Fig. 1a and Supplemental Material Fig. S2), and a wide range of unclassified Proteobacteria, dominated by the γ-Proteobacteria (Fig. 1a and Supplemental Material Fig. S3). Of the detected *Halobacteriaceae* in the CV sample the majority is unclassified (Fig. 1a). On the other hand, the SP sample contains only few bacterial members (all from the genus *Salinibacter*), and the majority of *Halobacteriaceae* representatives belong to the commonly isolated genera (Fig. 1b).

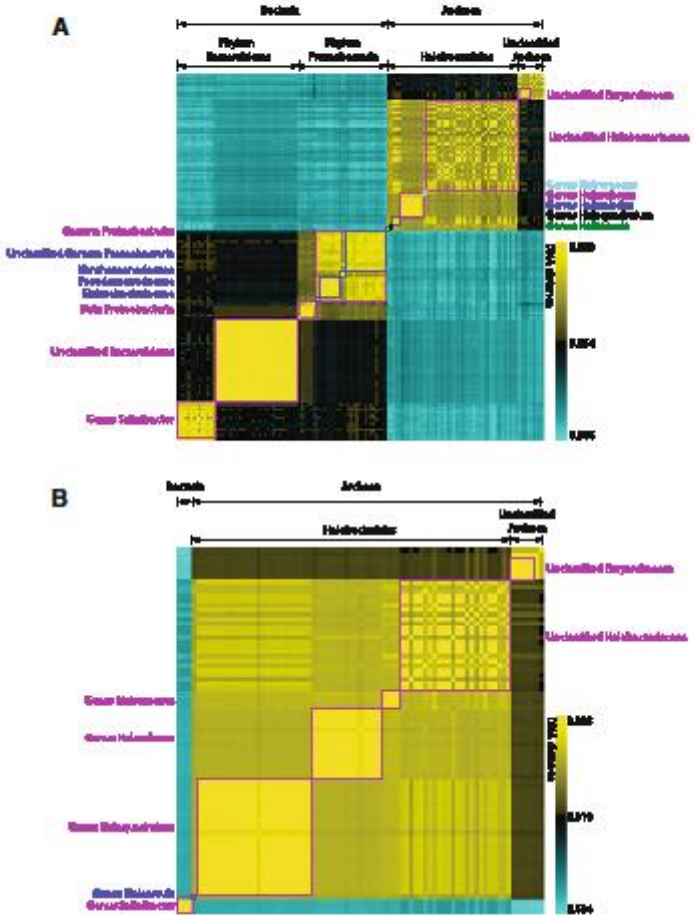


Fig. 1 Taxonomic classification of 16S rRNA sequences in a Chula Vista sample and b Santa Pola sample. 16S rRNA sequences are arranged on the axes according to their classification in RDP database and are *color-coded* by their DNA distance to each other (ranging from *yellow*, which corresponds to very closely related sequences, to *cyan*, which designates domain-level divergence). Note that *color-coded distances* have different scale in the two panels (color figure online)

We complemented our data set with archaeal and bacterial 16S rRNA sequences from a recent flow cytometry study of a saturated NaCl brine sample from a solar saltern in Tunisia (Trigui et al. 2011), referred throughout the manuscript as SFX sample. Cell sorting from SFX sample site, similar to ours, was based on DNA content and side scattering, making the dataset useful for direct comparison. Trigui et al. (2011), however, did not use whole genome amplification and PCR-amplified 16S rRNA genes directly from the lysed sorted cells. Regardless of 16S rRNA distance cutoffs used (0.01, 0.03 and 0.05) to circumscribe OTUs, the community compositions of CV, SP and SFX samples are significantly different ($p < 0.001$ in P test) for both bacterial and archaeal fractions (Fig. 2). SFX sample has very little overlap with CV, while SP shares more OTUs with both CV and SFX sites.

Phylogenetic analysis of SP and CV 16S rRNA sequences in the context of known haloarchaeal diversity (cultivated or not) is shown in Fig. 3. A group of Archaea identified in Fig. 1 forms a cluster on the phylogenetic tree separated from the rest of Halobacteria by a long branch. SFX crystallizer pond sample lacks representatives of this group. Search of the RDP database for similar sequences revealed that these divergent halophilic Archaea have been observed in the environmental samples from hypersaline environments in Kenya, Algerian Sahara, Tunisia and Australia (Grant et al. 1999; Baati et al. 2010; Oh et al. 2010). Grant et al. (1999) suggested that the two

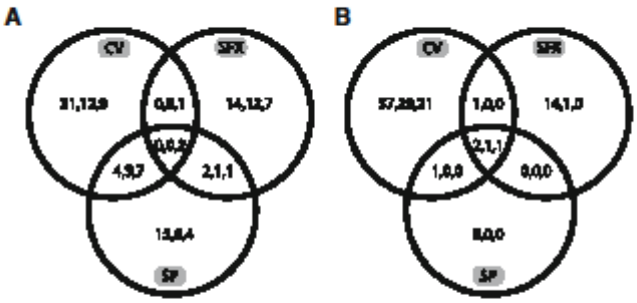


Fig. 2 Comparison of archaeal (a) and bacterial (b) community composition across three sampled sites. *Comma-separated numbers inside the Venn diagram* refer to number of OTUs defined at distance cutoffs of 0.01, 0.03 and 0.05, respectively. The archaeal and bacterial populations of these three geographically separated saltern communities are significantly different in composition (*P* test; $p < 0.001$). *SP* Santa Pola, *CV* Chula Vista, *SFX* Sfax sample sites

phylotypes observed in East African salterns were representatives of a divergent euryarchaeal branch. Analysis of three Australian crystallizer ponds further pointed out the abundance of these deep-branching lineages of *Halobacteria* and referred to the group as "MSP8-clade" (Oh et al. 2010). Our subsequent search of GenBank database retrieved additional matches from recent metagenomic studies of Chula Vista and Santa Pola salterns and of Lake Tyrell (Ghai et al. 2011; Narasingarao et al. 2011). Narasingarao et al. (2011) noted that the lineage is divergent from known *Halobacteria* and made a proposition for a novel class 'Nanohaloarchaea'.

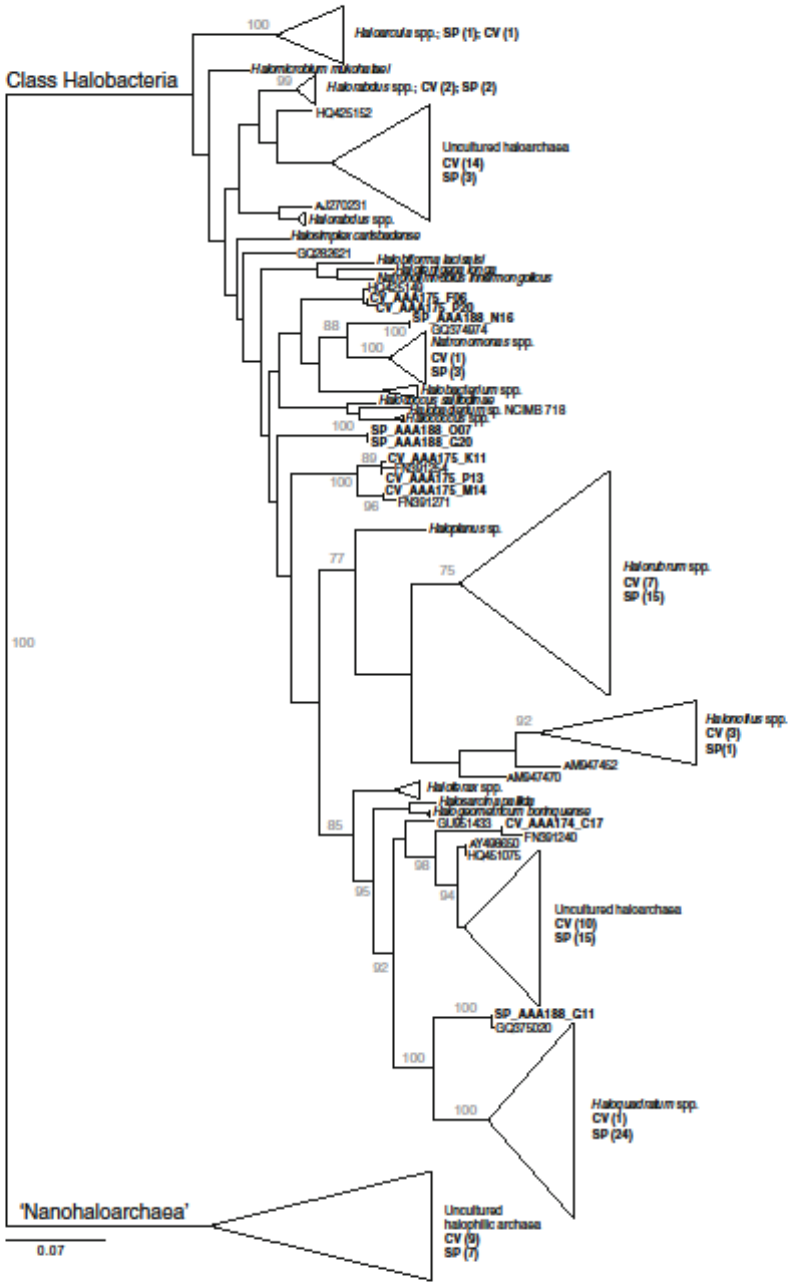
Our further phylogenetic analysis with new sequence data confirms that these divergent sequences indeed do not group within *Halobacteria* or as its deep-branching members, but instead form a sister clade with the class (Fig. 4). On the phylogenetic tree, the putative 'Nanohaloarchaea' clade appears as distant from Halobacteria as it is from other euryarchaeal classes. To quantify this conjecture, we examined distances within as well as between Euryarchaea classes (Table 1). The minimum observed distance between members of the novel clade and *Halobacteria* is roughly equal to the largest observed distance within *Halobacteria*, and the range of distances between *Halobacteria* and the novel clade is comparable to distances between classes *Halobacteria* and *Methanomicrobia*. This analysis indicates that putative 'Nanohaloarchaea' clade is sufficiently separated from *Halobacteria* to be elevated to the level of a novel class of halophilic Archaea, should its representatives be cultivated.

Discussion

Notorious difficulty in microbial culturability (Rollins and Colwell 1986), bias in 16S rRNA primers (Suzuki and Giovannoni 1996; Lueders and Friedrich 2003) and limitations to assembling complete genomes from metagenomic sequence data (Rusch et al. 2007) makes the cell sorting technique followed by genome amplification an attractive approach to examine microbial diversity in halophilic environments. Our study focused only on the fraction of the community that displayed high nucleic acid content. Given a limited number of cells that could be collected and analyzed, we were concerned that without a sorting criterion, we will examine only the numerically dominant cells, and hence underestimate microbial diversity. High DNA content due to polyploidy was suggested to be common among Halobacteria (Breuert et al. 2006), and we conjectured that this sorting criterion would allow us to preferentially focus on a halobacterial fraction of the community. The screen was only partially successful in that respect, since a large bacterial fraction was recovered in one of the analyzed salterns (Fig. 1a). However, it demonstrates that high DNA content sorting may help highlighting less dominant (and therefore less studied) members of a halophilic community, such as yet unclassified bacteria and archaea. The cell sorting technique will allow future studies of gene content of these unclassified lineages (such as Nanohaloarchaea) without extra effort of gathering and assembling fragmented metagenomic sequence data.

Our observation of different prokaryotic communities in two geographically separated hypersaline environments is qualitatively in agreement with other studies of the same

Fig. 3 Unrooted maximum likelihood phylogenetic tree of Archaeal ▶ 16S rRNA sequences from CV and SP samples and of related sequences from RDP database. Large clades were collapsed into *wedges*. Selected bootstrap supports above 80 % are shown. Environmental sequences are designated by their GenBank accession numbers. Sequences obtained in this study are shown in *bold*


Class Halobacteria
'Nanohaloarchaea'
0.07
Haloarcula spp.; SP (1); CV (1)
Halomicrobium mukohataei
Halorabdus spp.; CV (2); SP (2)
HQ425152
Uncultured haloarchaea
CV (14)
SP (3)
AJ270231
Halorabdus spp.
Halosimplex carlsbadense
GQ283621
SP_AAA188_N16
GQ374074
Natronomonas spp.
CV (1)
SP (3)
Halobacterium spp.
Halobacterium sp. NCIMB 718
SP_AAA188_O07
SP_AAA188_C20
CV_AAA175_K11
CV_AAA175_P13
CV_AAA175_M14
FN391271
Halopiger spp.
Halorubrum spp.
CV (7)
SP (15)
Halonotius spp.
CV (3)
SP(1)
AM947452
AM947470
Halobaculum gomorrense
GU051433
CV_AAA174_C17
FN391240
AY496950
HQ451075
Uncultured haloarchaea
CV (10)
SP (15)
SP_AAA188_C11
GQ375020
Haloquadratum spp.
CV (1)
SP (24)
Uncultured
halophilic archaea
CV (9)
SP (7)


0.2
100
95
100
98
84
GQ374069
SP-AAA188-G17
AB012050
AB012050
CV-AAA175-N23
CV-AAA174-D18
GQ375000
EU860371
CV-AAA174-M04
GQ374944
GQ374932
GQ374936
GQ374938
GQ374930
FN391274
GQ374934
GQ374940
GQ374924
GQ374926
CV-AAA174-O07
GQ374923
GQ374922
CV-AAA175-B14
GQ374947
SP-AAA188-E11
CV-AAA175-B18
SP-AAA188-I03
SP-AAA188-C15
GQ375010
GQ375019
GQ375009
GQ375014
SP-AAA188-M06
CV-AAA174-C13
SP-AAA188-E09
GQ374979
CV-AAA175-C13
SP-AAA188-M04
CV-AAA174-O06
'Nanohaloarchaea'
Natronomonas pharaonis
Halorhabdus utahensis
Haloferax mukohataei
Halobacterium salinarum
Haloarcula quadrata
Class Halobacteria
Methanospirillum hungatei
Methanomicrobium mobile
Methanocorpusculum parvum
Methanoculleus palmolei
Methanosarcina barkeri
Methanococcus shengliensis
Methanosaeta thermophila
Thermogymnomonas acidicola
Picrophilus torridus
Ferroplasma acidiphilum
Thermoplasma acidophilum
Methanosphaera stadtmanae
Methanothermococcus thermolithotrophicus
Methanocaldococcus vulcanius
Archaeoglobus profundus
Pyrococcus horikoshii
Methanopyrus kandleri
Aeropyrum pernix
Other Euryarchaea

◀ **Fig. 4** Maximum likelihood phylogenetic tree of 16S rRNA sequences from Nanohaloarchaea and from representatives of Euryarchaea. The tree is rooted with *Aeropyrum pernix*. Environmental sequences are designated by their GenBank accession numbers, and all named species are type strains. Sequences obtained in this study are shown in *bold*. Only selected bootstrap support values are depicted

sites: while in SP Archaea comprised 87 % of the community (Ghai et al. 2011), in the CV, they constituted only 54 % (Rodriguez-Brito et al. 2010); in CV, *Haloquadratum* spp. were either co-dominant among halobacterial genera (Rodriguez-Brito et al. 2010), or not observed (Bidle et al. 2005), while they were described as dominant in SP (Anton et al. 1999; Ghai et al. 2011). Both CV and SP salterns are considered stable in their community composition (Rodriguez-Brito et al. 2010; Ghai et al. 2011). Quantitatively, however, our study produces different relative abundance of commonly found bacterial and archaeal genera, which may reflect differences in methodology (e.g., we screened for cells with high DNA content, while the above-mentioned studies did not) and in the extent of sampling across the studies.

Observed compositional difference in geographically separated hypersaline environments is discordant with the evidence for extensive wind-driven atmospheric dispersal of microorganisms (e.g., Kellogg and Griffin 2006). The immense sizes of prokaryotic populations combined with their small cellular dimensions are hypothesized to promote dispersal between similar environments, and to prevent endemism and localized extinctions (Finlay 2002; Fenchel 2003). Furthermore, genera including *Haloquadratum*, *Halorubrum* and *Haloarcula* are found worldwide in hypersaline systems with different ionic composition (e.g., Ward et al. 2000; Benlloch et al. 2001; Radax et al. 2001; Burns et al. 2004; Purdy et al. 2004; Bidle et al. 2005; Sorensen et al. 2005; Walsh et al. 2005; Papke et al. 2007; Mutlu et al. 2008; Oh et al. 2010; Dyall-Smith et al. 2011), suggesting unencumbered dispersal and niche invasion of halophiles.

How do we reconcile the seemingly contradictory observations that site composition is unique, yet similar halophilic OTUs are widely dispersed? To explain the disagreement, we hypothesize that the rate of evolution (i.e., rates of mutation, gene gain/loss, and recombination) is faster than the combined rate of dispersal and invasion. Three studies support our hypothesis (Whitaker et al. 2003, 2005; Dyall-Smith et al. 2011). Highly recombinogenic populations of *Sulfolobus islandicus* strains within geographically isolated hot springs are ~0.1 % divergent, yet strains between geographically distant sites accumulated larger amounts of divergence (~1 %) over short geological time periods (Whitaker et al. 2003, 2005). This indicates that population diversity is maintained by dispersal but reduced through recombination. Analogously, genomes of two closely related *Haloquadratum walsbyi* strains from Spain and Australia have variable gene content and are on average 1.4 % nucleotide sequence divergent across shared genes (Dyall-Smith et al. 2011). We suggest that these two strains have been geographically separated for a long time, since co-occurring Halobacteria are highly recombinogenic in both the laboratory (Naor et al. 2012) and nature (Boucher et al. 2004; Cuadros-Orellana et al. 2007; Rhodes et al. 2011; Andam et al. 2012) and can undergo genetic homogenization that keeps divergence below 1 % (Papke et al. 2007; Papke 2009). This hypothesis emphasizes that dispersal is an ongoing and frequent aspect of halophile

Table 1 Distances within and between selected Euryarchaeal classes and the proposed novel Archaeal class of 'Nanohaloarchaea'

Archaeal class[a]	Number of sequences[b]	Distance observed within the group
Maximum distances		
Archaeoglobi	20	0.08
Halobacteria	1062	0.26
Methanobacteria	156	0.13
Methanococci	68	0.18
Methanomicrobia	241	0.28
Methanopyri	3	0.01
Thermococci	201	0.08
Thermoplasmata	23	0.28
'Nanohaloarchaea'	39 (16 are from this study)	0.17
Halobacteria + 'Nanohaloarchaea'	1101 (16 are from this study)	0.48
Halobacteria + Methanomicrobia	1303	0.39
Minimum distances		
Halobacteria + 'Nanohaloarchaea'	1101 (16 are from this study)	0.25
Halobacteria + Methanomicrobia	1303	0.22

[a] 'Nanohaloarchaea' is a putative novel class

[b] For the established Archaeal classes, only cultured representatives were used. For 'Nanohaloarchaea', all available sequences were used

biology, and that historical contingency (i.e., colonization order; Jousset et al. 2011; Langenheder and Szekely 2011) might be an important determinant of halophilic community composition.

Acknowledgments This work was supported by grants to R.T.P from the National Science Foundation (award numbers 0919290 and 0800024), the U.S.–Israel Binational Science Foundation (award number 2007043), and NASA Astrobiology: Exobiology and Evolutionary Biology Program (award number NNX12AD70G). We extend special thank you to Forest Rohwer (San Diego State University, USA) and Francisco Rodríguez-Valera (University Miguel Hernández, Spain) who sampled the Chula Vista and Santa Pola salterns, respectively.

References

Andam CP, Harlow TJ, Papke RT, Gogarten JP (2012) Ancient origin of the divergent forms of leucyl-tRNA synthetases in *Halobacteriales*. BMC Evol Biol 12:85

Anton J, Llobet-Brossa E, Rodriguez-Valera F, Amann R (1999) Fluorescence in situ hybridization analysis of the prokaryotic community inhabiting crystallizer ponds. Environ Microbiol 1:517–523

Anton J, Oren A, Benlloch S, Rodriguez-Valera F, Amann R, Rossello-Mora R (2002) *Salinibacter ruber* gen. nov., sp. nov., a novel, extremely halophilic member of the Bacteria from saltern crystallizer ponds. Int J Syst Evol Microbiol 52:485–491

Baati H, Guermazi S, Gharsallah N, Sghir A, Ammar E (2010) Novel prokaryotic diversity in sediments of Tunisian multipond solar saltern. Res Microbiol 161:573–582

Benlloch S, Acinas SG, Anton J, Lopez-Lopez A, Luz SP, Rodriguez-Valera F (2001) Archaeal biodiversity in crystallizer ponds from a solar saltern: culture versus PCR. Microb Ecol 41:12–19

Bidle K, Amadio W, Oliveira P, Paulish T, Hicks S, Earnest C (2005) A phylogenetic analysis of haloarchaea found in a solar saltern. Bios 76:89–96

Boucher Y, Douady CJ, Sharma AK, Kamekura M, Doolittle WF (2004) Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. J Bacteriol 186:3980–3990

Breuert S, Allers T, Spohn G, Soppa J (2006) Regulated polyploidy in halophilic archaea. PLoS One 1:e92

Burns DG, Camakaris HM, Janssen PH, Dyall-Smith ML (2004) Combined use of cultivation-dependent and cultivation-independent methods indicates that members of most haloarchaeal groups in an Australian crystallizer pond are cultivable. Appl Environ Microbiol 70:5258–5265

Casamayor EO, Schafer H, Baneras L, Pedros-Alio C, Muyzer G (2000) Identification of and spatio-temporal differences between microbial assemblages from two neighboring sulfurous lakes: comparison by microscopy and denaturing gradient gel electrophoresis. Appl Environ Microbiol 66:499–508

Casamayor EO, Massana R, Benlloch S, Ovreas L, Diez B, Goddard VJ et al (2002) Changes in archaeal, bacterial and eukaryal assemblages along a salinity gradient by comparison of genetic fingerprinting methods in a multipond solar saltern. Environ Microbiol 4:338–348

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ et al (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res 37:D141–D145

Cuadros-Orellana S, Martin-Cuadrado AB, Legault B, D'Auria G, Zhaxybayeva O, Papke RT, Rodriguez-Valera F (2007) Genomic plasticity in prokaryotes: the case of the square haloarchaeon. ISME J 1:235–245

de la Haba RR, Marquez MD, Papke RT, Ventosa A (2011) Multilocus sequence analysis (MLSA) of the family *Halomonadaceae*. Int J Syst Evol Microbiol 62:520–538

Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P et al (2002) Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci USA 99:5261–5266

Dyall-Smith ML, Pfeiffer F, Klee K, Palm P, Gross K, Schuster SC et al (2011) *Haloquadratum walsbyi*: limited diversity in a global pond. PLoS One 6:e20968

Fenchel T (2003) Microbiology. Biogeography for bacteria. Science 301:925–926

Ferris MJ, Muyzer G, Ward DM (1996) Denaturing gradient gel electrophoresis profiles of 16S rRNA-defined populations inhabiting a hot spring microbial mat community. Appl Environ Microbiol 62:340–346

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

Finsinger K, Scholz I, Serrano A, Morales S, Utibe-Lorio L, Mora M et al (2008) Characterization of true-branching cyanobacteria from geothermal sites and hot springs of Costa Rica. Environ Microbiol 10:460–473

Fleming EJ, Langdon AE, Martinez-Garcia M, Stepanauskas R, Poulton NJ, Masland ED, Emerson D (2011) What's new is old: resolving the identity of *Leptothrix ochracea* using single cell genomics, pyrosequencing and FISH. PLoS One 6:e17769

Ghai R, Pasic L, Fernandez AB, Martin-Cuadrado AB, Mizuno CM, McMahon KD et al (2011) New abundant microbial groups in aquatic hypersaline environments. Sci Rep 1:135

Grant S, Grant WD, Jones BE, Kato C, Li L (1999) Novel archaeal phylotypes from an East African alkaline saltern. Extremophiles 3:139–145

Gunde-Cimerman N, Ramos J, Plemenitas A (2009) Halotolerant and halophilic fungi. Mycol Res 113:1231–1241

Heywood JL, Sieracki ME, Bellows W, Poulton NJ, Stepanauskas R (2011) Capturing diversity of marine heterotrophic protists: one cell at a time. ISME J 5:674–684

Hongmei J, Aitchison JC, Lacap DC, Peerapornpisal Y, Sompong U, Pointing SB (2005) Community phylogenetic analysis of moderately thermophilic cyanobacterial mats from China, the Philippines and Thailand. Extremophiles 9:325–332

Ionescu D, Hindiyeh M, Malkawi H, Oren A (2010) Biogeography of thermophilic cyanobacteria: insights from the Zerka Ma'in hot springs (Jordan). FEMS Microbiol Ecol 72:103–113

Jiang H, Dong H, Zhang G, Yu B, Chapman LR, Fields MW (2006) Microbial diversity in water and sediment of Lake Chaka, an Athalassohaline lake in northwestern China. Appl Environ Microbiol 72:3832–3845

Jousset A, Schulz W, Scheu S, Eisenhauer N (2011) Intraspecific genotypic richness and relatedness predict the invasibility of microbial communities. ISME J 5:1108–1114

Kellogg CA, Griffin DW (2006) Aerobiology and the global transport of desert dust. Trends Ecol Evol 21:638–644

Langenheder S, Szekely AJ (2011) Species sorting and neutral processes are both important during the initial assembly of bacterial communities. ISME J 5:1086–1094

Lau MC, Aitchison JC, Pointing SB (2009) Bacterial community composition in thermophilic microbial mats from five hot springs in central Tibet. Extremophiles 13:139–149

Lozupone C, Hamady M, Knight R (2006) UniFrac—an online tool for comparing microbial community diversity in a phylogenetic context. BMC Bioinformatics 7:371

Lueders T, Friedrich MW (2003) Evaluation of PCR amplification bias by terminal restriction fragment length polymorphism

analysis of small-subunit rRNA and mcrA genes by using defined template mixtures of methanogenic pure cultures and soil DNA extracts. Appl Environ Microbiol 69:320–326
Martin AP (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. Appl Environ Microbiol 68:3673–3682
Martinez-Garcia M, Swan BK, Poulton NJ, Gomez ML, Masland D, Sieracki ME, Stepanauskas R (2011) High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. ISME J 6:113–123
Maturrano L, Santos F, Rossello-Mora R, Anton J (2006) Microbial diversity in Maras salterns, a hypersaline environment in the Peruvian Andes. Appl Environ Microbiol 72:3887–3895
Muthu MB, Martinez-Garcia M, Santos F, Pena A, Guven K, Anton J (2008) Prokaryotic diversity in Tuz Lake, a hypersaline environment in inland Turkey. FEMS Microbiol Ecol 65:474–483
Naor A, Lapierre P, Mevarech M, Papke RT, Gophna U (2012) Low species barriers in halophilic archaea and the formation of recombinant hybrids. Curr Biol 22:1444–1448
Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ et al (2011) De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J 6:81–93
Oh D, Porter K, Russ B, Burns D, Dyall-Smith M (2010) Diversity of *Haloquadratum* and other haloarchaea in three, geographically distant, Australian saltern crystallizer ponds. Extremophiles 14:161–169
Oren A (1994) The ecology of the extremely halophilic archaea. FEMS Microbiol Rev 13:415–440
Oren A (2005) A hundred years of Dunaliella research: 1905–2005. Saline Syst 1:2
Oren A, Kuhl M, Karsten U (1995) An endoevaporitic microbial mat within a gypsum crust: zonation of phototrophs, photopigments, and light penetration. Mar Ecol Prog Ser 128:151–159
Page KA, Connon SA, Giovannoni SJ (2004) Representative freshwater bacterioplankton isolated from Crater Lake, Oregon. Appl Environ Microbiol 70:6542–6550
Papke RT (2009) A critique of prokaryotic species concepts. Methods Mol Biol 532:379–395
Papke RT, Ramsing NB, Bateson MM, Ward DM (2003) Geographical isolation in hot spring cyanobacteria. Environ Microbiol 5:650–659
Papke RT, Zhaxybayeva O, Feil EJ, Sommerfeld K, Muise D, Doolittle WF (2007) Searching for species in haloarchaea. Proc Natl Acad Sci USA 104:14092–14097
Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26:1641–1650
Purdy KJ, Cresswell-Maynard TD, Nedwell DB, McGenity TJ, Grant WD, Timmis KN, Embley TM (2004) Isolation of haloarchaea that grow at low salinities. Environ Microbiol 6:591–595
Radax C, Gruber C, Stan-Lotter H (2001) Novel haloarchaeal 16S rRNA gene sequences from Alpine Permo-Triassic rock salt. Extremophiles 5:221–228
Raghunathan A, Ferguson HR Jr, Bornarth CJ, Song W, Driscoll M, Lasken RS (2005) Genomic DNA amplification from a single bacterium. Appl Environ Microbiol 71:3342–3347
Rhodes ME, Spear JR, Oren A, House CH (2011) Differences in lateral gene transfer in hypersaline versus thermal environments. BMC Evol Biol 11:199
Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M et al (2010) Viral and microbial community dynamics in four aquatic environments. ISME J 4:739–751
Rodriguez-Valera F, Ruiz-Berraquero F, Ramos-Cormenzana A (1981) Characteristics of the heterotrophic bacterial populations in hypersaline environments of different salt concentrations. Microb Ecol 7:235–243
Rollins DM, Colwell RR (1986) Viable but nonculturable stage of *Campylobacter jejuni* and its role in survival in the natural aquatic environment. Appl Environ Microbiol 52:531–538
Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S et al (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5:e77
Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB et al (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541
Sieracki ME, Poulton NJ, Crosbie N (2005) Automated isolation techniques for microalgae. In: Anderson R (ed) Algal culturing techniques. Elsevier Academic, New York
Sorensen KB, Canfield DE, Teske AP, Oren A (2005) Community composition of a hypersaline endoevaporitic microbial mat. Appl Environ Microbiol 71:7352–7365
Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690
Stepanauskas R, Sieracki ME (2007) Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. Proc Natl Acad Sci USA 104:9052–9057
Suzuki MT, Giovannoni SJ (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl Environ Microbiol 62:625–630
Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D et al (2011) Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. Science 333:1296–1300
Trigui H, Masmoudi S, Brochier-Armanet C, Barani A, Gregori G, Denis M et al (2011) Characterization of heterotrophic prokaryote subgroups in the Sfax coastal solar saltern by combining flow cytometry cell sorting and phylogenetic analysis. Extremophiles 15:347–358
Walsh DA, Papke RT, Doolittle WF (2005) Archaeal diversity along a soil salinity gradient prone to disturbance. Environ Microbiol 7:1655–1666
Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267
Ward BB, Martino DP, Diaz MC, Joye SB (2000) Analysis of ammonia-oxidizing bacteria from hypersaline Mono Lake, California, on the basis of 16S rRNA sequences. Appl Environ Microbiol 66:2873–2881
Whitaker RJ, Grogan DW, Taylor JW (2003) Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science 301:976–978
Whitaker RJ, Grogan DW, Taylor JW (2005) Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. Mol Biol Evol 22:2354–2361
Woyke T, Xie G, Copeland A, Gonzalez JM, Han C, Kiss H et al (2009) Assembling the marine metagenome, one cell at a time. PLoS One 4:e5299
Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH et al (2011) Single-cell genomics reveals organismal interactions in uncultivated marine protists. Science 332:714–717
Zhang K, Martiny AC, Reppas NB, Barry KW, Malek J, Chisholm SW, Church GM (2006) Sequencing genomes from single cells by polymerase cloning. Nat Biotechnol 24:680–686

Chapter 3 : Analysis of geographically separated hypersaline environments reveals uniquely constituted haloarchaeal communities.

3.1 Abstract

Allopatric speciation is caused by geographic isolation, and is a dominant force in multicellular eukaryotes. To what extent this mechanism governs microbial speciation is less clear. Here we study halobacterial community diversity from geographically distant island-like saturated brine (~35% NaCl) environments and estimate levels of endemism and dispersal. We used PCR amplification of the bacteriorhodopsin from environmental DNA as a molecular marker since it is almost uniquely expressed in Halobacteria, it usually occurs as a single copy in genomes, and it is not as conserved as the 16S rRNA gene. We also compared metagenomes using the 16S rRNA gene, RNA polymerase B subunit, and bacteriorhodopsin as well. Alpha and beta diversity analysis within and between the sampling sites revealed that the vast majority of sequences were endemic, and that more rarely dispersal occurs between sites. Community composition is unique and this uniqueness is not credited to geographic proximity nor ecological similarity. Together, these observations suggest that dispersal occurs regularly, but that its rate is slow compared to the pace of evolution. Therefore, dispersal followed by endemism may be a major mechanism for escaping homogenization forces that maintain species cohesion for all members of the class Halobacteria.

3.2 Materials and Methods

3.2.1 Sequence acquisition from environmental DNA

3.2.1.1 DNA isolation and PCR amplification of *bop* gene

Saturated brine samples (~34%NaCl) were collected from five geographically distant hypersaline sites listed in Figure 3-1 and table 3-1. Environmental DNA from each site was isolated using the following protocol published in the Halohandbook [142]. Briefly, the saturated brine samples were spun at 15,000rpm for 10 minutes to pellet cells. 400µl of distilled and deionized water was added to the pellet to lyse cells by osmotic shock. Lysates were then placed in a heating block maintained at 70ºC for 10 minutes to inactivate proteins. A working solution of the template DNA for PCR amplification was prepared by making a 1:200 dilution of the crude environmental DNA stock.

The bacteriorhodopsin gene was amplified from the working template DNA solutions for each of the environmental sites. The primers bop401F and bop795R were adapted from [135] and modified to carry the M13 forward and reverse sequences respectively to give bopF_poly and bopR_poly. The polymerase Phusion (New England BioLabs) was used to produce high fidelity copies of template. The following PCR cycle protocol was used: One minute initial denaturation at 94°C, followed by 30 cycles of 30 seconds at 94°C, 30 seconds at 53°C and 45 seconds at 68°C. Final elongation occurred at 68°C for 5 minutes. PCR reactions contained Phusion polymerase [2 units/µl], 0.25 µl; 5x GC buffer, 5 µl; DMSO [100%], 2.5 µl; dNTPs [100 mM], 1 µl; forward and reverse primers [10 µM], 0.5 µl each; genomic DNA [20ng/µl], 1.0 µl; and 1.0 µl of dH_2O. Acetamide [25% w/v], 13.25 µl was added to the reaction to improve specificity of primer binding and DMSO was used in order to facilitate denaturation of the high G+C template.

3.2.1.2 Cloning, Plasmid Isolation and Sequence acquisition

PCR products of the desired length were isolated from a 1% (w/v) agarose gel using the SV Gel & PCR Clean-Up System (Promega) and then cloned using the Zero Blunt TOPO Cloning Kit (Invitrogen) according to the manufacturer's directions. The recombinant plasmid was isolated from the clones for each sample using the Wizard Plus SV Minipreps DNA Purification System (Promega) according to manufacturer's instructions. Plasmids were tested for presence of an insert by restriction digestion with *Eco*RI. The purified plasmids with cloned inserts were sent to GENEWIZ Inc. for sequencing. Sequences from earlier studies on hypersaline sites using bacteriorhodopsin were also included: Santa Pola, Spain [135, 143]; Secovlje solar salterns in Secovlje, Slovenia [136]; Exportadora de Sal (ESSA) evaporative saltern in Guerrero Negro, Baja CA S., Mexico [13], a saltern from Chiku, Taiwan, Eilat solar saltern (described in Chapter 4) and Chinese salt lakes. Accession numbers for this study are: Accession numbers from sequences obtained from NCBI are: The sequences will be submitted soon. .

3.2.2 Sequence acquisition from available metagenomes

3.2.2.1 Extraction of *bop,* 16S rRNA, and *rpoB* genes from metagenomes

The *bop*, 16S rRNA, and *rpoB* gene sequences from the following metagenomic sequence read archive (SRA) datasets: Isla Cristina Saltern (IC21) [144]; Santa Pola Salterns (SS33 and SS37) [63, 145]; Lake Tyrrell [86]; Cahuil Lagoon, Chile [146]; and Chula Vista Salterns [147] were extracted by performing stand-alone BLAST runs. Sequence data for each gene from the haloarchaeal genomes were used as inputs. BLASTN was used for extracting the 16S rRNA gene while TBLASTX was employed for *bop* and *rpoB.* Obtained rhodopsin sequences were aligned and everything except *bop* was filtered out.

3.2.3 Sequence analysis

The PCR amplified *bop* sequences (PCR-*bop*), 380bp in length, were edited using the commercial software Geneious 4.8.3 Sequences for each gene, both PCR amplified and metagenome derived, were aligned using MUSCLE [148] and alignments were edited using MACCLADE 4.08 [149].

3.2.4 Halobacterial diversity analysis

MOTHUR v.1.20.0 [138] was used to cluster the sequences into operational taxonomic units (OTUs) using the average neighbor method of clustering at 99 and 95% similarities for PCR amplified *bop,* and metagenome extracted *bop,* and *rpoB*, while 99 and 97% were used as cutoffs for metagenomic 16S rRNA sequences (met-*bop,* met-*rpoB,* and met-16S respectively). Sampling efficiency for each gene at both OTU definitions was determined by rarefying the data. R packages phyloseq [150] and phylogeo [151] were used for further analyses on the OTUs clustered by Mothur. Various alpha diversity indices were calculated within phyloseq including Chao, ACE, Shannon, and Simpson indices. Species richness estimators Chao [152] is nonparametric and bases richness on the number of singletons and doubletons, and ACE [153] is based on the number of rare groups of observed species (S_{obs}). The Shannon index (H) is a commonly used diversity index that takes into account both abundance and evenness of species observed in the community. H usually ranges between 1.5 and 3.5 and the higher value suggests a diverse and evenly distributed community. The closer H is to 0, the less diverse the community is and when H=0, there is a single species represented in the community. Simpson's index of diversity (1-D) ranges between 0 and 1, and describes the sample diversity: the closer D is to 1, the more diverse the sample is, and the closer it is to 0, the less diverse and more even the community is. This index is a measure of the probability that two randomly chosen individuals from the sample belong to different species.

3.2.5 Comparison of communities from distant sites

Representative sequences for each OTU at the different definitions of sequence similarity defined by MOTHUR were assembled and aligned. These were used to extract top BLAST hits from the 109 halobacterial genomes available on GenBank using a tBLASTx search [154]. The top BLAST hits of the representative sequences was used to determine the composition of each community at the genus level. BLAST hits with <50% nucleotide sequence identity were classified as unknown Haloarchaea. Communities were compared qualitatively based on the presence or absence of different genera and their relative abundance in the composition.

The LIBSHUFF command within Mothur was employed to determine statistical similarity between each pairwise comparison of communities. The command within Mothur implemented the original LIBSHUFF program [155]. It tested for similarity in structure between two or more communities by incorporating the Cramer-von Mises test statistic [156] and returning a significance value for the difference between each pair in consideration.

Pairwise community distances were measured employing 44 different methodologies within the phyloseq package and the Jaccard and Canberra distance methods were selected for further community comparisons and visualization of data. The Jaccard index of dissimilarity [157] measures the extent of overlap of species between two different communities by taking into consideration the total number of species present in both communities and the number of species that are exclusive to each community. The closer the value is to 1, the more dissimilar the communities are. The Canberra distance [158], similar to the Manhattan distance measure, determines the sum of a series of fraction of differences between two vectors. Two communities are distant if the value is close to 1. Phylogeo was used to estimate the effect of geographic distance on the genetic divergence between the communities.

3.2.6 Dispersal between sites

The list of OTUs was analyzed to identify clustering of sequences from different sites to estimate the extent of sharing. OTUs with sequences from different sampling sites were collected and the direction of dispersal was estimated using the software package Migrate-n [159]. Migrate-n predicts the immigration and emigration to and from a site using both gene frequencies and sequence data by assuming a migration matrix model with asymmetric migration rates to estimate the maximum likelihood of the occurrence of the event.

3.3 Results

3.3.1 Sampling efficiency and halobacterial diversity

Evidence from previous analysis of cultivated strains from the genus *Halorubrum* demonstrated that clusters with intra-cluster nucleotide similarity ≥99% contained only synonymous changes in protein coding loci, including the bacteriorhodopsin gene [84]. Therefore, we assumed any nonsynonymous nucleotide changes detected in OTUs with ≥99% similarity would be candidates for laboratory-induced mutations. Cutoffs of 99% and 95% sequence similarity were chosen to represent a stringent and liberal estimate, respectively.

Rarefaction curves were generated for PCR-*bop* at 99% and 95% as well as for the metagenome extracted gene sequences (figures 3-2 through 3-5). For PCR-*bop,* sampling for most communities seems to be sufficient since the rarefaction curves at both 99% and 95% either plateau or begin to plateau, which suggests our technique captured a very large fraction of the existing bacteriorhodopsin diversity. The exception to this are the SPS-I, SPS-II, CTS, SS communities that were not as deeply sampled. This, however, is not the case with the metagenomes. All the genes extracted from the metagenomes concur with the need for further sampling. The rarefaction

curves for meta-16S, meta-*rpoB*, and meta-*bop* do not plateau at the more stringent OTU definitions but start to plateau in certain cases with the liberal estimations.

3.3.2 Community comparisons

3.3.2.1 Phylogenetic Distribution Pattern

From each of the five sites sampled in this study, 100 putative bop containing plasmid clones were sequenced. However, not all plasmids inserts were successfully sequenced, leaving a total of 359 new sequences. Inclusion of bacteriorhodopsin sequence data from seven additional studies raised the total number of sites and sequences analyzed to twelve and 973 respectively. (Table 3-1). A ML tree was constructed with these sequences to see if there was site specific clustering within the tree. Figure 3-6 is the ML tree showing site specific clustering as depicted by the red colored collapsed branches. Many sequences cluster in a site specific manner while many others cluster with sequences from other sites.

3.3.2.2 Taxonomic Richness Estimation

Richness estimations Chao1 [152], and ACE [153], as well as the number of observed species (S_{obs}) for each site studied with PCR-*bop*, suggest that salterns are more species rich than hypersaline lakes. Apart from the Secovlje saltern in Slovenia and one study on Santa Pola saltern in Spain (SPS-I), all other salterns studied exhibit greater S_{obs} as well as Chao1 and ACE estimates (Figure 3-7). Both the Shannon and Simpson diversity indices corroborate the results from the species richness estimators. Chao1 estimates for the metagenomes analyzed, however, provide variable states. Estimates for 95% meta-*rpoB* and meta-*bop* shows that Lake Tyrrell is more species rich than the salterns. This is contradictory to the 97% meta-16S where the Chao1 estimate is lower than that for the salterns. Apart from the Chao1, the other indices of diversity and species

richness estimates agree when comparing the three genes used for metagenome analyses. Both methodologies of comparing hypersaline environments suggest that the man-made salterns are more rich and diverse in species than natural lakes.

3.3.2.3 Community fingerprints

Community fingerprints demonstrating the normalized composition as determined from the top BLAST hits from PCR-*bop* for each site were constructed. Figure 3-8 shows that each community is structured differently, even at the genus level, with different genera being present and in varying degrees of relative abundance. *Halorubrum* seems to be the dominant genus in ABL, CSL, CVS, ES, and SS while it is second in dominance in GSL and HS. GSL and HS are both dominated by *Haloarcula*. *Haloquadratum*, on the other hand is found to dominate in fewer sites including BSS, and SPS-I and II. *Halobacterium, Halobiforma, Haloferax, Halomicrobium, Haloplanus, Halorhabdus, Haloterrigena, Natrinema, Natronomonas,* and *Natronorubrum* are the other sparsely represented genera in varying abundances. Unknown members of haloarchaea are sporadically distributed between the sites. Apart from the community compositions, community fingerprints based on presence and absence of OTUs (Figure 3-9) supports the finding that the twelve sites studied using PCR-*bop* have unique community compositions. Similar analyses on the meta-genes from the 6 available metagenomes reveal unique OTU patterns (see figures 3-10, 3-11, and 3-12) further corroborating findings from PCR-*bop*, negating possible methodological biases.

3.3.2.4 LIBSHUFF analyses show statistical dissimilarity in OTUs from different sites.

LIBSHUFF carries out pairwise comparisons to determine if one data set is a subset of the other. Allowing for a 5% false detection rate before applying the Bonferroni correction for multiple

library comparisons, only p values less than 0.0025 are considered statistically significant for inferring that two samples are different. Bonferroni corrections were computed for each dataset and results from LIBSHUFF analyses are presented in table 3-2 for the 12 PCR-*bop* sites, tables 3-3 through 3-5 for the meta-genes compared. Most pairwise comparisons, be it PCR-*bop* or meta-genes, return significant p-values after applying the appropriate Bonferroni corrections suggesting that the communities are distinct from one another.

3.3.3 Genetic distance vs geographic distance

To test the hypotheses that geographically closer sites, or ecologically similar environments (e.g., industrial salterns vs. natural salt lakes) might produce comparable communities, we estimated the genetic distances among the twelve sites using the Jaccard and Canberra distances measured for PCR-*bop*. By and large evidence supporting either hypothesis is scarce: neither geographically closer sites nor ecologically similar ones appear to have similar communities (see figure 3-13). There is no distinct clustering of sites based on ecological conditions. Indeed, in figure 3-13, each community is about as dissimilar as possible, with the most similar two sites being Chula Vista, USA and Huelva, Spain. Comparable results were obtained from analyses on the metagenomic data (figure 3-14). We might not expect that natural salt lakes should contain similar communities with each other, or with salterns, as their environmental conditions dictated by ionic composition alone are often wildly different [160]. This is due to their salts being derived from their surroundings, which are composed of highly variable inorganic material. However, we might expect salterns to be more similar in community composition, because their ionic composition is derived from the same starting point; seawater. Our analyses, however, indicate no such bias in similarity. The Great Salt Lake community

resembles more the ones from the Eilat crystallizer and the Aran-Bidgol Lake, rather than the ones from other lakes or from the Chula Vista saltern in USA. Another example is the saltern community from Huelva, Spain that is more like the one in Chula Vista, USA rather than the ones in Santa Pola or Chiprana in Spain.

3.3.4 Dispersal between sites

Even though the communities are dissimilar and have unique compositions, they do share many genera and possibly species, indicating strong evidence for dispersal between them. Further, the sum of OTUs defined for individual sample sites is greater than the total number of OTUs for all twelve sites collectively (see Table 3-6), indicating OTUs are shared between sites. Our analysis demonstrates that in some cases sequences with 100% similarity were shared between sites. As expected, as the OTU definition is relaxed, the number of OTUs shared between sites increases. In order to discover patterns of dispersal between sites we analyzed the 95% PCR-*bop* OTUs using the software package Migrate-n [159], and summarize the results in figure 3-15. The arrows depict the direction of dispersal and the width of each sector represents the fraction of the likelihood estimate for each site. There is evidence for widespread dispersal events in the past irrespective of the geographic distance or the ecological type. CVS and HS have the largest dispersal events the past, with more OTUs leaving CVS than coming in, and vice versa for HS. There are 40 OTUs out of a total of 271 (~15%) OTUs defined at 99% that are shared between at least two sites, and 39 that were shared at 95% (out of 162 OTUs, ~24%). At 99%, 1 OTU is shared between 5 sites; 1 shared OTU between 4 sites; 9 OTUs shared between three sites; 29 OTUs shared between two sites. At 95%, there is 1 OTU shared between 6 sites; 2 OTUs shared between 5 sites; 3 OTUs shared between 4; 10 OTUs shared between 3 sites; 23 OTUs shared between 2 sites. These observations suggest that some organisms (the vast minority) may be particularly well adapted to

surviving dispersal and invading habitats. CVS and HS shared the most OTUs with 13 at 99% and 95% cutoff values, which explains why they resembled each other in our community similarity analyses, but does not offer an explanation for why they are shared: saltern crystallizer ponds located in SPS-I and II, SS, and ES are all much closer. There did not seem to be any patterns in terms of direction of dispersal, however CVS and ES, appeared to share more OTUs with more locations than any other single sample site.

Endemism is apparent as though there are shared OTUs between the sites, a total of 123 out of the 162 OTUs observed at 95% sequence similarity still remain unique to the respective sites (see table 3-7). Not all sites displayed evidence for dispersal between them as only half of the analyzed samples shared an OTU and there is a mixed pattern of dispersal to and from the six sites. CVS near San Diego in the United States of America and the crystallizer saltern in Huelva half way across the world in Spain share the most OTUs suggesting great dispersal between the two. Also, there is sharing between the naturally occurring salt lake, GSL, and the crystallizer pond ES, Israel. With apparent differences in the type of habitat, dispersal seems to be occurring on the global scale refuting the notion that dispersal is greater between two close niches especially since no shared OTUs between HS and SPS-I and II or CSL and SPS-I and II were recovered. However the extent of dispersal does not imply that the Halobacteriales are all the same in the hypersaline waters.

The metagenomes depict a similar scenario where endemism is true for haloarchaeal communities (see tables 3-8 and 3-9). No OTU is shared between all 6 sites analyzed. At 97%, the meta-16S OTUs show sharing only between 3 out of the 6 metagenomes analyzed. There are 3 OTUs shared among 3 sites, 2 of which are shared among the 21%, 33%, and 37% salinity salterns in Spain. The other OTU is shared among the 21%, 33%, Spain, and the saltern in the Cahuil

lagoon in Chile. There are 41 OTUs shared between two sites – 9 shared between 21% Spain and Chile; 8 shared between 21% and 33% salterns in Spain; 8 shared between 33% and 37% salterns in Spain; 6 between 21% Spain and CVS; 5 between 21% and 37% Spain; 3 between 33% Spain and CVS; 1 between 33% Spain and Chile; and 1 between 33% Spain and Lake Tyrrell in Australia. These 44 shared OTUs represent ~1.38% of the total OTUs obtained combining the 6 metagenomes. Interestingly, there are no shared OTUs at 99% meta-16S. Similarly with the 95% meta-*rpoB*, sharing is observed only between 3 of the 6 sites and a total of 66 (~3.17%) OTUs are shared between at least 2 sites – 7 shared among 3 sites; and 59 among 2 sites. Just like with the meta-16S, no shared OTUs are observed with the 99% meta-*rpoB*. The findings from the 95% meta-*bop* further confirms the results from meta-16S and meta-*rpoB*, most number of sites that share an OTU is 3. There are 33 (~4.01%) OTUs shared at least between 2 sites – 3 among 3 sites and 30 among 2 sites. All three genes used as markers to compare the six metagenomes corroborate the findings from PCR-*bop*, there is sharing of haloarchaea between sites but each site promoted endemism.

3.4 Discussion

Biogeographic patterns of Archaea can be studied on hypersaline environments such as the Dead Sea or the Great Salt Lake as these are geographically distant and impose very specific restrictions upon organisms that thrive within them. Compared to soil or other complex niches, these extreme environments house fewer species and can be studied relatively easily. Earlier diversity analyses on these environments have primarily been driven by the 16S rRNA [14, 63, 64, 67, 68, 76, 79, 161-163]. Here we used PCR amplification of the *bop* gene as previously used [13, 135, 136] as well as metagenomics comparisons of multiple geographically distant hypersaline environments to determine biogeographic patterning and the forces driving it. Frequent dispersal

events can act as a homogenizing force while adaptive mutations, community resistance to invasion, and differences in the ability to disperse can lead to endemism at different locations. Our analyses show that the bacteriorhodopsin gene shows results similar to the 16S rRNA and *rpoB* genes, and it gives better species richness estimation to observed species ratio (Figure 3-7b) as well as recovers similar dominant genera as some earlier studies on these environments using 16S rRNA [16, 79, 164]. Though some of the rarefaction curves do not plateau, most of the curves suggest sufficient sampling (Figure 3-2) and can therefore justify our finding that each community analyzed has a unique fingerprint (figures 3-8 and 3-9). Each of the twelve sites assayed using PCR-*bop* as well as the six metagenomes (figures 3-10 through 3-12) displayed the existence of unique communities, both based on community composition at the genus level as well as the presence/absence of OTUs. This suggests that the biogeographic patterning resulting in the global distribution of the haloarchaea is similar to earlier findings in other systems [8-25].

Comparing the species richness of the salterns to the lakes is particularly interesting. An earlier study described the difference in species richness between natural hypersaline lakes and man-made crystallizer ponds [16] and determined that hypersaline lakes were more species rich than the salterns. This, however, is not what is observed here. Every measure of species richness and diversity estimated for the PCR-*bop* as well as for meta-16S, meta-*rpoB*, and meta-*bop* suggests that the salt crystallizers are more species rich and diverse than natural lakes (figures 3-7a and b). This could be attributed to the relative stability of the lakes in comparison to salterns with respect to disturbance, the salterns are disturbed more frequently and derives greater species diversity. This must be the reason since though salterns are derived from sea water in contrast to lakes, both show stable communities maintained though time. In fact, a recent study on the saltern in Eilat, Israel showed the maintenance of diverse community through time with variations in

relative abundances of its members (described in Chapter 4) similar to Lake Tyrrell, Australia where there were seasonal fluctuations in the abundance of the members [59]. Another possible explanation would be that the hypersaline lakes are older in comparison to the salterns. With age, the hypersaline lake niche could have been saturated with the inhabitants making it difficult for newer species to successfully invade and survive here.

If ecological differentiation affected the haloarchaeal communities, we would expect to see saltern communities that are more similar to one another and distant from the lake communities. Interestingly, this expectation is not met. There is no clustering of sites in the MDS plots based on whether the sites were salterns or lakes as seen in figures 3-13 (PCR-*bop* on 12 sites) and 3-14 (meta-genes on 6 sites) using the Jaccard and Canberra distances. Figures 3-13 and 3-14 also show a lack of clustering on the NMDS plots based on geographic distance. Estimating the increase in genetic distance with increase in geographic distance, there are no real patterns observed. Unlike with eukaryotes, the prokaryotic genus *Roseiflexus* [21], and the archaeal and bacterial inhabitants of the Pantanal sediment [25], with the haloarchaea the distance between the sites doesn't seem to play a role in diverging them. In fact, the communities are all as different as can be except for CVS and HS communities, and GSL and ES communities that have Jaccard distances of less the 0.85 at 95%. CVS in USA and HS, half way across the world, in Spain are less distant than HS and the other sites in Spain. GSL, a naturally hypersaline lake is more similar to ES, a saltern in Eilat than the other lakes studied. The metagenomes second the findings from PCR-*bop*, however, while meta-16S shows that each community is completely different from the other, meta-*bop* and meta-*rpoB* identify less distant pairs of communities (Jaccard index is still greater than 0.9) and not completely dissimilar ones. This leads us to believe that neither the geographic isolation nor the ecological conditions with respect to whether saltern or lake play a role in overall communities.

Identification of distinct haloarchaeal communities, as also seen in earlier studies [12, 13, 66, 136, 165-167], raises the question whether geographical separation acts as a barrier to dispersal of haloarchaea, preventing the homogenization of the communities. This does seem to be an important factor, though geographic isolation does not completely inhibit dispersal, dispersal in limited by the separation. There is evidence for dispersal between sites as many sites share OTUs at 99% and 95% PCR-*bop* and yet most OTUs are unique. The discrepancy in the sum of all the individual OTUs and the number of OTUs with all 12 sites collectively (Table 2) was the first evidence for shared OTUs among sites. Though there is not one OTU shared by all 12 sites studied with PCR-*bop* or the 6 metagenomes, sharing between fewer sites is widespread. Estimates on the likelihood of dispersal between sites also suggests the haloarchaea are swept far and wide. The network of dispersal events in figure 3-6 at 95% PCR-*bop* shows that irrespective of geographic separation or ecological type, immigration and emigration to and from a site occurs. Though there is not a lot known about how the haloarchaea are dispersed, the identification of *Halococcus spp.* in the nostrils salt gland of *Calonectris diomedea* [168] illustrates a possible mode of dispersal from one site to the next aside from simply being blown about by the wind. Though geographic separation does not act as a barrier to any dispersal, the fraction of shared OTUs is much lower than the ones that are unique. The shared OTUs defined at 95% PCR-*bop* only account for ~24% of the overall OTUs, 75% are endemic (Table 3-7). With the metagenomes, this fraction of shared OTUs is even smaller (~1.38% meta-16S (see table 3-8), ~3.17% meta-*rpoB*, and ~4.01% meta-*bop* (see table 3-9)) suggesting that the haloarchaeal communities are still mostly endemic. Similar to most of the eukaryotes, the haloarchaea do in fact maintain unique communities at each hypersaline environment.

Table 3-1: Total number of sequences obtained from each sampling site -

Site	*Country*	*Abbreviation*	*Number of sequences*	*Source*
Aran-Bidgol Lake	Iran	ABL	58	This study
Guerrero Negro Saltern	Mexico	BSS	161	[13]
Chinese Salt Lakes	China	CHL	19	unpublished
Chiprana Salt Lake	Spain	CSL	74	This study
Chiku Saltern	Taiwan	CTS	23	Lin et al., unpublished
Chula Vista Saltern	USA	CVS	95	This study
Eilat Saltern	Israel	ES	349	Ram-Mohan et al., 2016
Great Salt Lake	USA	GSL	72	This study
Huelva Saltern	Spain	HS	60	This study
Santa Pola Saltern	Spain	SPS-I	23	[135]
Santa Pola Saltern	Spain	SPS-II	29	[143]
Secovlje Saltern	Slovenia	SS	10	[136]
Total			973	

Table 3-2. LIBSHUFF pairwise comparison of the 12 sampling sites assayed with the PCR-*bop*. For 12 sites, applying the Bonferroni correction, p-values <0.0004 are statistically significant.

COMPARISON	DCXYSCORE	SIGNIFICANCE
ABL-SPS-II	0.126529	0.0001
SPS-II-ABL	0.080003	0.0001
ABL-BSS	0.139154	0.0001
BSS-ABL	0.263005	0.0001
ABL-CTS	0.196386	0.0001
CTS-ABL	0.099908	0.0001
ABL-CHL	0.123929	0.0001
CHL-ABL	0.151595	0.0001
ABL-CSL	0.039841	0.0001
CSL-ABL	0.106122	0.0001
ABL-CVS	0.021019	0.0001
CVS-ABL	0.03255	0.0001
ABL-ES	0.054293	0.0001
ES-ABL	0.120627	0.0001
ABL-GSL	0.051071	0.0001
GSL-ABL	0.00396	0.0358
ABL-HS	0.074221	0.0001
HS-ABL	0.114317	0.0001
ABL-SPS-I	0.22278	0.0001
SPS-I-ABL	0.244221	0.0001
ABL-SS	0.168294	0.0001
SS-ABL	0.077117	0.1137
SPS-II-BSS	0.001126	0.885
BSS-SPS-II	0.050318	0.0001
SPS-II-CTS	0.142427	0.0001
CTS-SPS-II	0.13463	0.0001
SPS-II-CHL	0.160591	0.0001
CHL-SPS-II	0.25011	0.0001
SPS-II-CSL	0.13631	0.0001
CSL-SPS-II	0.153366	0.0001
SPS-II-CVS	0.055476	0.0001
CVS-SPS-II	0.129656	0.0001
SPS-II-ES	0.025493	0.0258
ES-SPS-II	0.123393	0.0001
SPS-II-GSL	0.070015	0.0001
GSL-SPS-II	0.037124	0.0001
SPS-II-HS	0.103137	0.0001
HS-SPS-II	0.182865	0.0001

SPS-II-SPS-I	0.064163	0.0001
SPS-I-SPS-II	0.031762	0.0005
SPS-II-SS	0.091163	0.007
SS-SPS-II	0.047178	0.3202
BSS-CTS	0.301047	0.0001
CTS-BSS	0.118665	0.0001
BSS-CHL	0.313974	0.0001
CHL-BSS	0.229303	0.0001
BSS-CSL	0.316357	0.0001
CSL-BSS	0.141589	0.0001
BSS-CVS	0.201825	0.0001
CVS-BSS	0.128556	0.0001
BSS-ES	0.035979	0.0001
ES-BSS	0.056973	0.0001
BSS-GSL	0.245515	0.0001
GSL-BSS	0.042072	0.0001
BSS-HS	0.228557	0.0001
HS-BSS	0.11666	0.0001
BSS-SPS-I	0.10453	0.0001
SPS-I-BSS	0.013833	0.0028
BSS-SS	0.243747	0.0001
SS-BSS	0.042377	0.1211
CTS-CHL	0.065702	0.0001
CHL-CTS	0.061328	0.0001
CTS-CSL	0.064034	0.0001
CSL-CTS	0.283996	0.0001
CTS-CVS	0.036178	0.0621
CVS-CTS	0.201888	0.0001
CTS-ES	0.03788	0.0353
ES-CTS	0.269303	0.0001
CTS-GSL	0.049056	0.0001
GSL-CTS	0.26489	0.0001
CTS-HS	0.047718	0.0001
HS-CTS	0.203452	0.0001
CTS-SPS-I	0.107863	0.0001
SPS-I-CTS	0.159145	0.0001
CTS-SS	0.09744	0.0024
SS-CTS	0.104026	0.0086
CHL-CSL	0.184321	0.0001
CSL-CHL	0.144166	0.0001
CHL-CVS	0.154555	0.0001
CVS-CHL	0.154038	0.0001
CHL-ES	0.18329	0.0001

ES-CHL	0.20696	0.0001
CHL-GSL	0.197376	0.0001
GSL-CHL	0.197661	0.0001
CHL-HS	0.162868	0.0001
HS-CHL	0.179869	0.0001
CHL-SPS-I	0.28144	0.0001
SPS-I-CHL	0.306418	0.0001
CHL-SS	0.32031	0.0001
SS-CHL	0.100052	0.0001
CSL-CVS	0.04241	0.0001
CVS-CSL	0.034676	0.0001
CSL-ES	0.033685	0.0001
ES-CSL	0.147686	0.0001
CSL-GSL	0.109339	0.0001
GSL-CSL	0.135089	0.0001
CSL-HS	0.021562	0.0001
HS-CSL	0.062534	0.0001
CSL-SPS-I	0.313925	0.0001
SPS-I-CSL	0.291671	0.0001
CSL-SS	0.147567	0.0001
SS-CSL	0.042354	0.2952
CVS-ES	0.06005	0.0001
ES-CVS	0.02981	0.0001
CVS-GSL	0.100578	0.0001
GSL-CVS	0.009879	0.0001
CVS-HS	0.01443	0.0007
HS-CVS	0.001942	0.3223
CVS-SPS-I	0.219136	0.0001
SPS-I-CVS	0.128809	0.0001
CVS-SS	0.15593	0.0001
SS-CVS	0.031455	0.7156
ES-GSL	0.092233	0.0001
GSL-ES	0.011109	0.0001
ES-HS	0.157989	0.0001
HS-ES	0.082756	0.0001
ES-SPS-I	0.25967	0.0001
SPS-I-ES	0.095608	0.0001
ES-SS	0.204577	0.0001
SS-ES	0.037976	0.6455
GSL-HS	0.156068	0.0001
HS-GSL	0.127554	0.0001
GSL-SPS-I	0.264302	0.0001
SPS-I-GSL	0.222573	0.0001

GSL-SS	0.190568	0.0001
SS-GSL	0.068719	0.001
HS-SPS-I	0.220926	0.0001
SPS-I-HS	0.199501	0.0001
HS-SS	0.161018	0.0001
SS-HS	0.04155	0.5715
SPS-I-SS	0.111388	0.0019
SS-SPS-I	0.081083	0.0001

Table 3-3. LIBSHUFF pairwise comparison of the 6 metagenomes assayed with the meta-16S. For 6 samples, applying the Bonferroni correction, p-values <0.0016 are statistically significant.

COMPARISON	DCXYSCORE	SIGNIFICANCE
21S-33S	0.01964	<0.0001
33S-21S	0.02713	<0.0001
21S-37S	0.075957	<0.0001
37S-21S	0.021056	<0.0001
21S-CV	0.06564	<0.0001
CV-21S	0.060397	<0.0001
21S-CHILE	0.019661	<0.0001
CHILE-21S	0.000508	1
21S-TYRRELL	0.210566	0.004
TYRRELL-21S	0.042704	1
33S-37S	0.040874	<0.0001
37S-33S	0.002478	0.0445
33S-CV	0.071858	<0.0001
CV-33S	0.042749	<0.0001
33S-CHILE	0.07934	<0.0001
CHILE-33S	0.005586	0.9752
33S-TYRRELL	0.245779	0.0028
TYRRELL-33S	0.046474	0.9999
37S-CV	0.067811	<0.0001
CV-37S	0.093066	<0.0001
37S-CHILE	0.101102	<0.0001
CHILE-37S	0.040537	<0.0001
37S-TYRRELL	0.210865	0.0001
TYRRELL-37S	0.034647	0.9972
CV-CHILE	0.103455	<0.0001
CHILE-CV	0.025998	<0.0001
CV-TYRRELL	0.248725	<0.0001
TYRRELL-CV	0.026785	0.9977
CHILE-TYRRELL	0.156288	0.0084
TYRRELL-CHILE	0.031329	0.996

Table 3-4. LIBSHUFF pairwise comparison of the 6 metagenomes assayed with the meta-*bop*. For 6 samples, applying the Bonferroni correction, p-values <0.0016 are statistically significant.

COMPARISON	DCXYSCORE	SIGNIFICANCE
21S-33S	0.069827	<0.0001
33S-21S	0.062926	<0.0001
21S-37S	0.150437	<0.0001
37S-21S	0.08928	<0.0001
21S-CV	0.182829	<0.0001
CV-21S	0.02035	0.997
21S-CHILE	0.044833	0.0002
CHILE-21S	0.003256	0.9995
21S-TYRRELL	0.304989	<0.0001
TYRRELL-21S	0.002106	1
33S-37S	0.032748	<0.0001
37S-33S	0.003845	<0.0001
33S-CV	0.288904	<0.0001
CV-33S	0.053048	0.9046
33S-CHILE	0.227129	<0.0001
CHILE-33S	0.042161	0.0036
33S-TYRRELL	0.311317	<0.0001
TYRRELL-33S	0.009251	0.9993
37S-CV	0.298574	<0.0001
CV-37S	0.057086	0.2339
37S-CHILE	0.350845	<0.0001
CHILE-37S	0.087111	<0.0001
37S-TYRRELL	0.322567	<0.0001
TYRRELL-37S	0.009493	0.9795
CV-CHILE	0.040765	0.2852
CHILE-CV	0.124988	<0.0001
CV-TYRRELL	0.122777	<0.0001
TYRRELL-CV	0.024478	0.0197
CHILE-TYRRELL	0.18491	<0.0001
TYRRELL-CHILE	0.010514	0.4607

Table 3-5. LIBSHUFF pairwise comparison of the 6 metagenomes assayed with the meta-*rpoB*. For 6 samples, applying the Bonferroni correction, p-values <0.0016 are statistically significant.

COMPARISON	DCXYSCORE	SIGNIFICANCE
21S-33S	0.045737	<0.0001
33S-21S	0.04145	<0.0001
21S-37S	0.067058	<0.0001
37S-21S	0.036447	<0.0001
21S-CV	0.076615	<0.0001
CV-21S	0.042992	<0.0001
21S-CHILE	0.027024	<0.0001
CHILE-21S	0.003542	0.994
21S-TYRRELL	0.279563	<0.0001
TYRRELL-21S	0.000649	1
33S-37S	0.02492	<0.0001
37S-33S	0.006458	<0.0001
33S-CV	0.146475	<0.0001
CV-33S	0.09665	<0.0001
33S-CHILE	0.162688	<0.0001
CHILE-33S	0.03705	<0.0001
33S-TYRRELL	0.363345	<0.0001
TYRRELL-33S	0.006899	0.9988
37S-CV	0.175443	<0.0001
CV-37S	0.103398	<0.0001
37S-CHILE	0.178567	<0.0001
CHILE-37S	0.045224	<0.0001
37S-TYRRELL	0.339474	<0.0001
TYRRELL-37S	0.006515	0.9829
CV-CHILE	0.101383	<0.0001
CHILE-CV	0.066288	<0.0001
CV-TYRRELL	0.32023	<0.0001
TYRRELL-CV	0.030446	<0.0001
CHILE-TYRRELL	0.152888	<0.0001
TYRRELL-CHILE	0.002568	0.9965

Table 3-6: Comparison between sum of OTUs from individual sampling sites and collective testing.

	99%	**97%**	**95%**
Sum of OTUs from individual sampling sites	325	245	227
OTUs from nine sites collectively	271	203	162

Table 3-7: OTUs – total and unique at 95%

Site	No. of OTUs at 95%	Unique OTUs at 95%
Baja-Sur	20	9
Eilat	39	21
Great Salt Lake	13	5
Aran-Bidgol Lake	20	14
Chiprana salt Lake	20	12
Chiku	20	18
Chula Vista	37	16
Huelva	20	3
Chinese Salt Lakes	7	6
Secovlje	7	5
Santa Pola-I	6	2
Santa Pola -II	18	12
Total	227	123
Total no. of OTUs between all 9 sites	162	

Table 3-8: OTUs – total and unique at 97% meta-16S

Site	No. of OTUs at 97%	Unique OTUs at 97%
21s	1350	1319
33s	739	715
37s	552	537
Chile	201	192
CV	326	315
Tyrrell	22	21
Total	3190	3099

Table 3-9: OTUs – total and unique at 95% meta-*bop*

Site	No. of OTUs at 95%	Unique OTUs at 95%
21s	282	270
33s	195	169
37s	213	187
Chile	26	25
CV	76	72
Tyrrell	43	42
Total	835	765

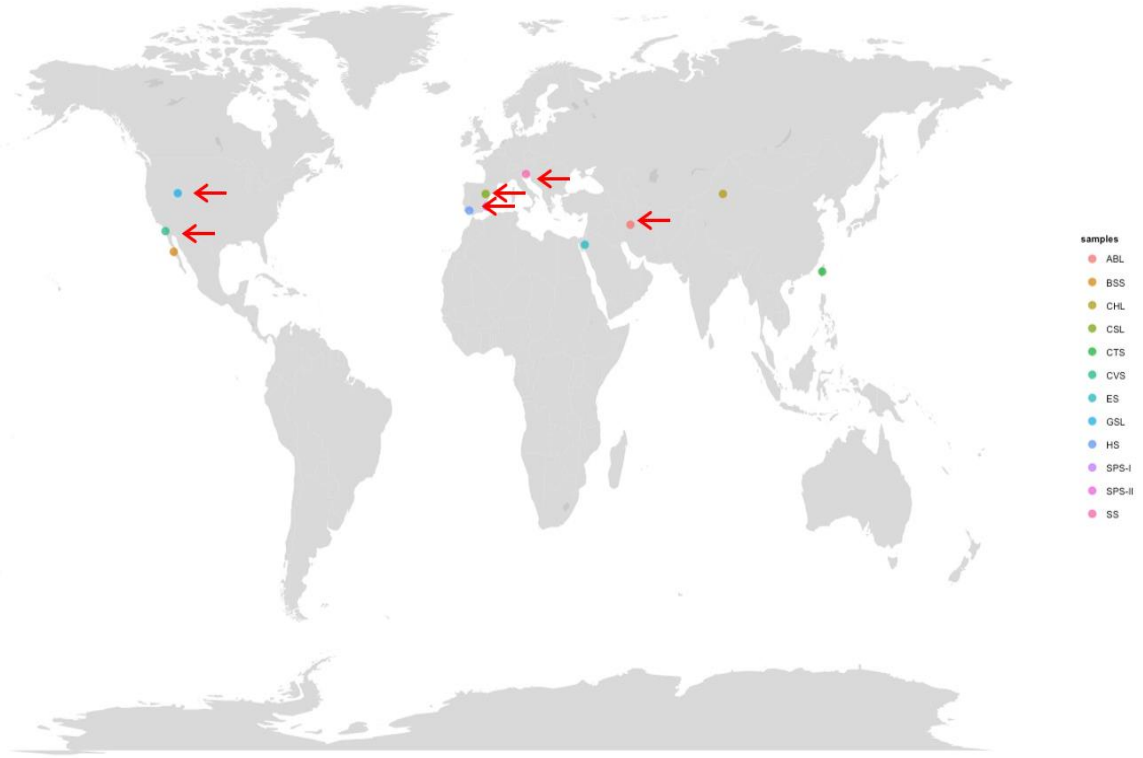


Figure 3-1 Sampling sites from where PCR-*bop* was amplified. Red arrows point to sites sampled in this study.

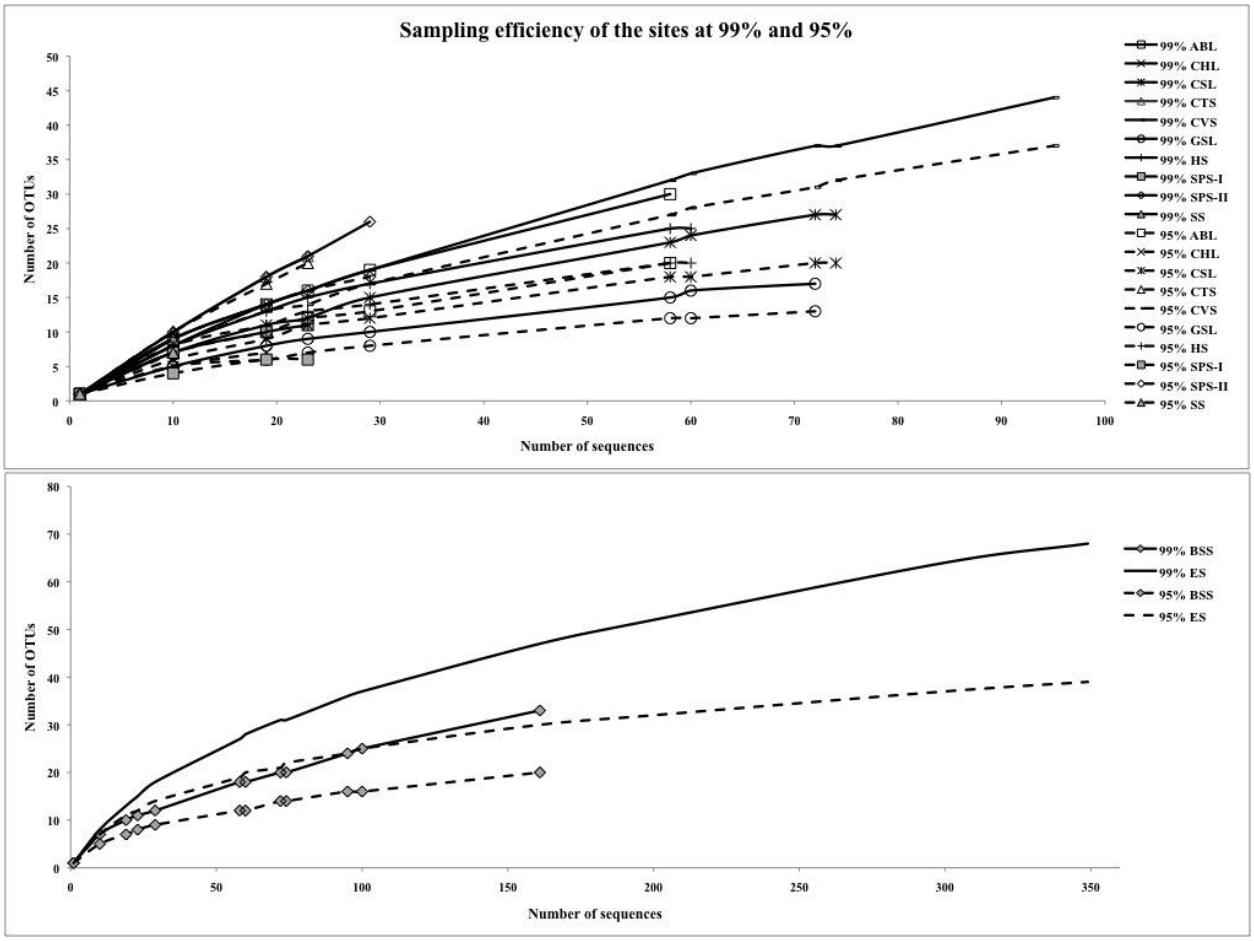


Figure 3-2. Rarefaction curves estimating the sampling efficiency. **Top**. Curves for amplified PCR-*bop* at 99% and 95% sequence similarity for 10 sites. **Bottom**. Curves for BSS and ES at 99% and 95%.

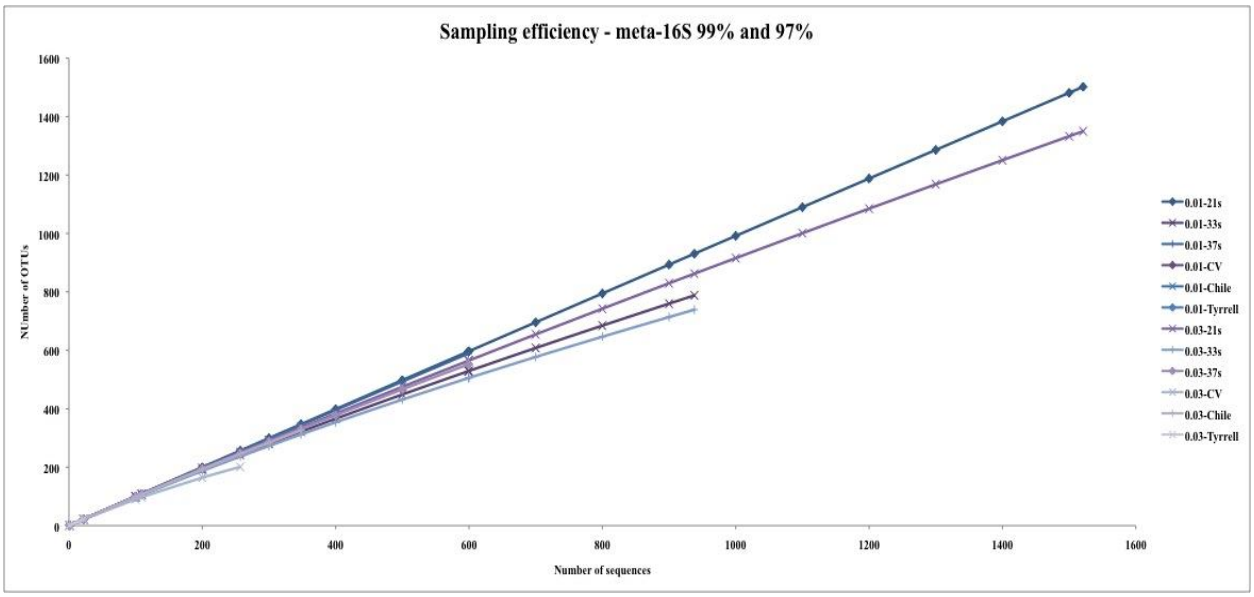


Figure 3-3. Sampling efficiency of the metagenomes analyzed as determined by the meta-16S at 99% and 97%.

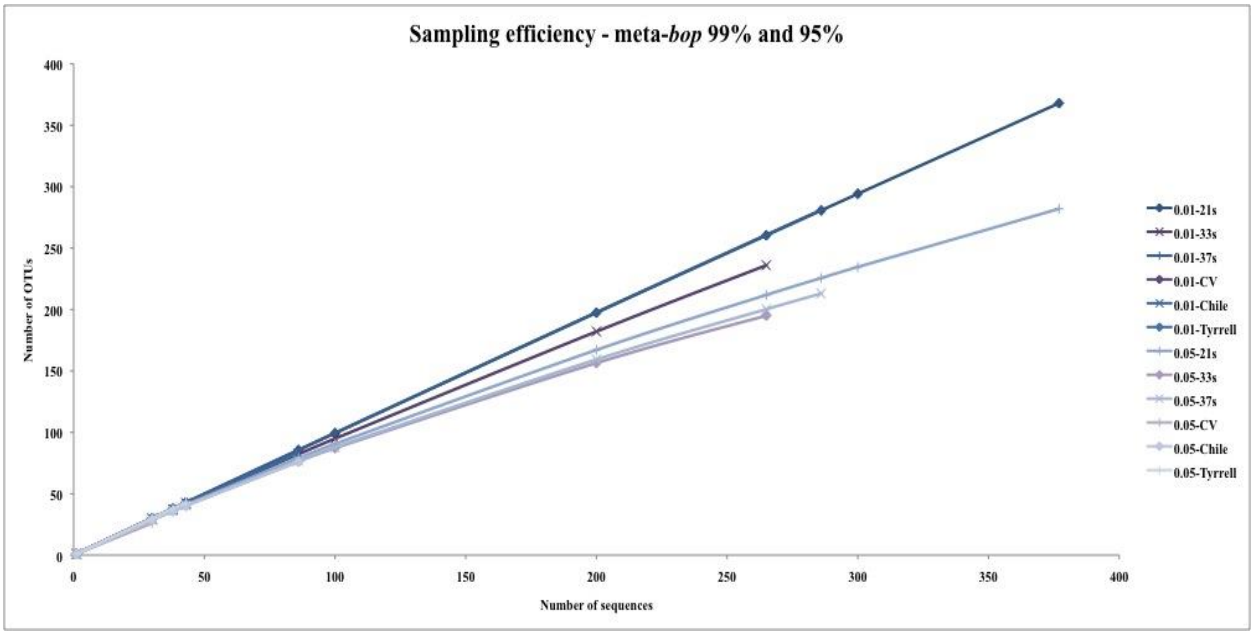


Figure 3-4. Sampling efficiency of the metagenomes analyzed as determined by the meta-*bop* at 99% and 95%.

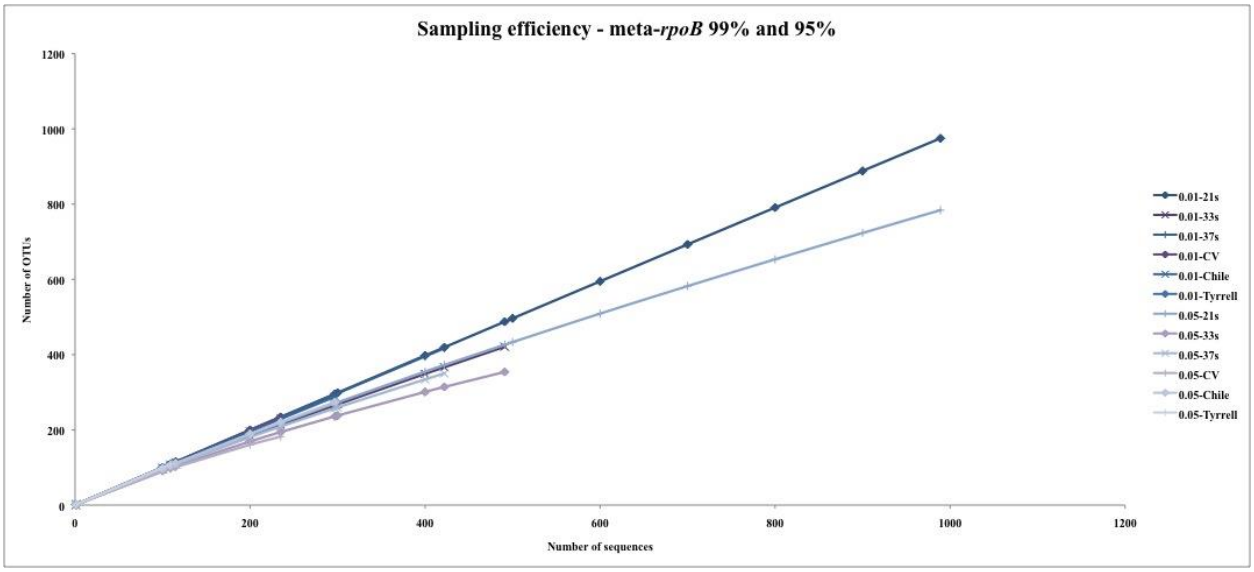


Figure 3-5. Sampling efficiency of the metagenomes analyzed as determined by the meta-*rpoB* at 99% and 95%.

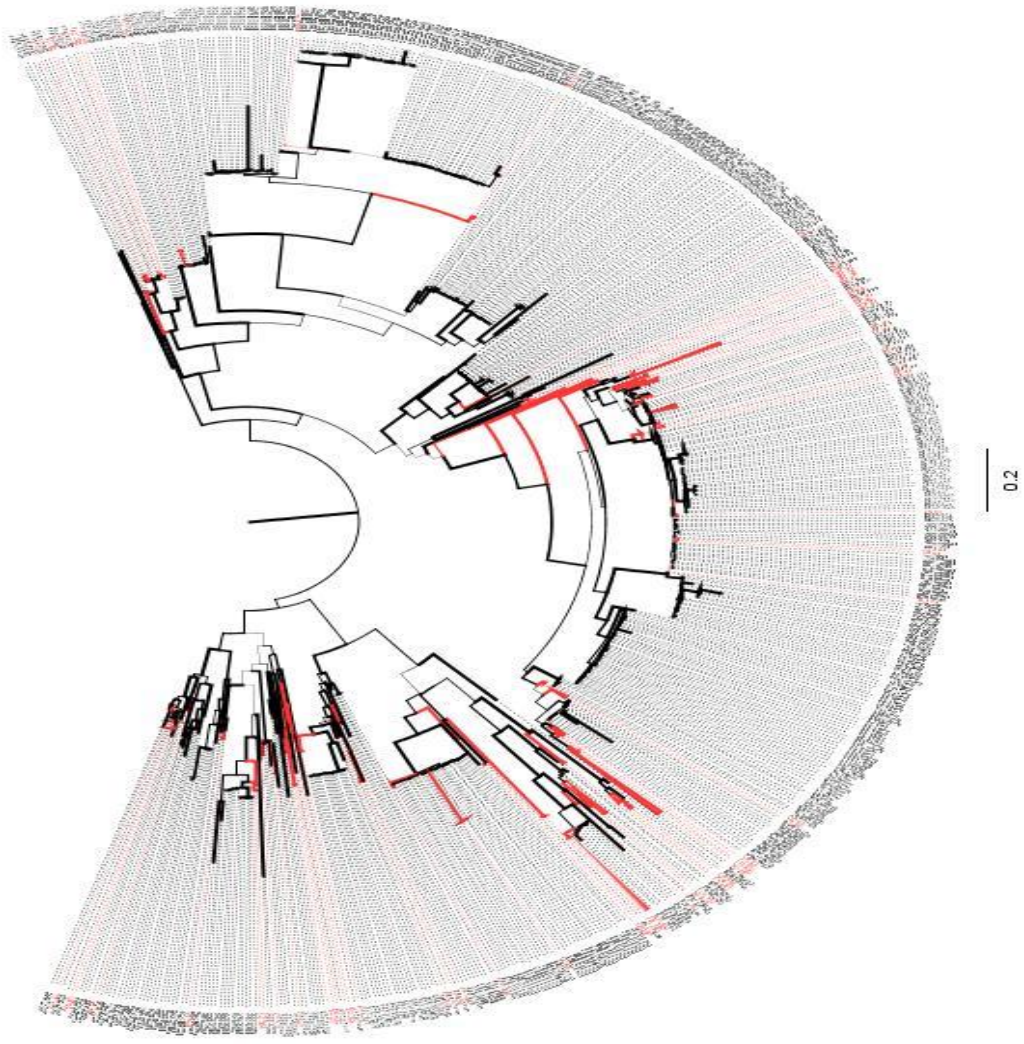


Figure 3-6. ML tree of 973 PCR-*bop* sequences. Clades of endemic sequences were collapsed and colored red. Indication of a combination of distribution and endemism.

Figure 3-7. Observed species, species richness, and diversity estimators. Circles represent natural hypersaline lakes and triangles represent salterns. A) PCR-bop on 12 sites. B) From the 6 metagenomes.

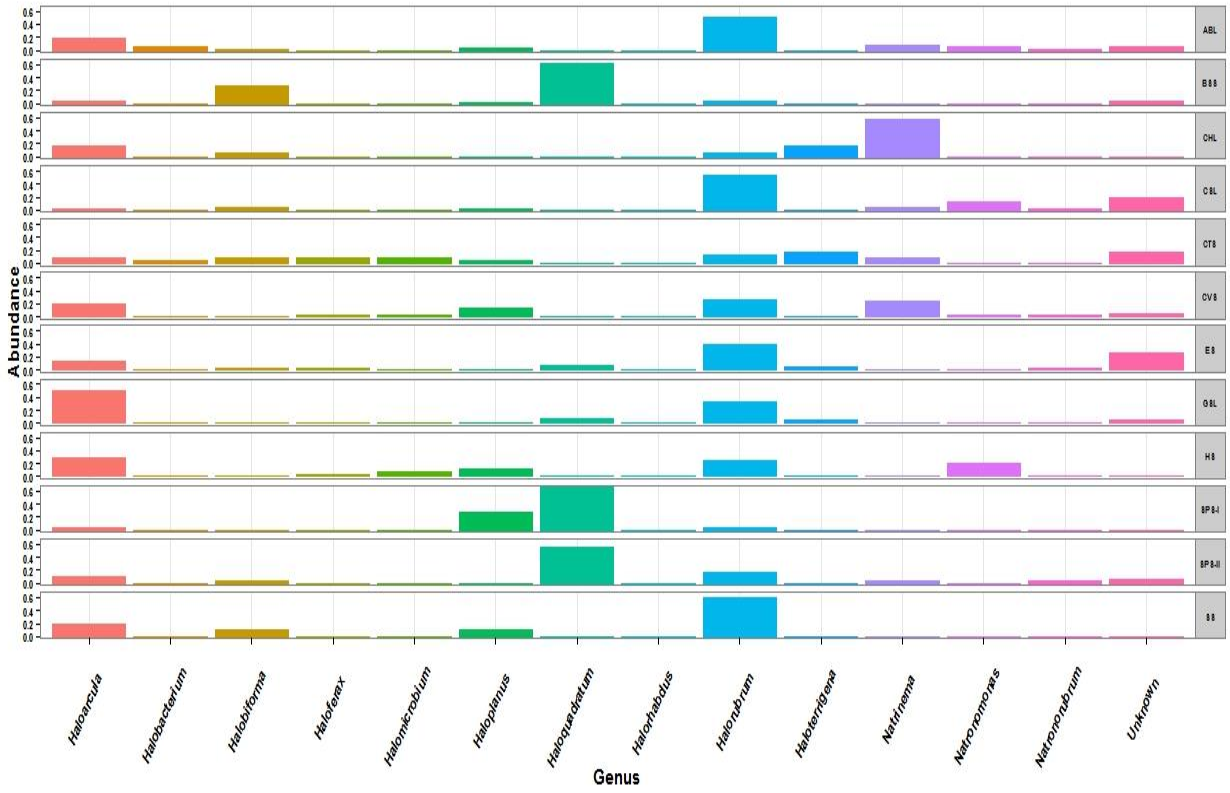


Figure 3-8 Community fingerprint. Composition as determined from the top BLAST hits derived from the 109 available Haloarchaeal genomes. <50% sequence identity is classified as unknown.

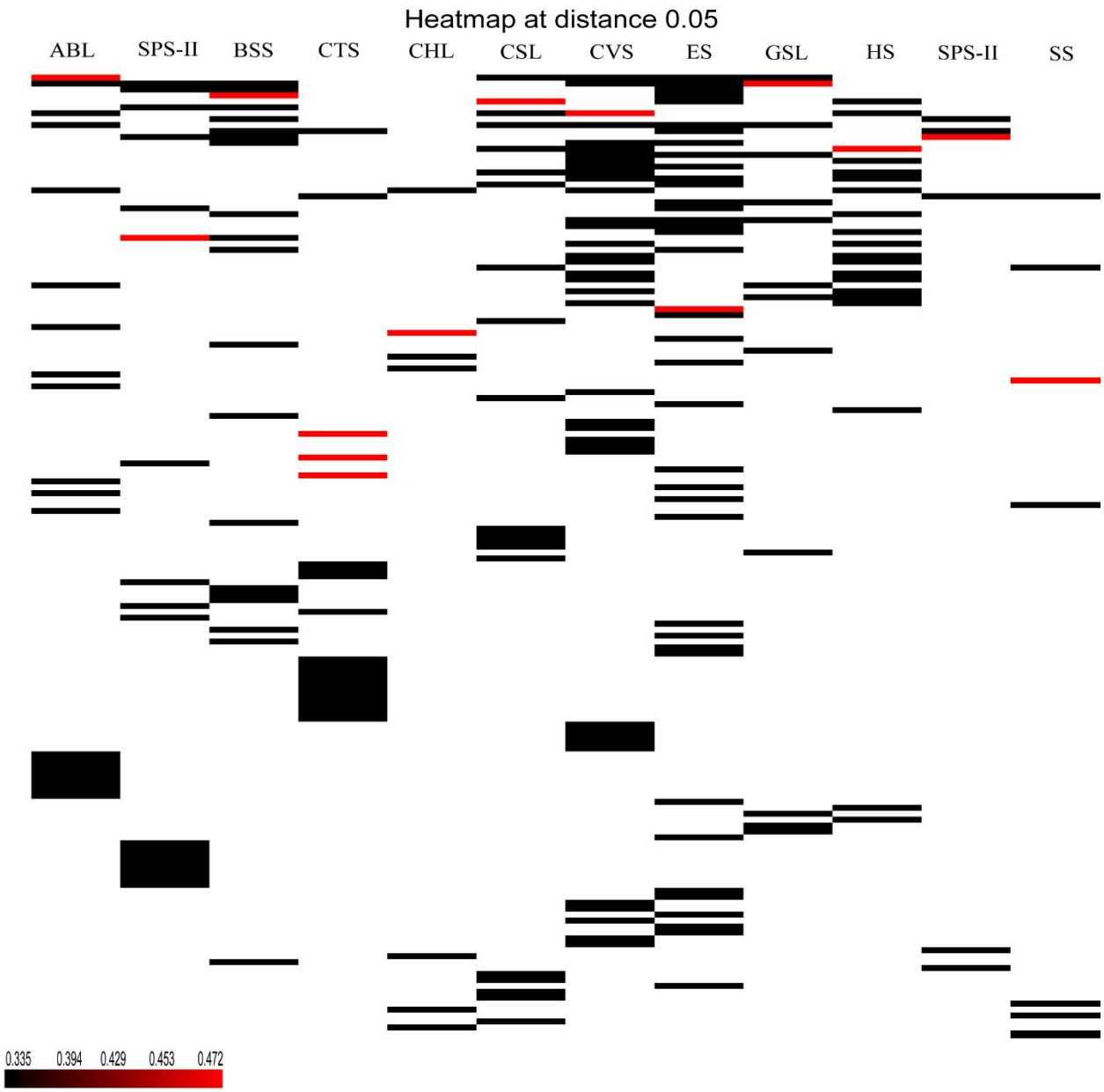


Figure 3-9. Heatmap of the presence/absence of PCR-*bop* OTUs at 95% sequence similarity. Each column represents a sampling site and every row in an OTU. OTUs are colored based on relative abundances.

Figure 3-10. Heatmap of the presence/absence of meta-16S OTUs at 97% sequence similarity in the analyzed metagenomes. Each column represents sampling site, from left – OTU number, 21s, 33s, 37s, Chile, CV, and Tyrrell. Owing to the large number of OTUs, the finer details are missed in the figure but overall, the presence/absence of OTUs at each site is different.

Figure 3-11. Heatmap of the presence/absence of meta-*bop* OTUs at 95% sequence similarity in the analyzed metagenomes. Each column represents sampling site, from left – OTU number, 21s, 33s, 37s, Chile, CV, and Tyrrell.

Figure 3-12. Heatmap of the presence/absence of meta-*rpoB* OTUs at 95% sequence similarity in the analyzed metagenomes. Each column represents sampling site, from left – OTU number, 21s, 33s, 37s, Chile, CV, and Tyrrell.

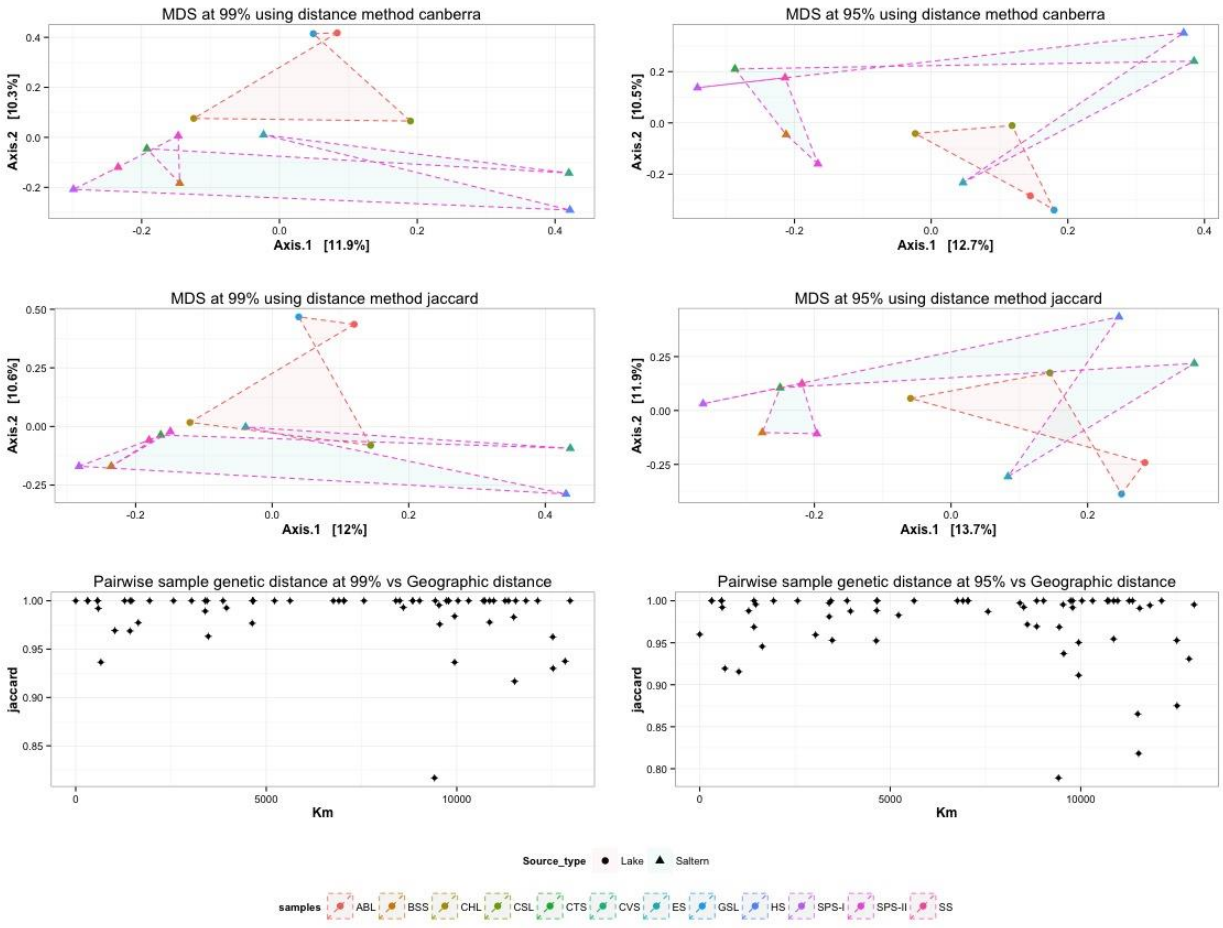


Figure 3-13. Clustering of sampling sites based on geographic proximity and ecological similarity. Distance-decay relationships. Canberra and Jaccard distance calculated between the communities at both 99% and 95%. For the 12 PCR-bop sites.

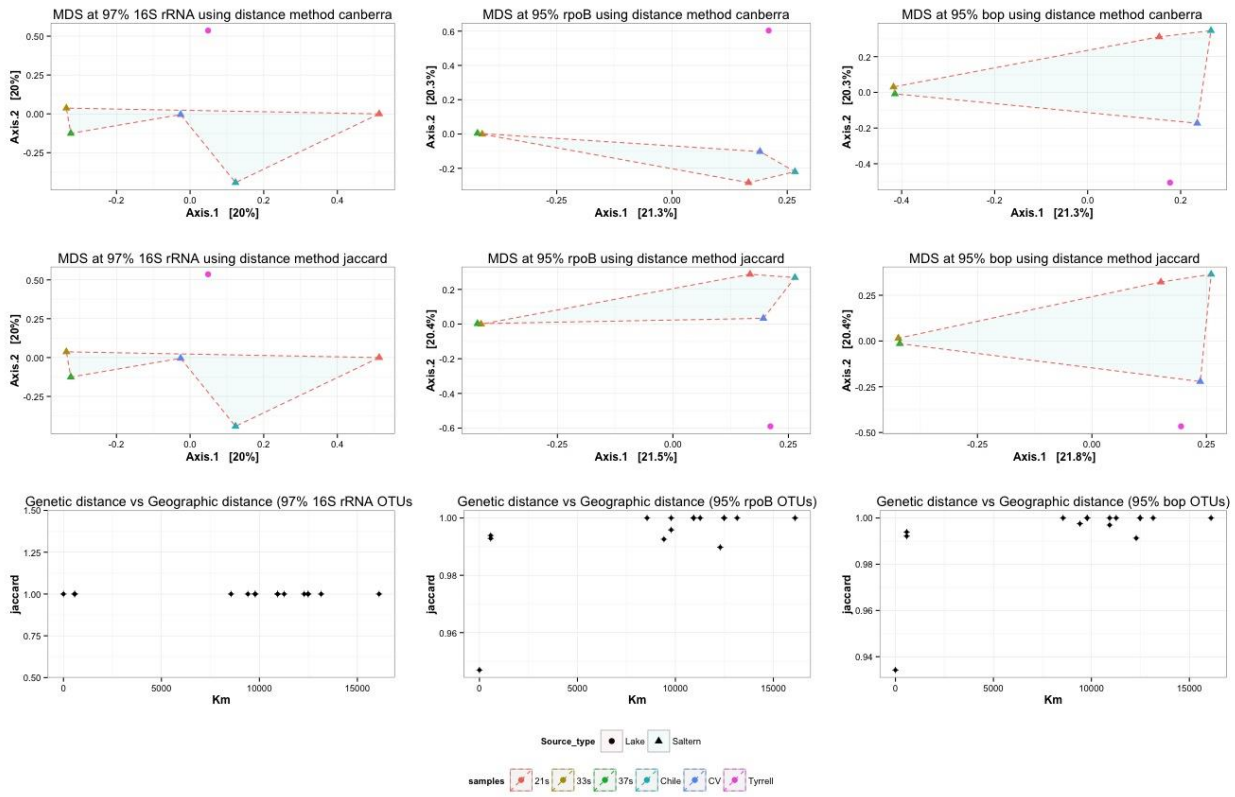


Figure 3-14. Clustering of sampling sites based on geographic proximity and ecological similarity. Distance-decay relationships. Canberra and Jaccard distance calculated between the communities at both 99% and 95% in the 6 metagenomes.

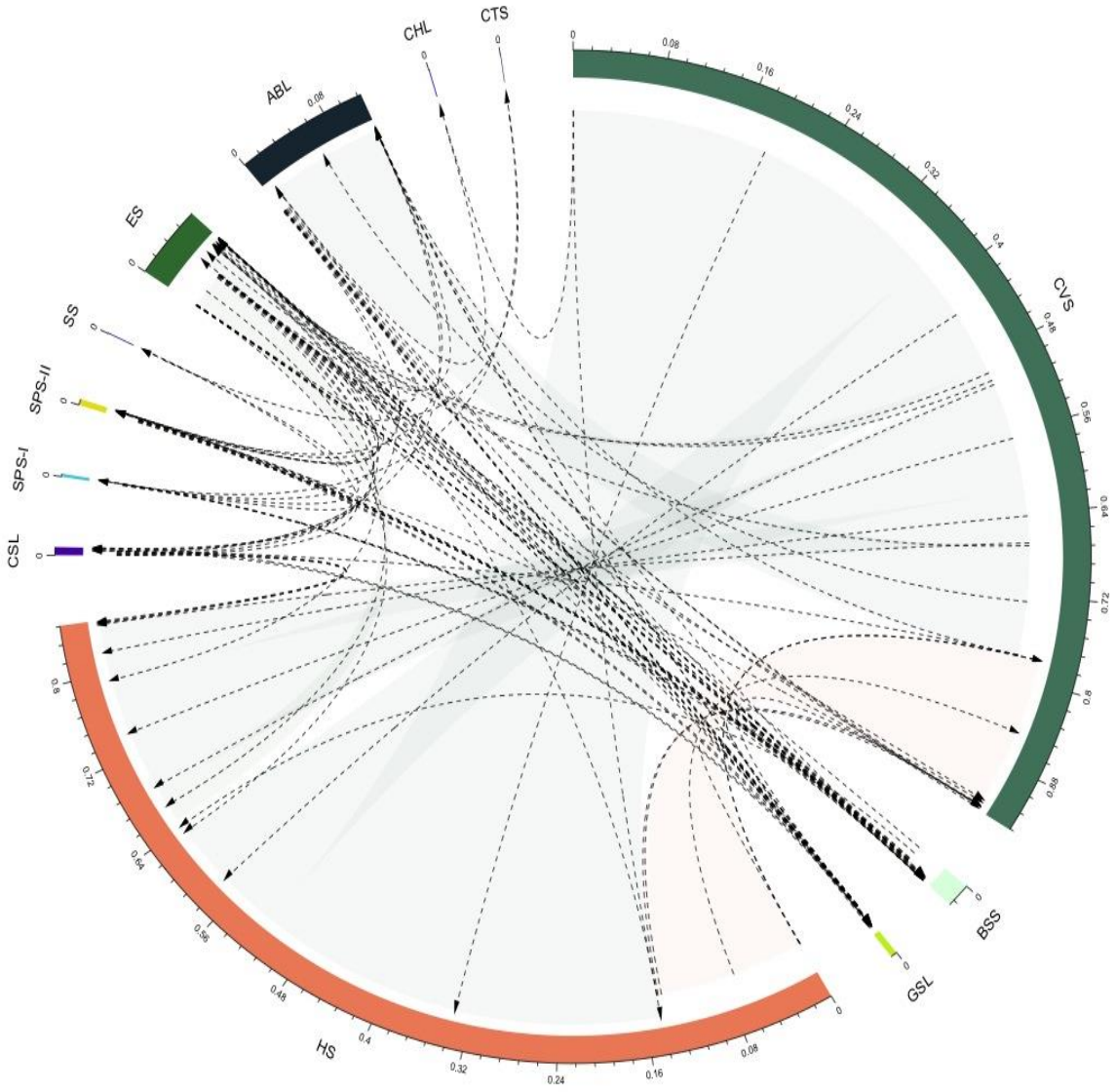


Figure 3-15. Dispersal patterns in the OTUs shared between different PCR-bop sites. Width of the section for each sampling site is a sum of the maximum likelihood estimates for all the immigration or emigration from that site. Dashed lines between sites represent the sharing of OTUs and the arrowhead depicts the predicted direction of dispersal. Going clockwise starting from CVS represents the placement on a map, closest geographic sites are next to each other.

Chapter 4 : Analysis of the bacteriorhodopsin-producing haloarchaea reveals a core community that is stable over time in the salt crystallizers of Eilat, Israel.

4.1 Abstract

Stability of microbial communities can impact the ability of dispersed cells to colonize a new habitat. Saturated brines and their halophile communities are presumed to be steady state systems due to limited environmental perturbations. In this study, the bacteriorhodopsin-containing fraction of the haloarchaeal community from Eilat salt crystallizer ponds was sampled five times over three years. Analyses revealed the existence of a constant core as several OTUs were found repeatedly over the length of the study: OTUs comprising 52% of the total cloned and sequenced PCR amplicons were found in every sample, and OTUs comprising 89% of the total sequences were found in more than one, and often more than two samples. LIBSHUFF and UNIFRAC analyses showed statistical similarity between samples and Spearman's coefficient denoted significant correlations between OTU pairs, indicating non-random patterns in abundance and co-occurrence of detected OTUs. Further, changes in the detected OTUs were statistically linked to deviations in salinity. We interpret these results as indicating the existence of an ever-present core bacteriorhodopsin-containing Eilat crystallizer community that fluctuates in population densities, which are controlled by salinity rather than the extinction of some OTUs and their replacement through immigration and colonization.

4.2 Materials and Methods

4.2.1 DNA isolation and PCR amplification

Brine samples were collected from the reddest pond among the salt crystallizers (301 – 304) at the salt works in Eilat over a period of three years at five time points (see table 1). Four liters of the water collected from the salt crystallizer pond was centrifuged at 6,500 rpm for 30 minutes in a large Sorvall rotor and the cell pellet was collected. DNA from these pellets was isolated using the protocol published in the Halohandbook [142]. Briefly, 400 µl of distilled and deionized water was added to the pellet to lyse cells by osmotic shock. Lysates were then placed in a heating block maintained at 70^{o}C for 10 minutes to inactivate proteins. A working solution of the template DNA for PCR amplification was prepared by making a 1:200 dilution of the crude environmental DNA stock.

The *bop* gene was amplified using primers bop401F and bop795R adapted from [169] and modified to carry the M13 forward and reverse sequences and renamed respectively to bopF_poly (5'-GTA AAA CGA CGG CCA GTG ACT GGT TGT TYA CVA CGC C-3') and bopR_poly (5'-AAC AGC TAT GAC CAT GAA GCC GAA GCC GAY CTT BGC-3'). Using *bop* as the molecular marker circumvents issues of low taxonomic resolution and multiple divergent copies of the 16S rRNA gene observed in many genera of halobacteria [89, 112]. Bacteriorhodopsins are present in significant quantities in solar salterns [134, 167] and are widely expressed among halobacteria living in light-filled environments. The advantages of the bacteriorhodopsin gene as a molecular marker for halobacterial communities has led to several publications [13, 136, 143, 169] and was applied here to each of our sampling time points using the primers to amplify bacteriorhodopsin directly from the community DNA.

The DNA polymerase Phusion (New England BioLabs) was used to produce high fidelity copies of the template. The following PCR cycle protocol was used: One minute initial denaturation at 94°C, followed by 30 cycles of 30 seconds at 94°C, 30 seconds at 53°C and 45 seconds at 68°C. Final elongation occurred at 68°C for 5 minutes. PCR reactions contained Phusion polymerase [2 units/µl], 0.25 µl; 5x GC buffer, 5 µl; DMSO [100%], 2.5 µl; dNTPs [100 mM], 1 µl; forward and reverse primers [10 µM], 0.5 µl each; genomic DNA [20ng/µl], 1.0 µl; and 1.0 µl of dH_2O. Acetamide [25% w/v], 13.25 µl was added to the reaction to improve specificity of primer binding and DMSO was used in order to facilitate denaturation of the high G+C template.

4.2.2 Cloning, plasmid isolation and sequence acquisition

PCR products of the expected length were isolated from a 1% (w/v) agarose gel using the SV Gel & PCR Clean-Up System (Promega) and then cloned using the Zero Blunt TOPO Cloning Kit (Invitrogen) according to the manufacturer's directions. The recombinant plasmid was isolated from the clones for each sample using the Wizard Plus SV Minipreps DNA Purification System (Promega) according to manufacturer's instructions. Plasmids were tested for presence of an insert by restriction digestion with *Eco*RI. The purified plasmids with cloned inserts were sent to GENEWIZ Inc. for sequencing. The sequences obtained in this study were submitted on GenBank under the following accession numbers: KT028773 - KT029121.

4.2.3 Sequence analysis

The obtained raw sequences were manually curated using the commercial software Geneious 4.8.3. Relevant phylogenetic context to be included in community composition detection was extracted from GenBank using TBLASTX [154] (http://blast.ncbi.nlm.nih.gov/Blast.cgi) against all of the currently available, 109 draft and closed halobacterial genomes and the non-redundant nucleotide database. BLAST hits with the lowest e-value and the highest query coverage were added to our analyses. All bacteriorhodopsin sequences were aligned using MUSCLE [148]. Multitaxa alignments were edited using MacClade 4.08 [149].

4.2.4 Bacteriorhodopsin producing Halobacterial community analysis

The aligned bacteriorhodopsin sequences were analyzed using Mothur v.1.20.0 [138] to estimate the species richness, community diversity indices and the similarity between the bacteriorhodopsin producing halobacterial communities, referred to simply as haloarchaeal communities, recovered from each sampling time point. Operational Taxonomic Units (OTUs) were determined at 100, 99, 97 and 95% sequence similarities using the average neighbor clustering, and the different OTUs were used for further analyses. 99% sequence similarity was designated as a stringent estimate for OTU clustering. A previous study showed that bop gene sequences with less than 1% variation formed species-like clusters [84]. 95% was chosen as a liberal estimate for the OTU definition. Presence of outliers in the observed number of OTUs over time was tested three ways – the extreme Studentized deviation (ESD identifier) [170]; Hampel identifier [171]; and the standard boxplot rule [172]. Species richness estimators Chao1 [152], which is nonparametric and bases values on the number of singletons and doubletons, and Ace [153], which is based on the number of rare groups of observed OTUs (OTU_{obs}), were calculated

for each time point. Shannon index (H) and Simpson index of diversity (1-D) were also estimated as a measure of the species diversity and evenness at each time point. Rarefaction curves were generated within Mothur to estimate the sampling completeness and efficiency for each time point.

4.2.5 Phylogenetic reconstruction

All 349 sequences were used to construct a maximum likelihood tree from distances calculated under the Generalized Time-Reversible model [173] within the PHYML 3.0 [174] phylogenetic program. Any OTU at 99% that contained sequences from at least two time points was termed a 'shared OTU'. An OTU was defined as a 'cumulative shared OTU' when it contained two or more 99% shared OTUs that combined into a single OTU at 95%. The shared OTUs at 99% and the cumulative shared OTUs at 95% were labeled 'sOTU number' and 'csOTU number' respectively. The tree was viewed, edited and midpoint rooted using FigTree (http://tree.bio.ed.ac.uk/software/figtree/), a tree editing software.

4.2.6 Community comparisons

The halobacterial communities recovered from the sampling time points were compared three ways. First, an online tool – UNIFRAC [140] was employed. The mid-point rooted maximum likelihood tree (not shown) of all the sequences and an environmental file listing the sequences from each sampling time point were uploaded onto the UNIFRAC tool to perform the JackKnife Environmental Clustering Analysis. Based on the clustering of the sequences observed on the maximum likelihood tree, UNIFRAC estimated a community level pairwise distance matrix. This was used to develop a UPGMA dendrogram of the communities and the JackKnifing

provided support for the clustering of the sample sites. Second, the LIBSHUFF command within Mothur was employed. The command within Mothur implemented the original LIBSHUFF program [155]. It tested for similarity in structure between two or more communities by incorporating the Cramer-von Mises test statistic [156] and returning a significance value for the difference between each pair in consideration. Finally, the OTUs defined at 99, 97, and 95% sequence similarity were manually curated to determine the OTUs with sequences from different time points clustering together. Pairwise OTU correlations were estimated by determining the Spearman's rank correlation coefficient [175] within Mothur. The correlation coefficient estimated the relatedness between the relative abundances of each pair of OTUs.

4.3 Results

4.3.1 Sample compositions and abundance of genera through time

The top BLAST hit for the representative sequence of each OTU at 95% was used to determine the taxonomic affiliations of the community members. Normalized bar plots (Figure 4-1) were constructed with these top BLAST hits from the non-redundant database (Figure 4-1a) and then only from available genomes (Figure 4-1b) with >50% identity to pictorially represent the observed community structure through time. While each time point retrieves top BLAST hits from the genus *Halorubrum*, the non-redundant database has many sequences from environmental studies and therefore the top BLAST hit for over 60% of each community is of unknown taxonomic affiliation (Figure 4-1a). Other recovered top BLAST hits from this database of known affiliation were *Haloarcula*, *Halomicrobium* and *Halosimplex* and each of these was recovered only once. To more accurately assign sequences to taxa, we retrieved the top BLAST hits querying only the sequenced haloarchaeal genomes, which revealed additional details regarding taxonomic

affiliations. Figure 4-1b shows that 11 groups were identified with two genera (*Halomicrobium* and *Halosimplex*) recovered only once, four genera were recovered twice (*Haloquadratum, Natronorubrum, Haloplanus* and *Halobiforma*), one genus was recovered thrice (*Haloferax*), one group was recovered four times (*Haloterrigena*) and three genera were recovered in every sample (*Halorubrum*, *Haloarcula,* and a taxonomically unclassified group with sequence identity <50% of the bacteriorhodopsin genes in the genomes). With the exception of *Halomicrobium* and *Halosimplex*, each genus was detected from more than one sample period, varying in their relative sequence abundances by our methods, which suggests that every genus detected is a long term member of the existing community, but below our ability to detect them in some samples.

4.3.2 OTUs are shared between samples

A five-way Venn diagram was constructed to demonstrate community overlap using different OTU definitions at 99, 97 and 95% (Figure 4-2). Identified in this analysis is the existence of a set of OTUs at the 95% definition that are found in all five samples through time. These four OTUs called the 'core' encompass 179 of the 349 overall sequences (~52%) obtained in this study and based on top BLAST hits belong to *Halorubrum* (cs03 and s01), *Haloarcula* (cs01), and one unknown (cs06). Many of the other 95% OTUs were detected at more than one time point. These mercurial OTUs represents ~37% of the overall sequence data. Within this 37%, ~3% cluster into 1 OTU (s08), belonging to *Haloterrigena*, that occurs in four out of five time points. Approximately 19% is shared between three time points and form five OTUs (s21, cs02, s30, s04, and cs07). ~15% cluster into 13 OTUs (s06, s07, s09, s10, s03, s31, s27, s22, s23, cs05, cs04, s19, and s20) and are found at two time points. A fraction (~11%) of the total sequences formed 15 95% OTUs that were only retrieved at one time point in our study (Aug'11: 9; May'12: 1; Apr'13:

4; Jan'10: 1). All sampling time points had unique OTUs detected, except in the June '10 sample. With one exception, all OTUs unique to a sample are represented by a single sequence (singleton) (12), two sequences (doubleton) (1) or three sequences (tripleton) (1): The exception being from the August '11 sample which had a unique 95% OTU (Aug'11-02) that was comprised of 22 sequences. The rarely retrieved OTUs, including *Halomicrobium*, *Halosimplex*, *Haloquadratum*, *Natronorubrum*, *Haloplanus*, *Halobiforma*, and *Haloferax* and unknown genera, are likely due to sampling limitations.

To determine if there was a dependent, non-random abundance relationship between members of core OTUs the Spearman's rank correlation coefficient (ρ) was estimated for each pair compared. ρ was determined for 100% identical sequences (100% OTUs) found within the core 95% OTUs and only p values < 0.05 were considered significant. The values range from -1 to 1, indicating a negative and positive correlation respectively between the relative abundances of the two 100% OTUs compared. Figure 4-4 is a plot of the statistically significant correlation coefficients, which shows most 100% OTU pairs have a strong positive correlation ($\rho = 0.88 – 1.00$). Three 100% OTU pair comparisons, which belong to the core 95% OTUs cs01 and cs03, returned a strong negative correlation (ranging from $\rho = -0.89$ to -0.91), while the remaining significant interactions of those two core OTUs showed a positive correlation. Comparison between other core OTUs resulted in negative correlation coefficients but were not statistically significant. The high numbers of shared OTUs and the strong correlation between them indicate that though PCR and cloning is not a quantitative technique, the sequence data retrieved were not random with respect to the abundances detected, and therefore dynamics seen in sequence abundance probably reflect real fluctuations in natural population sizes and not an indication of newly colonized cells invading the habitat.

4.3.3 Sampling efficiency and richness estimations

To estimate sampling efficiency, rarefaction curves were generated for each sampling time point for OTU definitions of 99% and 95% sequence similarity (figure 4-5). Though the curves never flatten for the August ‘11, May ‘12, and the April ‘13 samples at 99% suggesting further sampling is required, they do begin to level, and at 95% OTU definition, the plots plateau suggesting reasonable sampling at 95% sequence similarity. The January ‘10 and June ‘10 curves at both 99 and 95% sequence similarity indicates abundant sampling from these time points, suggesting that all time points are well sampled, but not completely, at 95% OTU definitions.

Species richness estimators for each sampling time point were calculated and plotted (figure 4-6). At the 95% sequence similarity definition, the observed number of OTUs is similar to the estimations of species richness for both Chao and Ace, whereas at the 99% and 97%, the observed number of OTUs was lower than the richness estimates. Though variation in richness estimations exists, the observed OTUs through time do not drastically change. There are no outliers in the data range as measured by the ESD identifier, Hampel identifier, and the standard boxplot rule, which together indicates that community diversity is statistically equivalent through time. The communities at each time point appear rich and diverse in OTUs (Table 2). The Shannon index (H) is a commonly used diversity index that ranges between 1.5 and 3.5 and takes into account both abundance and evenness of species observed in the community. H estimations suggest a diverse community at each time point, similar to the findings from the rarefaction curves. Simpson’s index of diversity (1-D) ranges between 0 and 1, and describes the sample diversity. The estimated 1-D values corroborate the findings from the rarefaction curves. Since sampling

efficiency is good but not complete, the absence of an OTU in some samples while present in others is not evidence for the absence of the sequence from the community. To test this, we performed a correlation test between the species richness estimations and salinity and showed that the Chao (correlation coefficient = 0.72) and Ace (correlation coefficient = 0.85) estimations of the low abundance members changed with respect to the salinity: therefore, increases in salinity possibly due to higher temperatures likely facilitated changes in population sizes as seen in other studies [59, 165, 176, 177] and had an effect on our ability to detect members of the community.

4.3.4 UNIFRAC and LIBSHUFF analyses show statistical similarity in OTUs through time

We statistically evaluated the observed between sample similarities at different times using UNIFRAC, which analyzed a maximum likelihood tree containing all 349 *bop* sequences and an environmental file listing the sampling time point each sequence. The pairwise UNIFRAC distances calculated are listed in table 4-3. The UNIFRAC distances between each sampled bacteriorhodopsin-containing community pair at different times ranged from 0.36 to 0.62, which statistically confirms the qualitative observation for the existence of core OTUs, and that a large number of non-core OTUs are shared between samples. Jackknifing analysis was performed and the UPGMA dendrogram that was derived from the UNIFRAC output data is shown in Figure 4-3. The small UNIFRAC distances, and the Jackknifing statistic for samples indicates that the January ’10 and May ‘12 samples, as well as the April ’13 and June ’10 samples are robustly and statistically similar. These analyses indicate the presence of similar bacteriorhodopsin-containing communities through the years, since the highly supported clustering seen between the January ‘10 and May ’12 samples, and the June ‘10 and April ‘13 samples is a function of detecting many of the same OTUs between those time points.

LIBSHUFF carries out pairwise comparisons to determine if one data set is a subset of the other. Allowing for a 5% false detection rate and applying the Bonferroni correction for multiple library comparisons, only p values less than 0.0025 are considered statistically significant for inferring that two samples are different. Results from LIBSHUFF analyses are presented in table 4-4 and are mapped onto the UNIFRAC derived similarity data in Figure 4-3. The January '10 and May '12, and August '11 and May '12 data sets are subsets of each other. Each sample except for August '11 is a statistically robust subset of April '13. These results indicate that the bacteriorhodopsin-containing haloarchaeal community composition is statistically similar through time and that differences seen in OTU absence/presence likely reflect the natural rise and fall in taxon abundances above and below our detection limits, rather than the colonization of new taxa.

4.4 Discussion

Bacteriorhodopsin, a member of the halobacterial rhodopsin protein family, is a light-driven proton-pump that generates an electrochemical gradient for the production of ATP [178], and it is present in significant quantities in hypersaline environments [133, 134, 179]. In this study, and others, it is shown to recover the familiar genera and diversity in halophilic environments when compared to the 16S rRNA gene [13, 78] and provides excellent support for binning haplotypes [13]. Similar to the studies using 16S rRNA genes to survey Haloarchaea in hypersaline environments (e.g., [12-14, 58, 63, 64, 76, 81, 166], we did find unknown diversity: ~85% of the sequences returned an uncultured haloarchaeon from the non-redundant database and ~24% were considered unknown from comparing against the 109 sequenced genomes. Haloarchaea have

multiple, often highly divergent 16S rRNA gene copies, that greatly biases the interpretation of data in environmental analyses [89, 180]. Further, because the 16S rRNA gene is highly conserved, and recombines easily between haloarchaeal species [84, 89] diversity is hidden, as two species could easily share the same sequence. By using the bacteriorhodopsin gene, we circumvented many of those complications.

Haloquadratum was previously suggested to be a dominant organism in this salt crystallizer pond using morphological criteria [181] and polar lipid composition profiling [182]. However, in this study, it was recovered from only two samples: August ’11 and May ’12. There seems to have been a possible bloom of *Haloquadratum* in the August ’11, correlating positively to the increase in salinity (Table 1). These data are in agreement with other studies showing that ion concentrations are correlated with *Haloquadratum* abundance [59]. Other ecological conditions like, rain, wind, and temperature remained stable for weeks prior to each sampling time point. One month prior to each sampling, there is no record of rain, wind speeds fluctuated from 6.4-19 kmh^{-1} and the difference in temperature between the coolest and warmest day was 19^{o} C. (see http://www.weatherundergound.com). PCR biases that cause differences in the ability to amplify DNA are known to exist [183]. However, the primers used have recovered *Haloquadratum* sequences in previous studies [13, 78, 169] with *Haloquadratum*-related sequences being the most abundant in [13]. In total, these results suggest that *Haloquadratum* is a member of this community, often below our detection limits, but that sometimes it dominates when the salinity conditions are favorable.

Many studies on soils, rivers, lakes or other aquatic microbial communities indicate a lack of temporal stability [47, 48, 50-54, 184, 185]. However, those communities also experienced large fluctuations in environmental conditions. Further, dramatic community shifts in response to perturbations may only cause the natural abundances of native taxa to rise and fall without the gain or loss of taxa (e.g., [56]). The seasonal environmental changes in southern Israel are markedly less fluctuating, and our results indicate the existence of a core set of bacteriorhodopsin OTUs in the Eilat salt crystallizer pond that experiences some abundance fluctuations in reaction to salinity changes. In a previous study surveying the impact of salinity and seasonal changes on microbial diversity in the Israel Salt Industries Ltd. in Eilat, similar findings were reported: two 16S rRNA gene OTUs were identified as present in every sample regardless of the salinity or season with varying relative ratios between the two [137]. In this study, four OTUs were retrieved from all five sampling time points over three years and comprised ~52% of the overall sequences obtained. Further, OTUs covering 89% of the total sequences were found in more than one time point and, the June ’10 sample shared every one of its OTUs with another sample suggesting the OTUs were present throughout, sometimes escaping detection. Statistical analyses support this interpretation. The Spearman’s correlation shows that the individuals detected are not present by chance; the detected number of OTUs at each sampling time is statistically neither over nor under abundant regardless of OTU definition; OTUs abundances are significantly negatively and positively correlated indicating detected sequences likely capture natural fluctuations of the bacteriorhodopsin-containing populations; both LIBSHUFF and UNIFRAC analyses determine that the datasets are highly similar and each is typically a subcomponent of the other. An alternative explanation for the data is that the differences seen in composition between sample times are due to the colonization of dispersed cells from other environments that then outcompete the established

Eilat Saltern cells for niche space and rise in population frequency to above our detection limits. This alternative explanation with frequent colonization events is non-parsimonious, and is not supported by the statistical analyses

Our interpretations above are further corroborated by other studies on hypersaline environments and indicate haloarchaeal communities are likely to be globally stable with minor oscillations in population abundances that are correlated with environmental factors. Temporal studies on the thalassohaline Lake Tyrrell in Australia [59]; the saturated brines from Sfax solar saltern in Tunisia [165]; the Bras del Port solar saltern in Santa Pola, Spain [143, 176, 186]; and the South Bay Salt Works in Chula Vista, USA [177], all showed that the detected haloarchaeal community membership at the sequence, OTU, and/or genus level was largely constant across sampled time points. Fluctuation in taxon relative abundances was observed in Lake Tyrrell, and there was an ion concentration-dependent negative correlation in the abundances of some genera (e.g., *Halorubrum* and *Haloquadratum*), while other community members like the Nanohaloarchaea [12, 63, 187] and *Halorhabdus* showed no correlation [59]. Similar results were observed in Sfax solar saltern where 95% of the OTU and sequence fluctuations correlated with ecological parameters [165]. This is consistent with what is observed in our study where abundances within core OTUs varied with salinity, and most non-core OTUs were found in multiple samples. Bacterial components of hypersaline communities also appear to have similar outcomes [143, 186], indicating a general phenomenon of extreme hypersaline environments. By and large, hypersaline environments appear to provide constant ecological conditions, probably because they are typically found in hot dry climates, and appear to promote locally stable microbial communities across the globe.

4.5 Conclusions

This study recovered the presence of a stable OTU core representing the bacteriorhodopsin-producing haloarchaeal community inhabiting the salt crystallizing ponds in Eilat, Israel. These four core OTUs comprised ~52% of overall sequences with only ~11% of the sequencings being unique to any one time point, and contained the following genera - *Halorubrum*, *Haloarcula*, *Haloterrigena*, *Halomicrobium*, *Halosimplex*, *Haloquadratum*, *Natronorubrum*, *Haloplanus*, *Halobiforma*, and *Haloferax*. Fluctuations in relative OTU abundances corresponded to natural variations of salinity. Because evidence from many temporal studies on hypersaline habitats, especially saturated brines, also demonstrate the existence of stable haloarchaeal and bacterial communities worldwide, we hypothesize that communities are resistant to dispersed invading species, at least temporarily, and might lead to the formation of endemic hypersaline adapted communities and populations.

Table 4-1: Sample information

Date	Year	Sample name	% Salinity	No. of clones sequenced
3rd January	2010	Winter ‘10	31.2	73
20th June	2010	Summer ‘10	32	70
25th August	2011	Fall ‘11	35.2	74
15th May	2012	Spring ’12	34.4	67
22nd of April	2013	Spring ‘13	33.4	65

Table 4-2: Species diversity and evenness estimations at the different sampling time points.

	99%			97%			95%		
	OTU_{obs}	Shannon	Simpson	OTU_{obs}	Shannon	Simpson	OTU_{obs}	Shannon	Simpson
January '10	16	2.17	0.18	9	1.50	0.35	9	1.50	0.35
June '10	14	2.20	0.12	12	1.99	0.17	11	1.97	0.17
August '11	28	2.69	0.11	24	2.51	0.12	22	2.45	0.12
May '12	28	2.59	0.14	18	2.11	0.19	17	2.06	0.20
April '13	33	3.23	0.03	24	2.90	0.05	21	2.77	0.06
Mean ± s.d	23.8 ± 8.32			17.4 ± 6.81			16 ± 5.83		

Table 4-3. Pairwise UNIFRAC distances between communities at different sampling time points.

January ‘10				
April ‘13	0.554425			
August ‘11	0.598268	0.574907		
June ‘10	0.379003	0.360183	0.621469	
May ‘12	0.386263	0.549493	0.533978	0.489596

Table 4-4. LIBSHUFF pairwise comparisons of community structures. Only significance values less than 0.0025 were considered statistically significant and denoting different communities. Community comparisons that do not fall within this threshold are listed.

Comparison	dCXYScore	Significance
Jan'10 – Apr'13	0.002789024	0.90257
Jan'10 – Jun'10	0.00351532	0.0633
Jan'10 – May'12	0.00462918	0.4743
May'12 – Jan'10	0.00850827	0.3013
Jun'10 – Apr'13	0.00073510	0.7540
May'12 – Apr'13	0.00375615	0.9122
Aug'11 – May'12	0.02594856	0.0306
May'12 – Aug'11	0.00754265	0.4193
May'12 – Jun'10	0.00408351	0.4289

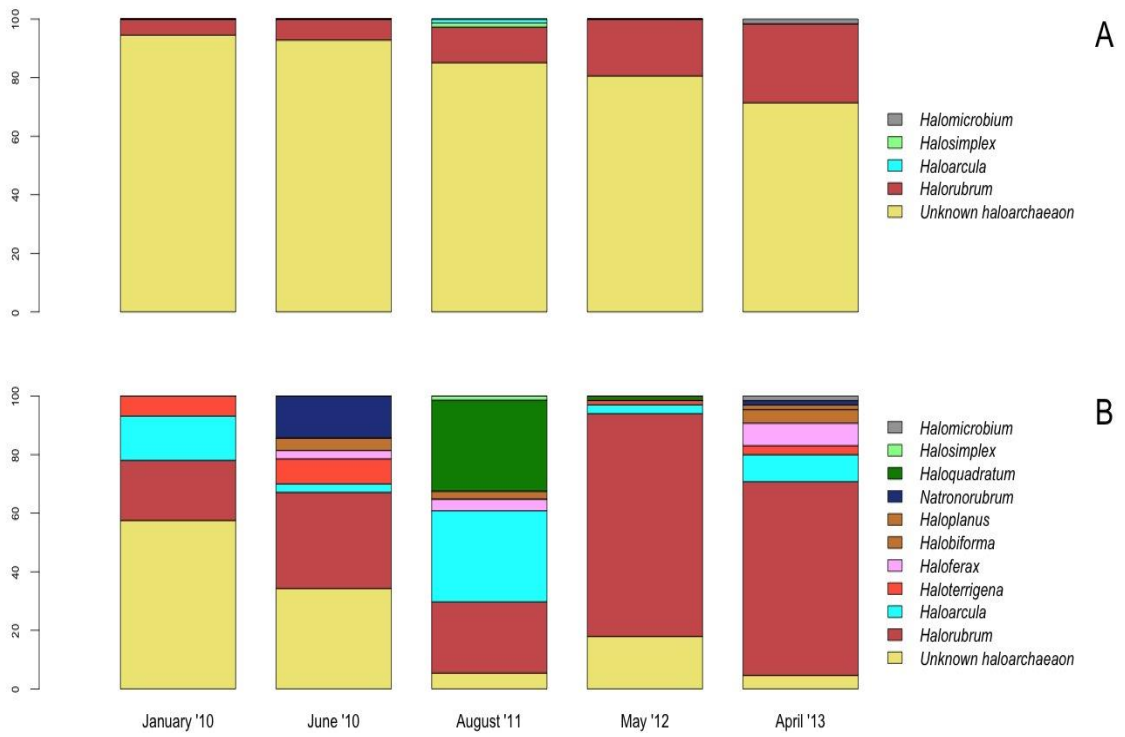


Figure 4-1. Community composition over time. Normalized bar plots were constructed at each time point based on the top BLAST hits for representative sequences of each OTU at 95%. **A.** Top BLAST hits from tBLASTx against entire non-redundant nucleotide database. **B.** Top BLAST hits from the 109 available genomes by performing a stand along tBLASTx.

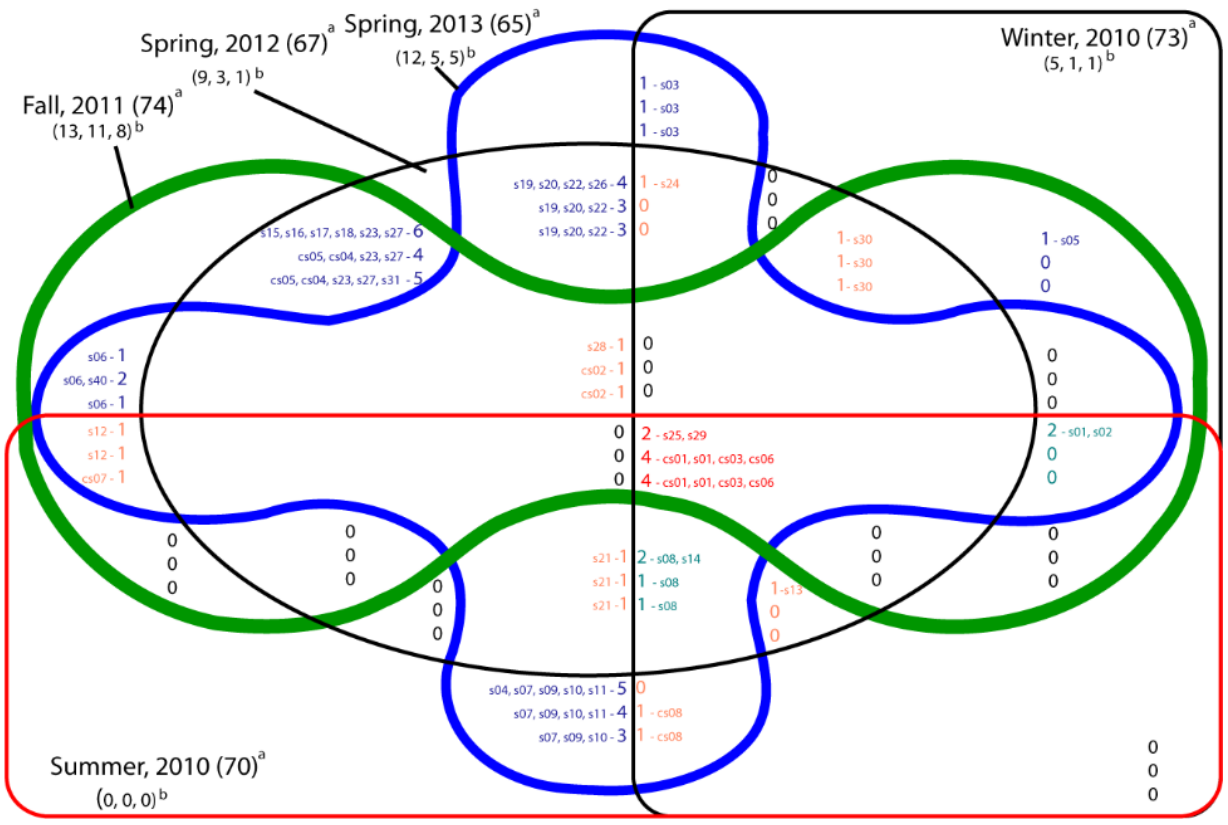


Figure 4-2. Sharing of OTUs between time points. Number of shared OTUs represented vertically at 99%, 97%, 95% sequence similarities respectively. (x)[a] – total number of sequences obtained at that time point: (x,y,z)[b] – number of OTUs unique to the sampling time point represented for 99%, 97%, 95% sequence similarities respectively.

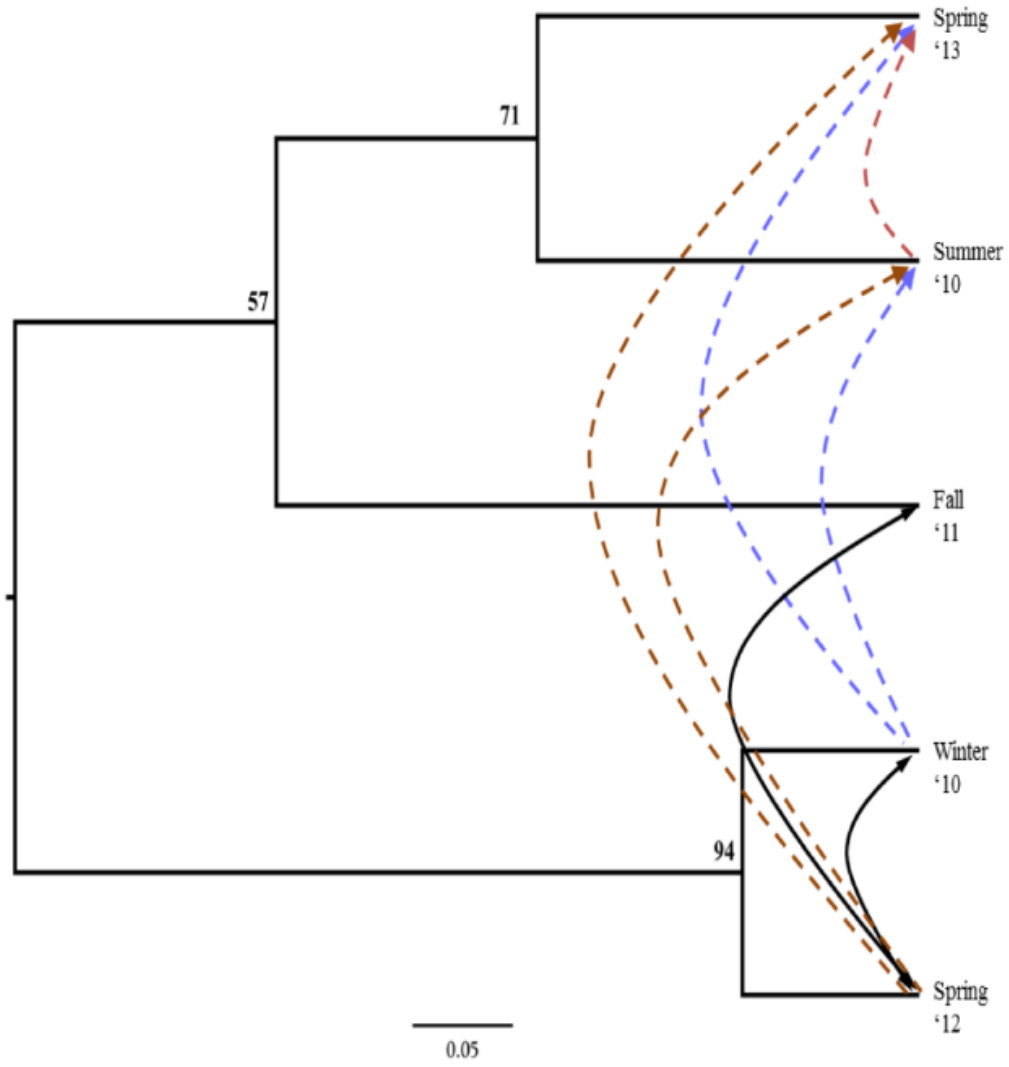


Figure 4-3. Community dynamics. UPGMA dendrogram of the environments based on the clustering of the sequences from each time point on the maximum likelihood tree (all 349 sequences). Jackknifing provided the branch support. Statistically insignificant values (>0.025) from LIBSHUFF pairwise analysis are plotted. Unidirectional dashed arrows indicate that one community is a subset of the other (brown: May '12; blue: January '10; red: June '10) and bidirectional solid arrows indicate two communities are subsets of each other.

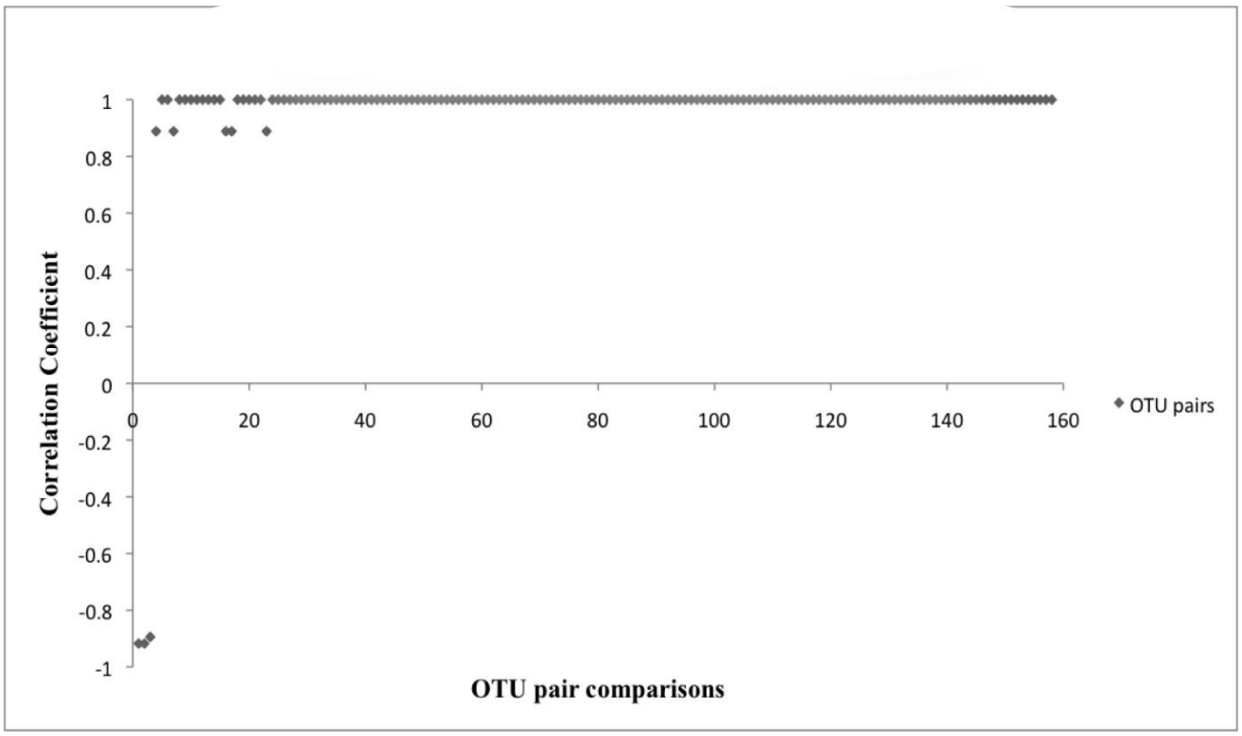


Figure 4-4. Plot of the Spearman correlation coefficients determined for the relative abundances of each pair OTUs within the core OTUs at 100% sequence similarity. Only correlation coefficient values with significance ($p < 0.05$) were considered statistically significant and plotted.

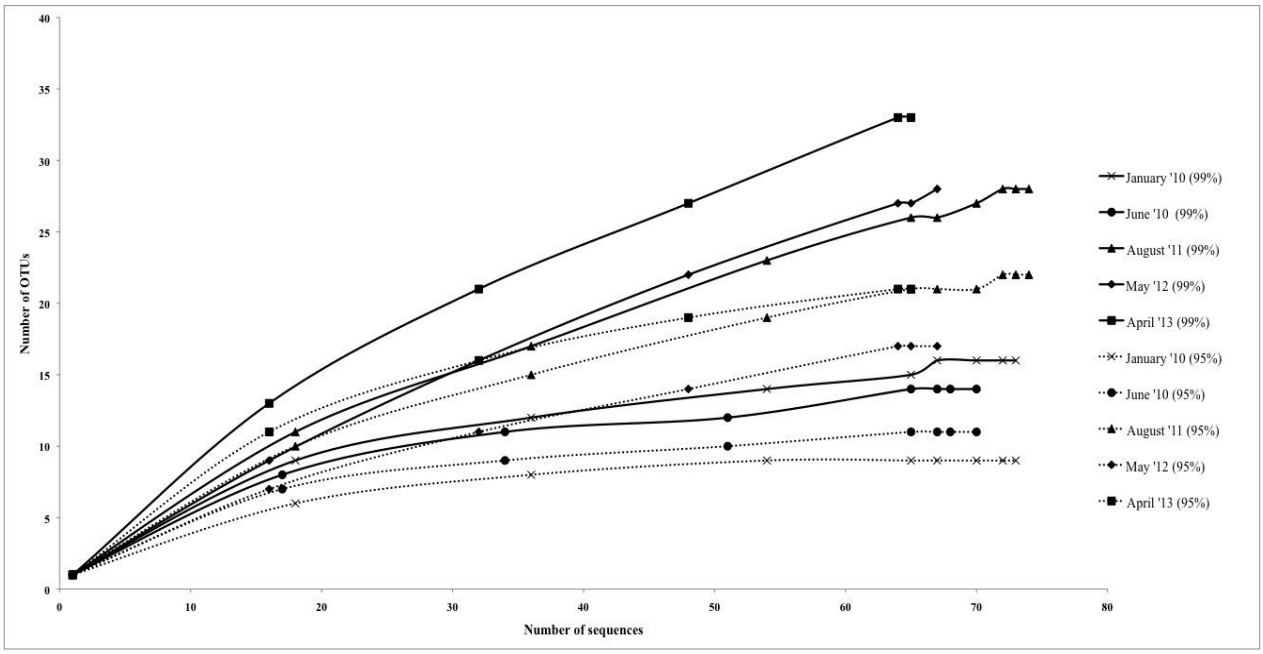


Figure 4-5. Rarefaction curves generated for each sampling time point at 99% and 95% sequence similarity.

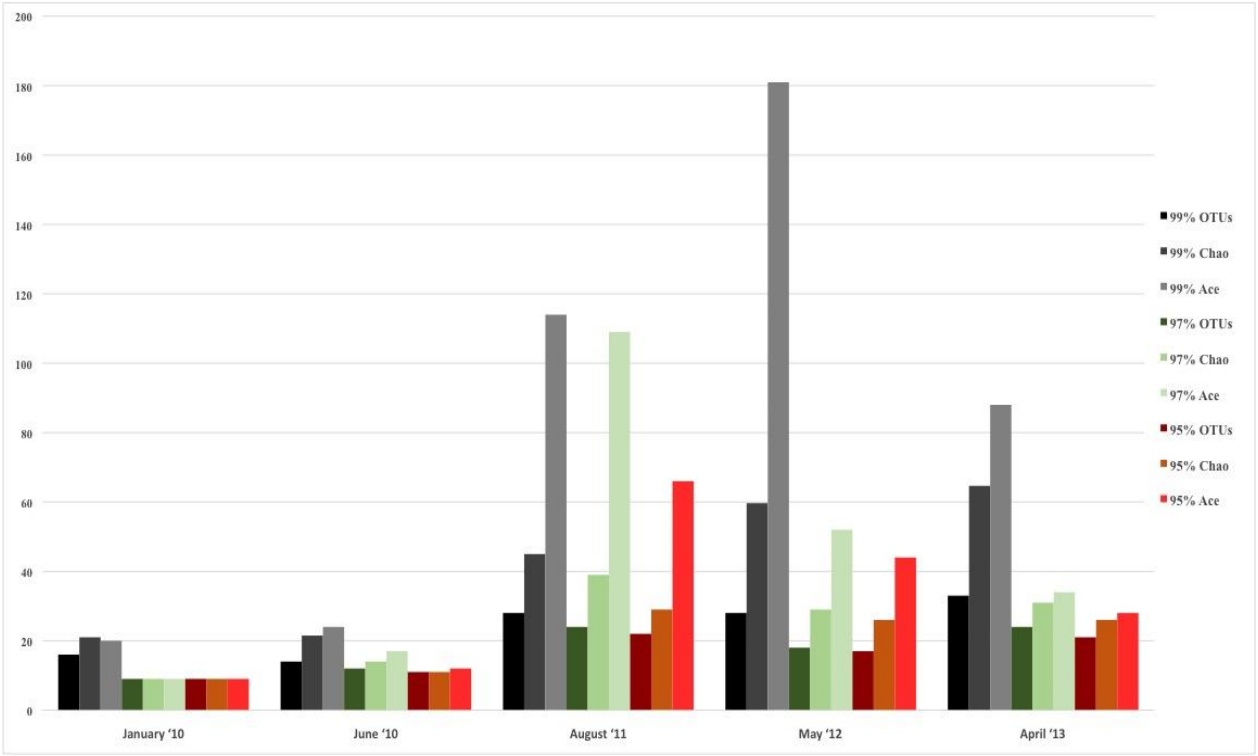


Figure 4-6. Species richness estimations. Plot of the observed OTUs, Chao and ace species richness estimators for each sampling time point at 99, 97, and 95% sequence similarity.

Chapter 5 : Evidence from phylogenetic and genome fingerprinting analyses suggests rapidly changing variation in *Halorubrum* and *Haloarcula* populations [100]

5.1 Abstract

Halobacteria require high NaCl concentrations for growth and are the dominant inhabitants of hypersaline environments above 15% NaCl. They are well documented to be highly recombinogenic, both in frequency and in the range of exchange partners. In this study, we examine the genetic and genomic variation of cultured, naturally co-occurring environmental populations of Halobacteria. Sequence data from multiple loci (~2500bp) identified many closely and more distantly related strains belonging to the genera *Halorubrum* and *Haloarcula*. Genome fingerprinting using a random priming PCR amplification method to analyze these isolates revealed diverse banding patterns across each of the genera and surprisingly even for isolates that are identical at the nucleotide level for five protein coding sequenced loci. This variance in genome structure even between identical multilocus sequence analysis (MLSA) haplotypes indicates that accumulation of genomic variation is rapid: faster than the rate of third codon substitutions.

5.2 Materials and Methods

5.2.1 Growth conditions and DNA extraction

Aran-Bidgol *Halorubrum* and *Haloarcula* spp. cultures were grown in Hv-YPC medium [188] at 37°C with agitation. DNA from Haloarchaea was isolated as described in the Halohandbook (http://www.haloarchaea.com/resources/halohandbook/). Briefly, stationary-phase cells were pelleted at 10000x*g*, supernatant was removed and the cells were lysed in distilled water. An equal volume of phenol was added, and the mixture was incubated at 65°C for one hour prior to centrifugation to separate the phases. The aqueous phase was reserved and phenol extraction was repeated without incubation, and followed with a phenol/chloroform/iso-amyl alcohol (25:24:1) extraction. The DNA was precipitated with ethanol, washed, and resuspended in TE (10mM tris, pH 8.0, 1mM EDTA). Type strains were grown, and DNA was purified as described by [31].

5.2.2 Sequence acquisition for MLSA

Five housekeeping genes were amplified using PCR. The loci were *atpB*, *ef-2*, *glnA*, *ppsA* and *rpoB* and the primers used for each locus are listed in Table 5-1. To more efficiently sequence PCR products, an 18bp M13 sequencing primer was added to the 5' end of each degenerate primer (Table 5-1). Each PCR reaction was 20µl in volume. Phire Hot Start II DNA polymerase (Thermo Scientific) was used in the amplification reactions. The PCR reaction was run on a Mastercycler Ep Thermocycler (Eppendorf) using the following PCR cycle protocol: 30 second initial denaturation at 98°C, followed by 40 cycles of 30 seconds at 98°C, 5 seconds at the annealing temperature for each set of primers and 15 seconds at 72°C. Final elongation occurred at 72°C for

1 minute. Table 5-2 provides a detailed list of reagents and the PCR mixtures for each amplified locus. The PCR products were separated by gel electrophoresis with agarose (1%). Gels were stained with ethidium bromide. An exACTGene mid-range plus DNA ladder (Fisher Scientific International Inc.) was used to estimate the size of the amplicons, which were purified using Wizard SV gel and PCR cleanup system (Promega). The purified amplicons sequenced by Genewiz Inc. The sequences obtained for the five genes in this study were submitted on Genbank under the following accession numbers: KJ152221 - KJ152260, KJ152261 - KJ152298, KJ152362 - KJ152397, KJ152398 - KJ152433, and KJ152323 - KJ152361.

5.2.3 Phylogenetic Analysis

Type strain genomes were obtained from the NCBI ftp repository. Blast searches identified DNA top hits for each MLSA target gene (*atpB, ef-2, glnA, ppsA* and *rpoB*) in each genome. Multiple-sequence alignments (MSAs) were created from the DNA genome hits as well as the PCR amplicons using MUSCLE [148] (alignments available upon request) with its refine function. The MSA length was manually trimmed down to the lengths of the PCR amplicons. In-house scripts created a concatenated alignment of all five genes. A model of evolution was determined using the Akaike Information Criterion with correction for small sample size (AICc). The jModelTest 2.1.4 [189] program was used to compute likelihoods from the nucleotide alignment and to perform the AICc test [190]. The AICc reported the best-fitting model to be GTR + Gamma estimation + Invariable site estimation. A maximum likelihood (ML) phylogeny was generated from the concatenated MSA using the PhyML v3.0_360-500 [174]. The model used in PhyML corresponded to the one favored by jModelTest: GTR model, estimated p-invar, 4 substitution rate

categories, estimated gamma distribution with 100 bootstrap replicates. The number of nucleotide differences in pairwise comparisons were determined using MEGA 5 [191].

5.2.4 Genomic Fingerprinting

In total, DNA from 81 haloarchaeal type strains and 43 isolates from the Aran-Bidgol Lake were tested. Each primer selected has successfully been used in genome fingerprinting in previous studies. Primers P1 and P2 were used to fingerprint *Vibrio harveyi* bacteriophages [192], primers OPA-9 and OPA-13 were used to asses marine viral richness [193]. The last primer, FALL-A was adapted from the primer used [193, 194] to study bacteriophages isolated from an industrial sauerkraut fermentation. Amplification conditions for each strain were equal to enable accurate comparison between banding patterns obtained. Each sample was diluted to 20ng μl^{-1} and amplified within the following reaction mixture: 12.5µl SYBR Universal Faststart Mastermix (Roche), 4.5µl dH_20, 1.5µl for each of five primers at 10ng μl^{-1} (see Table 5-3), and 0.5µl of template DNA. Two thermocycler programs were used in succession. The first included an initial 10 minute denaturation at 94ºC, followed by 4 cycles of a 45 second denaturation also at 94ºC, annealing at 30ºC for 2 minutes and extension at 72ºC for 50 seconds. This was followed by another 35 cycle program: 94ºC for 17 seconds, 36ºC for 30 seconds, and 72ºC for 45 seconds, and a final extension for 10 minutes at 72ºC. The aim of these repeated programs with low annealing temperatures and long annealing times is to produce as many nonspecific bands as possible for each sample, increasing the resolving power of the method. Strains were amplified in triplicate to ensure that a repeatable banding pattern could be obtained.

5.2.5 Gel electrophoresis

Reactions mixtures from PCR experiments were held at 4ºC prior to electrophoresis. Standard DNA electrophoresis was carried out with replicates from each strain. Gels were 1.5% agarose and run at 12v for 16 hours at 4ºC with the goal of producing crisp bands easily distinguishable by the analysis software. Gels were stained with ethidium bromide prior to imaging.

5.2.6 Imaging and Analysis

A digital image of each gel was created using a GelDoc (UVP). Images were then analyzed using the Phoretix 1D Pro program from the TotalLab Inc. (www.totallab.com). Banding patterns were standardized for cross gel comparisons by calibrating Rf lines on each individual gels. Phoretix 1D Pro converts banding patterns into a format that can be used to produce a dendrogram comparing the differences and similarities between the patterns of amplicons. The final dendrogram was created within Phoretix 1D Pro using UPGMA statistical analysis on Dice coefficients [195] for each of the lanes. A measure of the correlation between the matrix similarities and the dendrogram derived similarities, the cophenetic correlation coefficients [196] were determined for each sub-cluster of the dendrogram and displayed on the nodes of the constructed dendrograms to estimate the robustness of each cluster.

5.3 Results

5.3.1 Genomic Fingerprinting

The repeatability of banding patterns, and thus the success of the fingerprinting technique was tested on 81 haloarchaeal type strains. The PCR on each of the 81 was run in triplicate and the products were run on adjacent wells. Figure 5-1 demonstrates results of the banding pattern for 18 out of the 81 type strains, 15 from the genus *Halorubrum*, and one each from the genus *Halosarcina, Halosimplex,* and *Halostagnicola.* Repeatability for the other 63 was examined and they were consistent, as in Figure 1 (data are not shown). Repeatability of the technique indicated robustness of the conditions and primers used and provide confidence for estimating variation between strains.

We were interested to know if the random primers can be used as a screening technique. If banding patterns could reliably demonstrate similarity within genera for instance, newly cultured yet unidentified strains could be easily screened and a general taxonomic decision could be made. Therefore, the banding patterns for the 81 total haloarchaeal type strains were assessed using software that produced a dendrogram of the genomic fingerprints. Figure 5-2 is the UPGMA dendrogram determined for the above type strains. Compared to other studies [192, 193], our genome fingerprinting technique offers very little banding pattern complexity. There are two possible reasons - the primers were designed for systems other than the Haloarchaea and adopted for our purposes, and PCR bias, though if it occurs is reproducible (see figure 5-1). Yet, species specific banding patterns observed earlier in Haloarchaea [197] are also observed here; each species appears to have a unique banding pattern. However, there is very little clustering at the genus level. For instance, some species within the same genus have similar banding patterns according to the dendrogram analysis (e.g., *Natrinema ejinorense* and *Natrinema altunense*) but other species from the same genus are found elsewhere (e.g., *Natrinema pellirubrum* and

Natrinema versiforme). This pattern is observed for all the genera for which several species were analyzed (e.g., *Halorubrum*, *Haloferax*). Thus, this DNA fingerprinting should not be used to classify isolates to a genus level. The observed amount of variation displayed among species within the same genus, led to the hypothesis that this technique might also detect genomic variation among strains within the same species. Therefore we tested this fingerprinting technique on several populations of naturally co-occurring closely and distantly related strains.

5.3.2 MLSA on environmental strains

MLSA was performed in order to determine the genetic variation, and the evolutionary relationships of the isolates from Aran-Bidgol Lake. Multiple sequence alignments were constructed from individual locus data from the new isolates and from genome data deposited in the NCBI database of type strains. Concatenated alignments were made from these and then a phylogenetic tree was constructed. The Aran-Bidgol isolates clustered into two main genera; *Halorubrum* and *Haloarcula* (Figure 5-3). Two polytomous groups, A and B, were observed within the genus *Halorubrum* and depicts evidence for distinct phylogroups with low sequence diversity as first seen for Spanish and Algerian isolates [84]. Pairwise comparison of the number of nucleotides different within each of these phylogroups was carried out using MEGA 5 [191]. In both groups A and B, no two isolates had more than 10 nucleotide differences from one another across the concatenation of ~ 2500bp (i.e., <1% sequence divergence; Table 5-4). This also holds true for group C (Table 5-5) within the *Haloarcula* cluster.

5.3.3 Fingerprinting the Aran-Bidgol strains

Genomic fingerprint analysis was run on each of the Aran-Bidgol Lake environmental isolates. Banding patterns for each individual were generated and compared for similarity by dendrogram construction. The fingerprints and resulting dendrogram were then compared to the maximum likelihood tree constructed from the MLSA data (Figure 5-3) for relating genetic and genomic variation within populations. It is noteworthy that despite limited numbers of bands produced for fingerprinting analysis, closely related strains from a single phylogroup displayed numerous variations in banding patterns, many of which were dissimilar to each other as determined by the dendrogram analysis. These widely different banding patterns reflect the variation in individual genomes. Comparison between sequence and banding pattern similarity demonstrates a lot of variation and no discernable patterns of relatedness even between strains that have zero differences across ~2500 nucleotides. Banding patterns of isolates within the genus *Halorubrum* seem as different as the banding patterns of isolates between the genera *Halorubrum* and *Haloarcula.* In some cases identical MLSA haplotypes have identical fingerprint patterns. We believe this can be attributed to the relatively low complexity of fingerprint bands produced, rather than two strains having identical genomes, and in such cases other methods of comparison like genome sequencing might reveal additional differences.

5.4 Discussion

Our study employed DNA sequencing of multiple protein coding loci and random genomic amplification to test for variation in haloarchaeal isolates cultivated from the same location under the same conditions. The concatenated maximum likelihood tree in Figure 5-3, and the number of pairwise nucleotide polymorphisms in Tables 5-4 and 5-5, show that many isolates are closely

related to one another across the five loci and are more or less indistinguishable from each other by these methods. However, the DNA fingerprinting analysis on these same isolates revealed additional variation not captured by MLSA, indicating genomic changes occur faster than the rate of substitution in redundant codon positions. Unfortunately, the deeper branches of the UPGMA hierarchical clustering dendrogram are unreliable for determining relationships and do not provide a good description of the measured Dice coefficients. Yet, shallower branches in the clustering diagram that are a good representation of the banding pattern differences show conflict with the MLSA phylogeny (Figure 5-3). Though the fingerprinting technique did not yield patterns of relatedness at the species level or genus level, it did demonstrate the high probability that the genomes of each isolates are unique. Whether that uniqueness is based on gene content or in genomic arrangements is undeterminable from this analysis.

However, given the known propensity for HGT in Halobacteria [80-86], we surmise the fingerprint banding-pattern differences are largely due to gene transfer events. Discovery of recombinant hybrids [82] and the identification of enormous identical segments shared among the genomes of phylogenetically distant genera [81] indicates the Haloarchaea are subject to immense genomic variability from single gene transfer events. In another study, an influx of 303 transferred genes into *Haloferax mucosum* and *Haloferax mediterranei* were mostly of unknown function with some known transporters [99], which is similar to the types of genes observed in the highly recombinogenic genomic islands of *Haloquadratum waslbyi* [80]. The *H. waslbyi* genome is 47.9% GC, but its genomic islands are GC rich by comparison, and enriched in transposable and repeat elements [198] indicating a role for viruses in generating genomic diversity [80]. Similar to *H. waslbyi*, the genome of *Halobacterium* NRC-1 was interspersed with 91 insertion sequence

elements of diverse GC compositions [87, 199]. Apart from homologous recombination, IS elements have been attributed to inactivating the bacterio-opsin gene in *Halobacterium halobium* [200] and causing genomic rearrangements at AT-rich regions in *Halobacterium* NRC-1 [199]. Moreover, recent analysis indicates these Aran-Bidgol Lake isolates display enormous variation in whole genome content with differences in group A ranging from 0.01Mb up to 0.51Mb and from 0.07Mb up to 0.30Mb in group B [201]. Therefore, we hypothesize the drastic differences in fingerprints observed for the closest relatives (e.g., strains from groups A, B, and C) are more likely due to HGT, possibly mediated by insertion sequence elements [87, 199, 200], tRNAs [82], or other factors, rather than genome rearrangements.

We further suggest that the fingerprint banding patterns, especially for those within groups A, B, and C, were unlikely due to mutational events. Haloarchaea have low rates of spontaneous mutation, having been measured at 1.90×10^{-8} mutational events per cell division [202]. Furthermore, halobacteria are considered to have a high capacity for repairing DNA, as they have demonstrated the ability to survive radiation and desiccation damaged DNA [203, 204], which is probably due to the prevalence of polyploidy through the process of gene conversion [205]. Preliminary *in silico* analysis to determine the binding sites for each of the five primers in *Haloquadratum walsbyi* DSM 16790 and *Halorubrum lacusprofundi* revealed priming mostly in conserved loci, although a few phage related loci were also detected. Because many of the compared strains are very closely related, having only a few (or zero) nucleotide polymorphisms in the ~2500 sequenced base pairs, yet display enormous differences in fingerprint banding patterns, it would be unlikely that a few, or even one of the PCR binding sites in every strain within groups A, B, or C, would be mutated. Therefore, substitutions in PCR primer binding sites seem

unlikely to have played a role in generating all the observed differences in banding patterns, especially those from closely related strains.

Analysis of five housekeeping genes demonstrates the isolates form genetically similar and distinct populations in a single environmental community and yet each genome is apparently different. This observation agrees well with expectations from the distributed genome hypothesis [206]. According to this, the non-core genes available in the pan-genome pool are dispensed uniquely amongst the individual cells of a species. The differences in haloarchaeal genomic banding patterns suggests that in nature populations are made of highly varied individuals rather than clones of a single individual. The number of distinct genotypes observed, most likely due to gene flow, suggests that haloarchaeal cells are acquiring genomic variation within populations at a rate faster than redundant codon position substitutions, and possibly at every replication event. Distribution of the non-core genes within a highly recombining population (defined by MLSA phylogeny) theoretically enables the individual to quickly adapt to new environmental selection conditions, especially virus predation [80] but may also result from random processes like neutral drift [207].

Table 5-1. Degenerate primers used to PCR amplify and sequence the atpB, ef-2, glnA, ppsA and rpoB genes for MLSA

MLSA primer sequence 5'-3'		
Locus	**Forward**	**Reverse**
atpB	tgt aaa acg acg gcc agt aac ggt gag scv ats aac cc	cag gaa aca gct atg act tca ggt cvg trt aca tgt a
Ef-2	tgt aaa acg acg gcc agt atc cgc gct bta yaa stg g	cag gaa aca gct atg act ggt cga tgg wyt cga ahg g
glnA	tgt aaa acg acg gcc agt cag gta cgg gtt aca sga cgg	cag gaa aca gct atg acc ctc gcs ccg aar gac ctc gc
ppsA	tgt aaa acg acg gcc agt ccg cgg tar ccv agc atc gg	cag gaa aca gct atg aca tcg tca ccg acg arg gyg g
rpoB	tgt aaa acg acg gcc agt tcg aag agc cgg acg aca tgg	cag gaa aca gct atg acc ggt cag cac ctg bac cgg ncc

Table 5-2. PCR conditions for each locus

	atpB	**ef-2**	**glnA**	**ppsA**	**rpoB**
water (µl)	11.6	8.2	11.8	7.9	11.9
5x phire reaction buffer (µl)	4.0	4.0	4.0	4.0	4.0
DMSO (µl)	0.6	0	0.4	0.6	0.6
Acetamide (25%)	0	4.0	0	4.0	0
dNTP mix (10mM)	0.4	0.4	0.4	0.4	0.4
forward primer (10mM)	1.0	1.0	1.0	1.0	1.0
reverse primer (10mM)	1.0	1.0	1.0	1.0	1.0
Phire hot start II DNA polymerase (µl)	0.4	0.4	0.4	0.4	0.4
template DNA (20ng μl^{-1})	1.0	1.0	1.0	0.7	0.7
annealing temperature (˚C)	60.0	61.0	69.6	66.0	63.7

Table 5-3. Random primers for genomic fingerprinting.

Primers	
Primer Name	Sequence
P1	5'-CCGCAGCCAA-3'
P2	5'-ACGGGCAGC-3'
OPA-9	5'-GGGTAACGCC-3'
OPA-13	5'-CAGCAGCCAC-3'
FALL-A	5'-ACGCGCCCTG-3'

Table 5-4. Pairwise comparison of number of nucleotide differences within polytomous Groups A and B defined on the maximum likelihood tree.

GROUP A	Ga27					0	7	8	5	7	10	9	6	Ea8	GROUP B
	Ec5	5					7	8	5	7	10	9	6	Ea4p	
	Ec15	8	5					5	2	8	7	6	5	Ea10	
	Ga66	7	8	7					5	9	8	7	4	Hd13	
	Fc2	5	4	5	6					6	9	8	3	Ib24	
	Fa2	1	1	1	0	1					5	4	7	Eb13	
	Fa5	7	8	7	4	6	0					1	8	Ib25	
	Fa17	2	2	2	0	2	0	1					7	Ea1	
	C191	8	9	8	5	7	0	1	1					Ib43	
	Fb21	8	9	8	5	7	0	1	1	0					
	Ga36	4	3	6	7	3	1	7	3	8	8				
	G37	8	3	4	7	5	0	5	1	6	6	6			
	Ga2p	6	5	4	5	3	0	5	1	6	6	4	4		

Table 5-5. Pairwise comparison of number of nucleotide differences within polytomous Group C defined on the maximum likelihood tree. Cells in blue represent members of Group C and cells in black represent the neighboring cluster on the ML tree.

GROUP C	Dc10						
	A14	0					
	Fb5	0	0				
	Dg17	0	0	0			
	Fb19	2	2	2	2		
	Hd14	30	30	30	30	32	
	Cc8	29	29	29	29	31	7
	Hd4	38	38	38	38	39	26	19

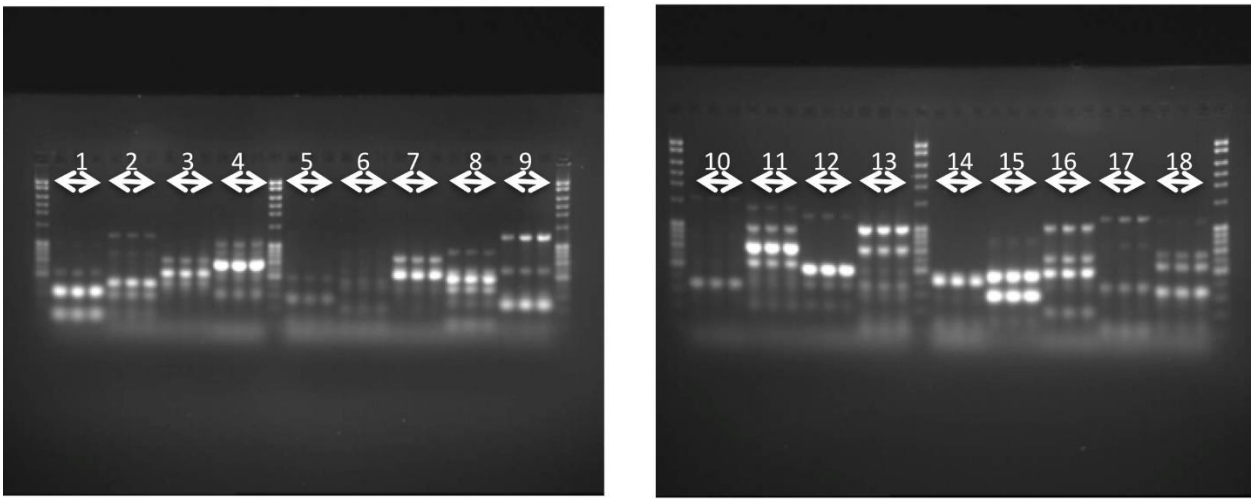


Figure 5-1. Repeatability of the fingerprinting technique. Each number represents a type strain analyzed in triplicate. 1) *Halorubrum arcis* JCM 13916 2) *Halorubrum coriense* DSM 10284 3) *Halorubrum distributum* JCM 9100 4) *Halorubrum ejinorense* JCM 14265 5) *Halorubrum lacusprofundi* ATCC 49239 6) *Halorubrum lipolyticum* DSM 21995 7) *Halorubrum litoreum* JCM 13561 8) *Halorubrum saccharovorum* DSM 1137 9) *Halorubrum sodomense* JCM 8880 10) *Halorubrum tebenquichense* DSM 14210 11) *Halorubrum terrestre* JCM 10247 12) *Halorubrum tibetense* JCM 11889 13) *Halorubrum trapanicum* JCM 10477 14) *Halorubrum vacuolatum* JCM 9060 15) *Halorubrum xinjiangense* JCM 12388 16) *Halosarcina pallida* JCM 14848 17) *Halosimplex carlsbadense* JCM 11222 18) *Halostagnicola larsenii* JCM 13463.

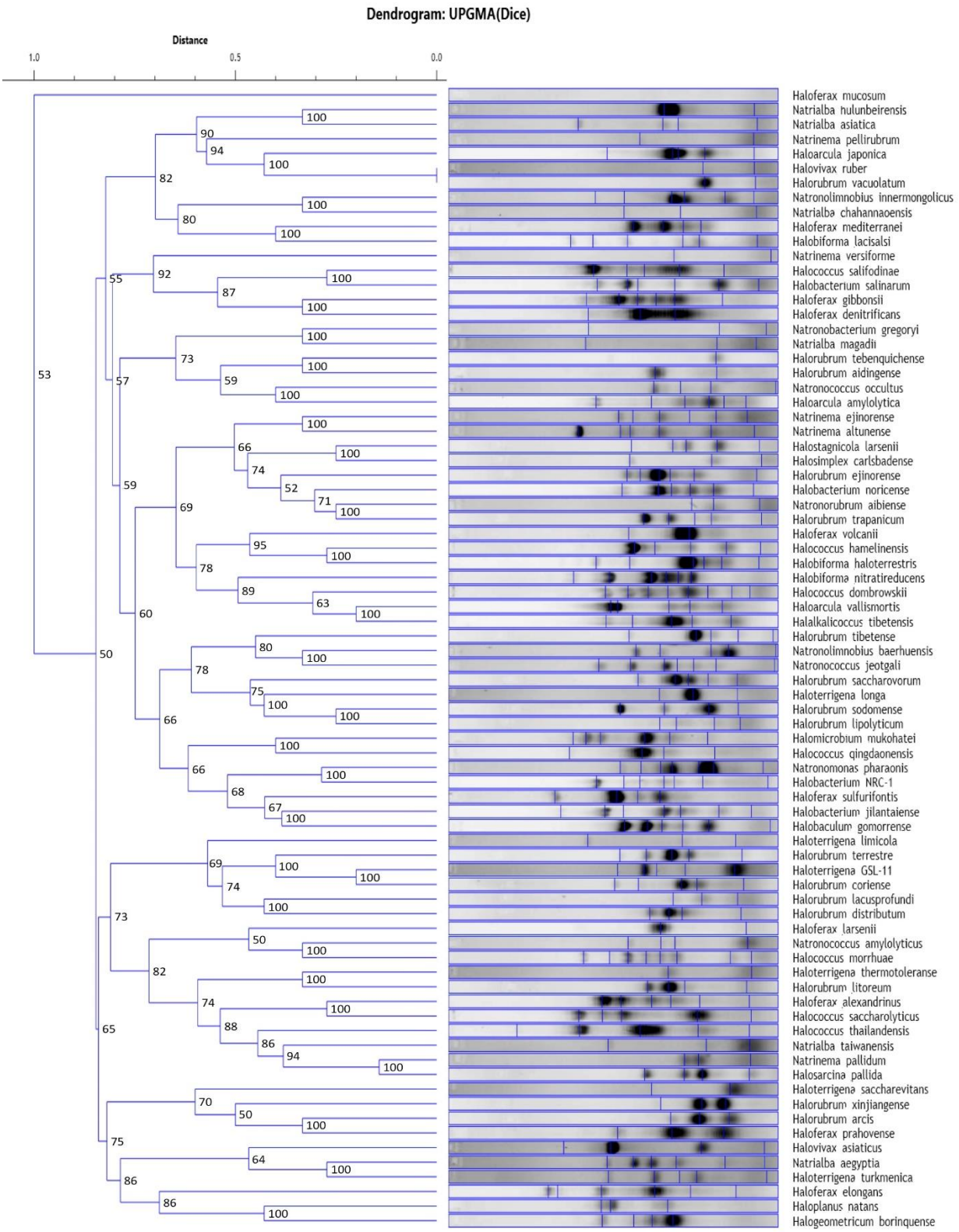


Figure 5-2. UPGMA dendrogram comparing banding patterns between type strains. The numbers displayed at the nodes represent the cophenetic correlation coefficients.

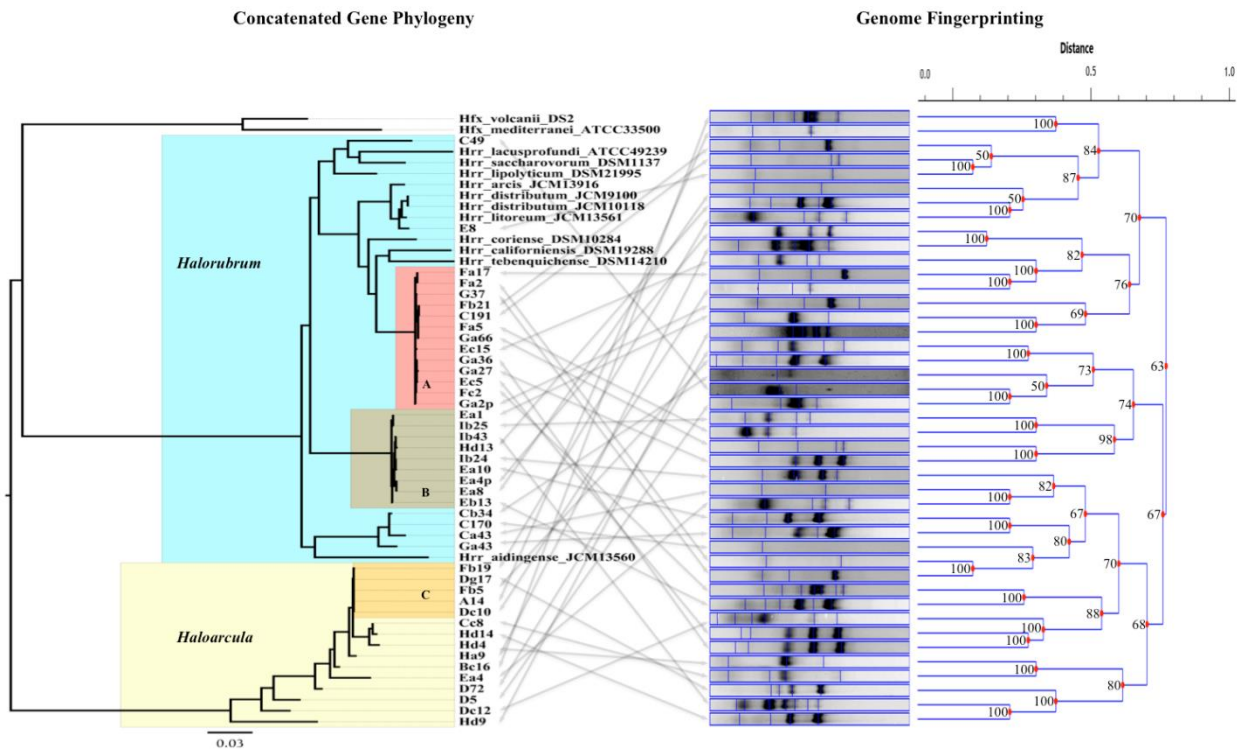


Figure 5-3. MLSA vs Genome fingerprinting. Comparison of the maximum likelihood tree computed from the concatenation of five housekeeping genes, and the UPGMA dendrogram determined from the banding patterns of the genome fingerprinting. Lines between tree and dendrogram connect the same strain in the different analyses.

5.5 Additional evidence

Genome sequence for the isolates listed in table 5-6, 17 from this study and 7 from collaborators, were obtained to various levels of completion. The obtained contigs were rearranged with respect to the genome sequence of Hrr_Fb21 using Mauve version 2.4.0 [208]. The rearranged contigs for each isolate were combined and aligned using progressiveMauve [209]. The genome alignments were compared against the *atpB* gene phylogeny of these isolates.

The *atpB* phylogeny recreates the clusters observed in figure 5-3. Closely related isolates have varying genome arrangements and presence or absence of Locally Collinear Blocks (LCBs) detected by Mauve figure 5-4. These variations are drastic between phylogroups and though not as severe within the phylogroups, many differences exist nonetheless. The findings from the genome sequence comparisons corroborate the results from the whole genome fingerprinting described earlier.

Table 5-6. List of isolates with genomes sequenced.

Isolate
Hrr_C49
Hrr_Ea1
Hrr_Eb13
Hrr_Ib24
Hrr_Hd13
Hrr_Ea8
Hrr_Ga36
Hrr_Fb21
Hrr_LG1
Hrr_LD3
12-10-3
Hrr_G37
Hrr_Ga2p
Hrr_Ec15
Hrr_Sp3
Hrr_Sp9
Hrr_C2
Hrr_167
Hrr_E8
Hrr_Sp5
Hrr_Sp7
ASP1
ASP121
ASP200

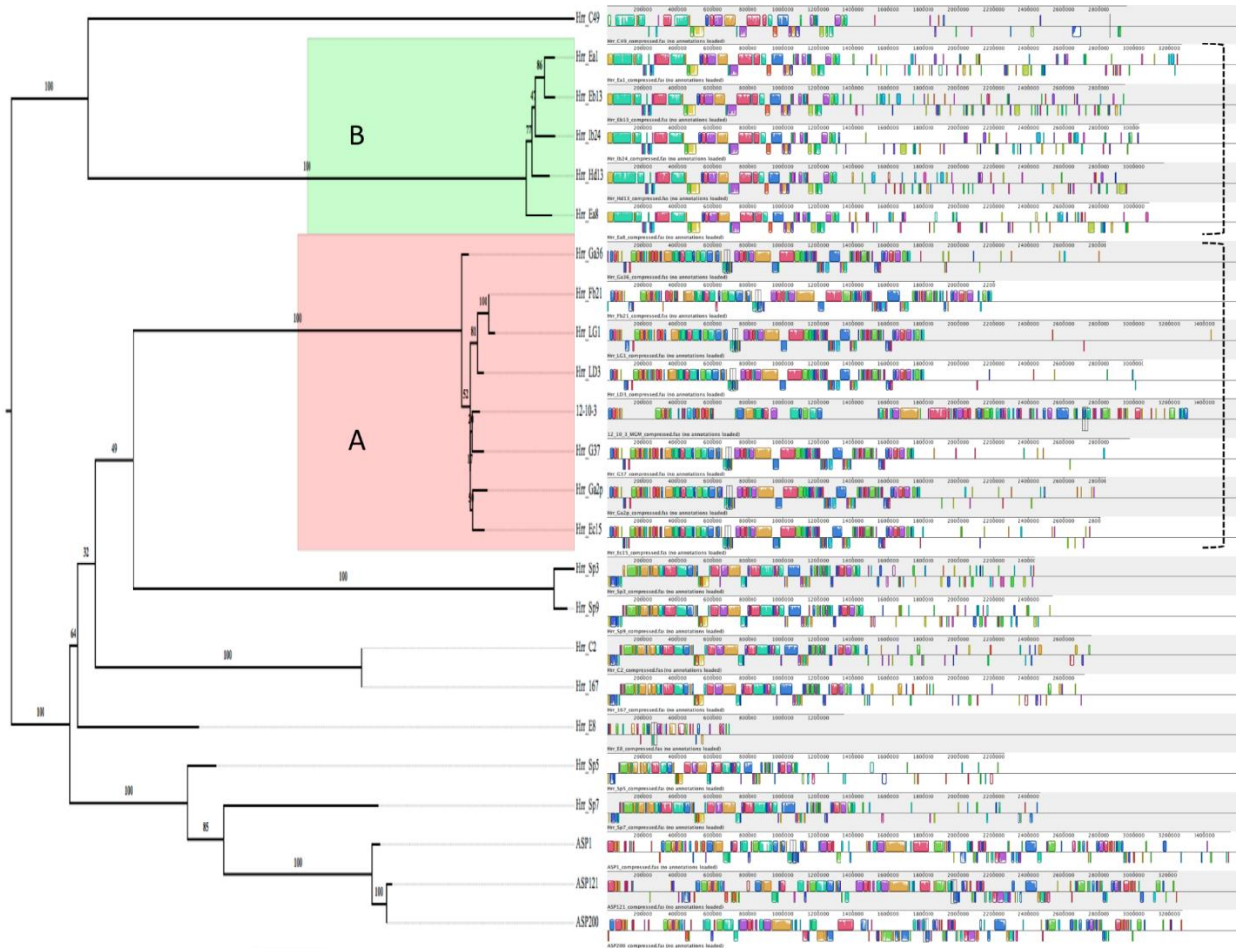


Figure 5-4. Comparison of *atpB* gene phylogeny and genome alignments. ML tree on the left and Mauve alignment of the genes on the right. Two phylogroups A and B on the ML tree similar to what is described from MLSA. Each block on the genome alignment represents an LCB which denotes a conserved segment that is free of internal genome rearrangements.

Chapter 6 : Extensive intragenic recombination in the highly conserved haloarchaeal 16S rRNA and *rpoB* genes

6.1 Abstract

The 16S rRNA and the ß subunit of the RNA polymerase (rpoB) are widely employed as markers of choice for diversity studies, taxonomy, and phylogenetic analyses in the Halobacteriales. However, evidence for the existence of multiple divergent copies of the 16S rRNA gene in several genera (e.g., *Haloarcula, Natrinema, Halomicrobium, and Halosimplex*) suggests the gene is transferred between species and genera. These extra copies are genetically distant from the others in the genome, with divergent copies of the rRNA operons differing by up to ~7% of the nucleotide positions in some strains. In addition, the Haloarchaea have been demonstrated to undergo recombination at high frequencies with the 16S rRNA and rpoB gene being transferred between diverged species, including the discovery of identical or nearly identical 16S rRNA genes shared between separate phylogenetically defined groups, or species. In this study, we surveyed the 16S rRNA and rpoB genes from 109 haloarchaeal genomes to examine the extent of recombination within these highly conserved genes. Comparison of phylogenetic topologies derived from the two genes revealed distinct bipartitioning, indicating both genes experienced different evolutionary histories within the Haloarchaea. Moreover, phylogenetic reconstruction of gene evolutionary history using the two halves of the gene indicated that each was distinct and divergent from that of the full-length gene indicating independent evolutionary histories for separate gene segments within genes. Greater incidences of chimeric sequences were observed in the 16S rRNA than *rpoB*, at a higher frequency between closely related species with every site being subject to recombination, some more so than others. Maximum likelihood mapping of each gene revealed that *rpoB* provided fewer discordant quartets compared to the 16S

rRNA, however, quartet puzzling as well as the Phi test corroborate that these haloarchaeal genes are not exempt from recombination. The Haloarchaea undergo extensive recombination, often giving rise to stable chimeras that result in fuzziness in the taxonomic delineation of this entire class. Our results extend those observations and indicate that the highly conserved and most prominently utilized markers for taxonomic and evolutionary studies in the Halobacteriales have extensive intragenic recombination indicating a high probability for misclassification when used exclusively for taxonomic purposes.

6.2 Materials and Methods

6.2.1 Gene sequence acquisition and alignments

The 16S rRNA and *rpoB* gene sequences were extracted from the 109 available Haloarchaeal genomes that are listed in table 6-1 by using BLAST [154] (http://blast.ncbi.nlm.nih.gov/Blast.cgi). 109 full length 16S rRNA and *rpoB* sequences from Top BLAST hits were aligned using MUSCLE [148] and alignments were manually edited using MACCLADE 4.08 [149]. The alignments for each gene were further split halfway the length of the gene to result in the alignments of the first half and second half of each gene.

6.2.2 Phylogenetic reconstruction and tree topology comparison

Maximum likelihood phylogenetic trees were constructed for the 16S rRNA and *rpoB* genes, and the partial 16S and *rpoB* segments from distances calculated under the Generalized Time-Reversible model [210] within the PHYML 3.0 [174] phylogenetic program. Inferred phylogenies were compared in a pairwise manner using R packages APE [211] and distory [212] to determine the topological distance between two measured as a branch length score and to

identify unique branching in each tree respectively. The distance returned reflects the square root of the sum of the squared differences of the internal branch lengths that define similar bipartitions in the trees being compared [213].

6.2.3 Quartet puzzling

Topology of the aligned full length gene sequences were further tested using the quartet puzzling algorithm [214] included within TREE-PUZZLE [215], a phylogenetic reconstruction program. Briefly, parameters for substitution process and rate variation were estimated from quartet sampling and Neighbor Joining tree respectively using the HKY model of substitution [216]. First, likelihood mapping analysis was performed on 5563251 quartets without clustering the taxa to determine the overall phylogenetic signal provided by each gene. Second, the taxa were clustered into four groups representing the four major clades formed by the 16S rRNA phylogenetic tree and likelihood mapping was performed on 384930 to obtain evidence for recombination.

6.2.4 Estimation of recombination

The extent of intragenic recombination in both full length genes was estimated two ways. First, statistically significant evidence for recombination events within the dataset was estimated by implementing the phi test [217] within the Splitstree program [218]. Second, the recombination detection program (RDP4) [219] was used to identify putative recombination events. RDP4 houses a suite of algorithms to determine recombination including the RDP method [220], GENECONV method [221], Bootscan/Recscan method [222], MaxChi [223], Chimaera [224], SiScan [225], and 3Seq [226] . Each full length gene dataset was analyzed for recombination by

all the above algorithms with 100 bootstrap replicates. All putative recombination events detected were compiled for further estimation of overall recombination events across the length of the sequence and to identify recombination hotspots. Rate of recombination was determined by running 10000000 MCMC updates.

6.3 Results

6.3.1 Comparison of phylogenies

Assessing the midpoint rooted maximum likelihood trees derived for both the 16S rRNA and *rpoB* shows discrepancies in the phylogenetic (figure 6-1). Red lines in the tree depict branching that is unique to each tree in the comparison as determined by the R package distory. The 16S rRNA and *rpoB* gene phylogenies portray variations in relationships between closely related taxa as well as in deeper branches with an overall distance of 0.824. Appraisal of the partial gene fragment topologies to one another and to the complete gene provides further interesting insights. 16S_a (bases 1 to 742) and 16S_b (bases 743 to 1484) topologies are distinct, distance of 0.7645159, from one another (figure 6-2) suggesting that the two halves of the 16S rRNA gene have different evolutionary histories and result in discrepancies in clustering of taxa, there are 79 crisscrossing connections between the two trees. The topology for both 16S_a and 16S_b are approximately equidistant from that of the full length 16S rRNA gene with a distance of 0.5553157 and 0.5954106 respectively. The full length 16S rRNA gene provides an average phylogenetic signal in the tree reconstruction. This is the case with the full length *rpoB* gene as well. The full length topology is approximately equidistant from that of the two halves rpoB_a and rpoB_b, 0.6050257 and 0.5403120 respectively. However, unlike with 16S_a and 16S_b, topologies of the two *rpoB* halves, from bases 1-915 (*rpoB*_a), and 916-1830 (*rpoB*_b), are closer to each other than

the full length gene (0.4385393) (figure 6-3) and depict fewer crisscrossing connections between the taxa on the two trees than the two halves of the 16S rRNA gene.

6.3.2 Maximum likelihood maps of the two genes

Each possible quartet in the phylogenetic tree for the 16S rRNA and *rpoB* were assayed. The vertices of the triangle were equally loaded in both cases, however, the 16S rRNA returned a greater percentage of unresolved quartets out of the 5563251 combinations, ~ 4.5% partially resolved and ~3.1% completely unresolved. *rpoB* returned smaller percentages of both partial and unresolved quartets, ~1.8% and ~1.5% respectively.

Quartet puzzling was redone with predetermined groups of taxa clustered based on the four major clades in the 16S rRNA phylogeny (figure 6-1) to determine inter-clade transfer events as predicted by the phylogenetic signal. Maximum likelihood maps were constructed for each gene by assaying 384930 quartets (figures 6-4 and 6-5). Once the taxa were pre-classified into groups, both 16S rRNA (figure 6-4) and *rpoB* (figure 6-5) showed better grouping of quartets into one of the three configurations. Neither gene had any unresolved quartets. The proportion of partially resolved quartets diminishes as well. However, neither gene topology is 100% supported. There is evidence for recombination between the deep branching groups defined as 6.9% of the resolved 16S rRNA quartets do not conform to the predominant phylogenetic signal. This is observed with the *rpoB* as well but only a smaller portion, ~3.1%, of the quartets are discordant.

Since there was evidence for recombination between the deep branching groups, quartet puzzling was rerun on each group by further dividing the taxa into four subgroups. Figures 6-6 through 6-9 represent the ML maps for groups A, B, C, and D of the 16S rRNA phylogeny. There are greater percentages of discordant quartets within groups A, B, and D than observed for between these groups suggesting much more recombination within groups than between. Group C is an

outlier in that there were no discordant quartets, 100% of the quartets tested supported one quartet configuration. Similar results were observed in the *rpoB* groups as well. There is a lot of recombination within groups A, B, and D but none in group C (figures 6-10 through 6-13). Group C is made up of way fewer taxa than the other groups and could show more recombination if it were bigger.

6.3.3 Extent of recombination

6.3.3.1 16S rRNA

Recombination within the 16S rRNA is evident from the discrepancies in the phylogeny. Computing the phi test on the dataset further corroborated this returning a statistically significant value ($p = 0.03925$). Apart from this, detecting recombinant sequences and estimation of the rate of recombination (rho) carried out by RDP4 provided validation for extensive recombination within this gene. 37 recombination events were detected by at least one algorithm out of which 24 were supported by more than one algorithm. The segment of the sequence that is recombinant, as well as the major and minor parents for each of the recombinant is listed in table 6-2. Various sizes of recombinant fragments are observed, ranging from 20 bps to 707 bps, within the 16S rRNA in Haloarchaea. In some cases, there is evidence for multiple recombination events within a single gene sequence. The 16S rRNA of the following Haloarchaea were subject to two recombination events at separate points on the gene - *Halobaculum gomorrense* strain JCM 9908, *Halopiger xanaduensis* strain: JCM 14033, *Haloplanus natans* strain: JCM 14081, and *Halorussus rarus* strain: JCM 16429. Figure 6-14a is the predicted secondary structure of the 16S rRNA depicting the conservation in the positons along the length of the gene as well as the covariation between the paired bases. Figure 6-14b plots all the recombination events along the length of the gene, the size

of the recombinants, and the –log(p-value) for each event. Recombination between the sequences is evident irrespective of the structure in the rRNA, be it conserved stems or variable loops. The number of recombination between closely related taxa is greater than those with more distantly related taxa as seen by the red and blue lines respectively. Recombination does not seem to be constrained by position on the gene, however the rate of recombination varies by site as seen in figure 6-15a. The average rho and mutation (theta) per site were estimated to be 5.211×10^{-2} and 3.826×10^{-2} respectively, that is, the rate of recombination is ~ 1.362x what the mutation rate is per site on the 16S rRNA. The greatest rates of recombination are realized between positions 933 and 953 which is within the V-5 region.

6.3.3.2 *rpoB*

Analyzing the *rpoB* sequences revealed similar results as the 16S rRNA. Phi test for recombination returned a statistically significant p value ($8.329e^{-4}$) and RDP4 detected 46 recombination events determined by one algorithm. Out of the 46 identified recombination events, only three were supported by more than one algorithm (see table 6-3). The three recombination events supported by more than one algorithm occur in different taxa and there is no evidence for multiple recombination events within one gene. As with the 16S rRNA, varying lengths of recombinant sequences are observed in *rpoB*. These range from 104 bps to 1644 bps in length. Figure 6-16 displays all the recombination events across the *rpoB* gene sequence. Recombination events in *rpoB* seem more frequent between closely related taxa than distantly related ones, as represented by the red and blue lines respectively, similar to 16S. The recombination rate plot (figure 6-15b) shows that the rate is almost even for a major portion of the gene. However, the rate increases ~4 fold between positions 845 and 1212 on the gene. Average theta per site calculates is

$4.361x10^{-2}$ and average rho per site is $9.717x10^{-1}$. The rate of recombination in *rpoB* is ~22.283 times the rate of mutation.

6.4 Discussion

This study measured recombination, evidence for its presence, and its effect on evolutionary histories of the 16S rRNA and *rpoB* genes from 109 available Haloarchaeal genomes. Employing the Phi test for recombination with each dataset showed that neither gene was exempt from recombination in the Haloarchaea. Other analyses corroborate this finding. Comparisons of the clustering on taxa on the 16S_a and 16S_b phylogenetic trees (figure 6-2) shows extensive within clade rearrangements and evidence for transfers between distant clades as well. This is also apparent from the maximum likelihood maps (figures 6-4 and 6-6 through 6-9) where discordant quartets are observed when quartet puzzling is carried out after clustering taxa into four major groups. Many recombination events were also detected in both genes by RDP4, across the length of the entire genes and in sizes ranging between 20 and 707 nucleotides in the 16S rRNA gene (figure 6-14b), and from 104 to 1644 nucleotides in the *rpoB* gene (figure 6-16). Our findings support the hypothesis that recombination between species must attribute to the presence of divergent copies of the 16S rRNA [89] causing the intragenic heterogeneity observed in many Haloarchaeal species [89, 112-116].

Given that the Haloarchaea are highly promiscuous [80-86, 99-101], it is not surprising to see that every genus of the Haloarchaea depicts evidence for recombination in their 16S rRNA gene (see table 6-2). It was hypothesized that the rate of genome variation in individuals in a population was greater than accruement of mutations at the third codon position since isolates that were identical across five loci had varying whole genome patterns [100]. Our findings support

this hypothesis. The estimated average rate of recombination per site is much greater than the rate of mutation for each gene. While the rate of recombination is ~ 1.362x mutation rate per site in the 16S rRNA gene, the same in *rpoB* is ~22.283x rate of mutation per site. The drastic difference in rates between the two genes can be attributed to the maintenance of protein function. It is shown that recombination between structurally related proteins increases the probability of preserving function over random mutation [227]. The recombination events detected in the two genes are also unrestricted by the size and location on the gene, and slightly hindered by the sequence divergence between the recombining sequences. These gene-centric results tie well with findings from whole genome comparisons [81, 82, 85, 99] and leads us to conclude that the promiscuity of Haloarchaea highly impact individuals at the individual gene level.

Comparing the 16S rRNA and *rpoB* genes, both show varied evolutionary patterns of the taxa analyzed (figure 6-1). This discrepancy in the evolutionary history of the Haloarchaea is not restricted to comparison between the two genes. Similar results were observed when the same analyses were done with partial segments of each gene. Each half of the two genes displayed marked differences in their evolutionary histories. This is more apparent in the 16S rRNA (figure 6-2) than *rpoB* (figure 6-3), the two halves of the 16S rRNA share a more divergent evolutionary history than those of *rpoB* as measured from the topological distances of 0.7645159 and 0.4385393 respectively. Phylogenetically, *rpoB* provided better signal than the 16S rRNA gene for the Haloarchaea. Quartet puzzling pre and post clustering of taxa into groups showed that *rpoB* (figure 6-5) returned fewer unresolved, partially resolved, and discordant quartets than the 16S rRNA gene (figure 6-4). The greater percentage of discordant quartets in the 16S rRNA gene must reflect the larger number or recombination events detected in the gene that were supported by more than one algorithm. Under these constraints, only three recombination events were identified in *rpoB*

whereas twenty four events were identified in the 16S rRNA (figure 6-14b) and in some cases, there were multiple events in one gene.

Despite the presence of multiple, frequently divergent, copies of the 16S rRNA in the genome, it has been used extensively as a gold standard gene to study taxonomy and diversity of bacteria and archaea in the environment. In the Haloarchaea, the 16S rRNA gene is subject to frequent recombination that sometimes leads to the formation of stable chimeras that can blur the taxonomic delineation in this class of organisms. The single copy *rpoB* gene is proven to be advantageous over the 16S rRNA gene for such purposes in many systems [103, 117-128], and this study shows that there is also evidence for fewer recombination events in the *rpoB* gene of the Haloarchaea thereby reducing the fraction of discordant quartets surveying the entire class. However, both genes offer a high probability for misclassification as a result of the extensive recombination, when used exclusively for taxonomic purposes.

Table 6-1. List of 109 Haloarchaeal genomes analyzed in this study.

Haladaptatus cibarius	*Haloferax larsenii* strain JCM 13917	*Halorubrum tibetense* strain JCM 11889
Haladaptatus litoreus	*Haloferax lucentense* strain JCM 9276	*Halorubrum trapanicum* strain JCM 10477
Haladaptatus paucihalophilus strain DX253	*Haloferax mediterranei* strain JCM 8866	*Halorubrum vacuolatum* strain JCM 9060
Halalkalicoccus jeotgali strain B3	*Haloferax mucosum* strain JCM 14792	*Halorubrum xinjiangense* strain JCM 12388
Halalkalicoccus tibetensis strain JCM 11890	*Haloferax prahovense* strain JCM 13924	*Halorussus rarus* strain JCM 16429
Halapricum salinum strain CBA1105	*Haloferax sulfurifontis* strain JCM 12327	*Halosimplex carlsbadense* strain JCM 11222
Halarchaeum acidiphilum strain JCM 16109	*Haloferax volcani*i strain JCM 8879	*Halostagnicola larsenii* strain JCM 13463
Halarchaeum nitratireducens strain MH1-136-2	*Halogeometricum borinquense* strain JCM 10706	*Haloterrigena limicola* strain JCM 13563
Halarchaeum rubridurum	*Halomicroarcula pellucida* strain BNERC31	*Haloterrigena longa* strain JCM 13562
Halarchaeum salinum strain MH1-34-1	*Halomicrobium mukohatae*i strain JCM 9738	*Haloterrigena saccharevitans* strain JCM 12889
Haloarcula amylolytica strain BD-3	*Halopiger xanaduensis* strain JCM 14033	*Haloterrigena thermotolerans* strain JCM 11050
Haloarcula argentinensis strain JCM 9737	*Haloplanus natans* strain JCM 14081	*Haloterrigena turkmenica* strain JCM 9101
Haloarcula californiae	*Haloquadratum walsbyi* C23	*Halovivax asiaticus* strain JCM 14624
*Haloarcula hispanic*a strain JCM 8911	*Haloquadratum walsby*i DSM 16790	*Halovivax ruber* strain JCM 13892
Haloarcula japonica strain JCM 7785	*Halorhabdus tiamatea* strain JCM 14471	*Natrialba aegyptia* strain JCM 11194
Haloarcula quadrata strain JCM11048	*Halorhabdus utahensis* strain JCM 11049	*Natrialba asiatica* strain JCM 9576
Haloarcula sinaiiensis strain JCM 8862	*Halorubrum aidingense* strain JCM 13560	*Natrialba chahannaoensis* strain JCM 10990

Haloarcula vallismortis strain JCM 8877	*Halorubrum alkaliphilum* strain JCM 12358	*Natrialba hulunbeirensis* strain JCM 10989
Halobacterium jilantaiense strain JCM 13558	*Halorubrum aquaticum* strain JCM 14031	*Natrialba magadii* strain JCM 8861
Halobacterium noricense strain JCM 15102	*Halorubrum arcis* strain JCM 13916	*Natrialba taiwanensis* strain JCM 9577
Halobacterium salinarum	*Halorubrum californiense* strain JCM 14715	*Natrinema altunense* strain JCM 12890
Halobaculum gomorrense strain JCM 9908	*Halorubrum chaoviator* strain DSM 19316	*Natrinema ejinorense* strain JCM 13890
Halobellus clavatus strain JCM 16424	*Halorubrum cibi* strain JCM 15757	*Natrinema pallidum* strain JCM 8980
Halobiforma haloterrestris strain JCM 11627	*Halorubrum coriense* strain JCM 9275	*Natrinema pellirubrum* strain JCM 10476
Halobiforma lacisalsi strain JCM 12983	*Halorubrum distributum* strain JCM 9100	*Natrinema versiforme* strain JCM 10478
Halobiforma nitratireducens strain JCM 10879	*Halorubrum ejinorense* strain JCM 14265	*Natronobacterium gregoryi* strain JCM 8860
Halococcus dombrowskii strain JCM 12289	*Halorubrum ezzemoulense* strain CECT 7099	*Natronococcus amylolyticus* strain JCM 9655
Halococcus hamelinensis strain JCM 12892	*Halorubrum kocurii* strain BG-1	*Natronococcus jeotgali* strain JCM 14583
Halococcus morrhuae strain NRC 16008	*Halorubrum lacusprofundi* strain JCM 8891	*Natronococcus occultus* strain JCM 8859
Halococcus qingdaogense	*Halorubrum lipolyticum* strain JCM 13559	*Natronolimnobius baerhuensis* strain JCM 12253
Halococcus saccharolyticus strain JCM 8878	*Halorubrum litoreum* strain JCM 13561	*Natronolimnobius innermongolicus* strain JCM 12255
Halococcus salifodinae strain JCM 9578	*Halorubrum luteum* strain CECT 7303	*Natronorubrum aibiense* strain JCM 13488
Halococcus sp. 197A	*Halorubrum orientale* strain CECT 7145	*Natronorubrum bangense* strain JCM 10635
Halococcus thailandensis strain JCM 13552	*Halorubrum saccharovorum* strain JCM 8865	*Natronorubrum sulfidifaciens* strain JCM 14089

Haloferax alexandrinus strain JCM 10717	*Halorubrum sodomense* strain ATCC 33755	*Natronorubrum tibetense* strain JCM 10636
Haloferax denitrificans strain JCM 8864	*Halorubrum tebenquichense* strain JCM 12290	
Haloferax gibbonsii strain JCM 8863	*Halorubrum terrestre* strain JCM 10247	

Table 6-2. Recombinant sequences identified in the 16S rRNA gene by RDP4 using more than one algorithm. The table lists the position of the recombination event in the sequence and the major and minor parents deriving the recombinant sequence. * after the nucleotide position represents the approximate position on the sequence. * after the species name represents the closest known species that is a part of the recombinant sequence.

In Recombinant Sequence				
Begin	End	Recombinant Sequence(s)	Minor Parental Sequence(s)	Major Parental Sequence(s)
178	407	*Halarchaeum acidiphilum* strain JCM 16109	*Halorussus rarus* strain JCM 16429*	*Halapricum salinum* strain CBA1105
511	816	*Haloarcula amylolytica* strain BD-3	*Halobellus clavatus* strain JCM 16424*	*Haloarcula sinaiiensis* strain JCM 8862
1064*	1239	*Haloarcula sinaiiensis* strain JCM 8862	*Natrialba magadii* strain JCM 8861	*Halorhabdus utahensis* strain JCM 11049
48*	366	*Haloarcula vallismortis* strain JCM 8877	*Halorubrum luteum* strain CECT 7303	*Halobacterium noricense* strain JCM 15102
940	978	*Halobaculum gomorrense* strain JCM 9908	*Halobacterium salinarum**	*Halorubrum vacuolatum* strain JCM 9060
1180	1393*	*Halobaculum gomorrense* strain JCM 9908	*Halorubrum xinjiangense* strain JCM 12388	*Halalkalicoccus tibetensis* strain JCM 11890
547	567	*Halococcus qingdaogense*	*Halorubrum lacusprofundi* strain JCM 8891	*Natronorubrum bangense* strain JCM 10635
625*	830	*Haloferax lucentense* strain JCM 9276	*Halomicroarcula pellucida* strain BNERC31	*Halorubrum kocurii* strain BG-1

459	1166	*Halogeometricum borinquense* strain JCM 10706	*Halorubrum xinjiangense* strain JCM 12388	*Halalkalicoccus tibetensis* strain JCM 11890
404	564	*Halopiger xanaduensis* strain JCM 14033	*Halorhabdus utahensis* strain JCM 11049	*Natronorubrum aibiense* strain JCM 13488
937	1015	*Halopiger xanaduensis* strain JCM 14033	*Halococcus qingdaogense**	*Natrialba aegyptia* strain JCM 11194
702	935	*Haloplanus natans* strain JCM 14081	*Halorubrum litoreum* strain JCM 13561	*Haloferax lucentense* strain JCM 9276
210	701*	*Haloplanus natans* strain JCM 14081	*Halovivax ruber* strain JCM 13892	*Haladaptatus paucihalophilus* strain DX253*
410*	587	*Haloquadratum walsbyi* C23	*Halorubrum alkaliphilum* strain JCM 12358*	*Haloferax lucentense* strain JCM 9276
885	1286*	*Halorhabdus tiamatea* strain JCM 14471	*Halomicrobium mukohataei* strain JCM 9738	*Haloterrigena saccharevitans* strain JCM 12889*
290	458	*Halorubrum ejinorense* strain JCM 14265	*Halorubrum vacuolatum* strain JCM 9060*	*Halorubrum lacusprofundi* strain JCM 8891
755	1233*	*Halorussus rarus* strain JCM 16429	*Halapricum salinum* strain CBA1105*	*Halococcus saccharolyticus* strain JCM 8878
406	495	*Halorussus rarus* strain JCM 16429	*Halorubrum lacusprofundi* strain JCM 8891	*Halalkalicoccus jeotgali* strain B3
1287	419*	*Haloterrigena turkmenica* strain JCM 9101	*Haloarcula vallismortis* strain JCM 8877	*Halobiforma lacisalsi* strain JCM 12983
936	1016	*Natrialba magadii* strain JCM 8861	*Halococcus saccharolyticus* strain JCM 8878	*Halopiger xanaduensis* strain JCM 14033
159	246	*Natrinema altunense* strain JCM 12890	*Natrialba chahannaoensis* strain JCM 10990	*Halapricum salinum* strain CBA1105*

936	1064	*Natronococcus jeotgali* strain JCM 14583	*Halorussus rarus* strain JCM 16429	*Haloterrigena longa* strain JCM 13562
1015	1258*	*Natronolimnobius baerhuensis* strain JCM 12253	*Halobiforma lacisalsi* strain JCM 12983*	*Halopiger xanaduensis* strain JCM 14033
586	776	*Natronorubrum sulfidifaciens* strain JCM 14089	*Halobiforma lacisalsi* strain JCM 12983*	*Natronorubrum bangense* strain JCM 10635

Table 6-3. Recombinant sequences identified in the rpoB gene by RDP4 using more than one algorithm. The table lists the position of the recombination event in the sequence and the major and minor parents deriving the recombinant sequence. * after the nucleotide position represents the approximate position on the sequence. * after the species name represents the closest known species that is a part of the recombinant sequence.

In Recombinant Sequence				
Begin	End	Recombinant Sequence(s)	Minor Parental Sequence(s)	Major Parental Sequence(s)
12*	1656	*Natronorubrum bangense* strain JCM 10635	*Natronorubrum sulfidifaciens* strain JCM 14089	*Halococcus morrhuae* strain NRC 16008
204*	455	*Haloarcula quadrata* strain JCM11048	*Haloarcula argentinensis* strain JCM 9737	*Haloarcula californiae*
6*	110	*Halopiger xanaduensis* strain JCM 14033	*Haladaptatus paucihalophilus* strain DX253	*Halalkalicoccus tibetensis* strain JCM 11890

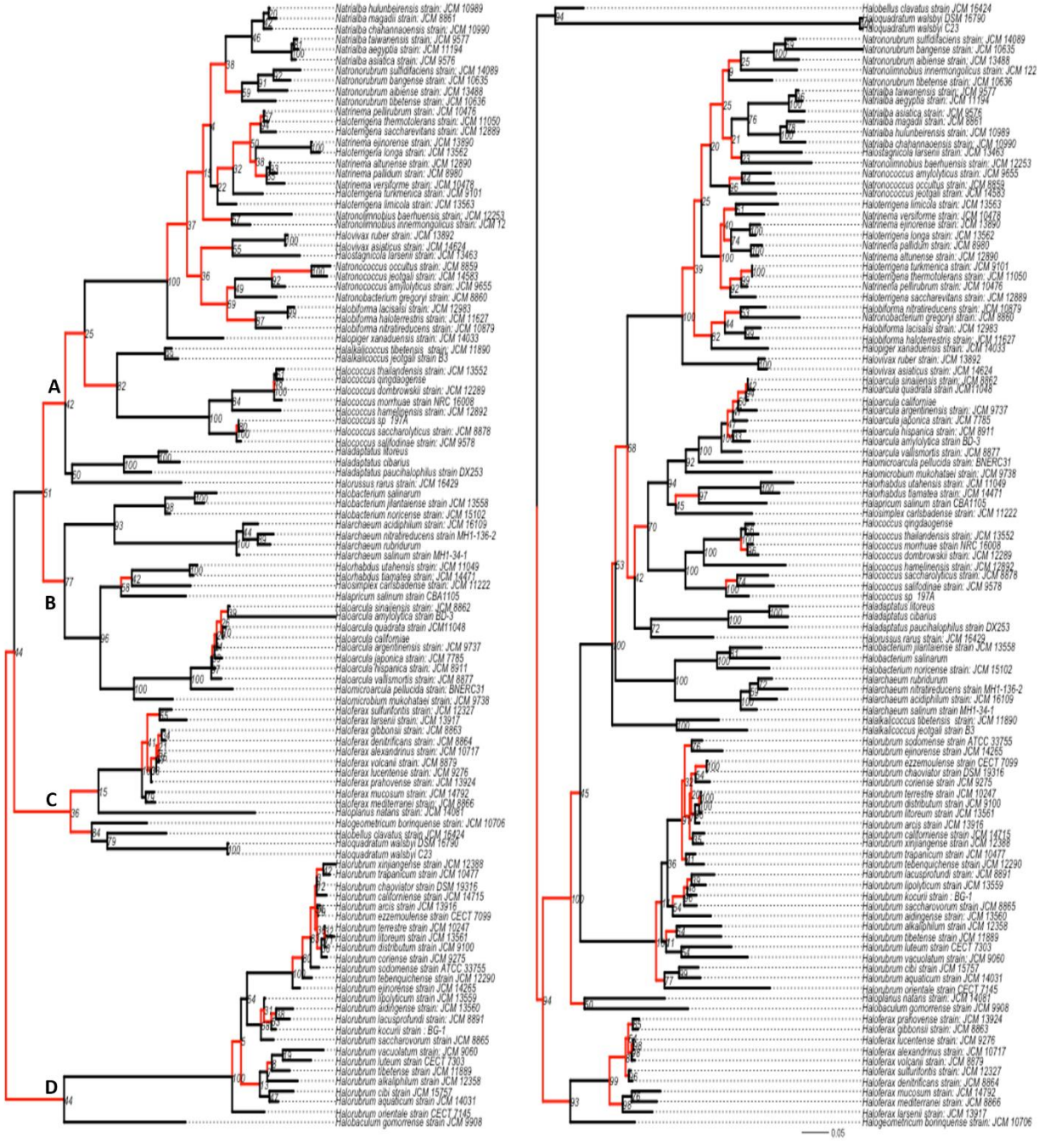


Figure 6-1. Maximum Likelihood (ML) trees constructed for the full length 16S rRNA (left) and rpoB (right) gene sequences using PHYML 3.0 phylogenetic program. Comparison of the topologies using R packages: ape and distory. Black lines represent branches and bipartitioning common to both trees and red lines represent the same that are unique to each tree. ML trees for both full length gene sequences depict distinct evolutionary patterns.

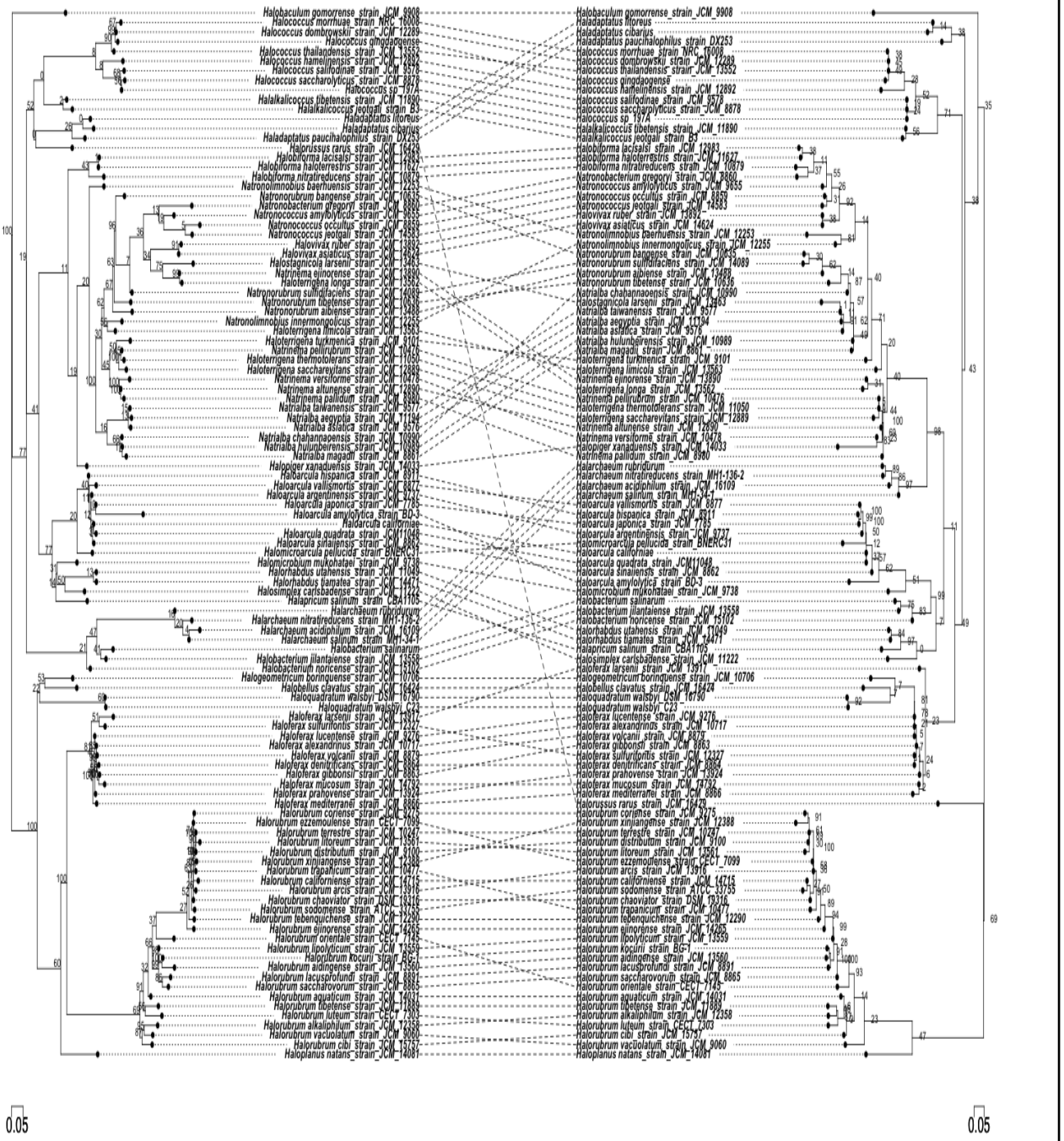


Figure 6-2. ML trees constructed from partial 16S rRNA gene fragments 16S_a and 16S_b. ML tree (left) constructed with the first half of the gene from bases 1 to 742. ML tree (right) constructed from second half of the gene from bases 743 to 1484. Dashed lines in the center connect the same taxa on the two trees. Evidence for within and between clade clustering of taxa on the two trees.

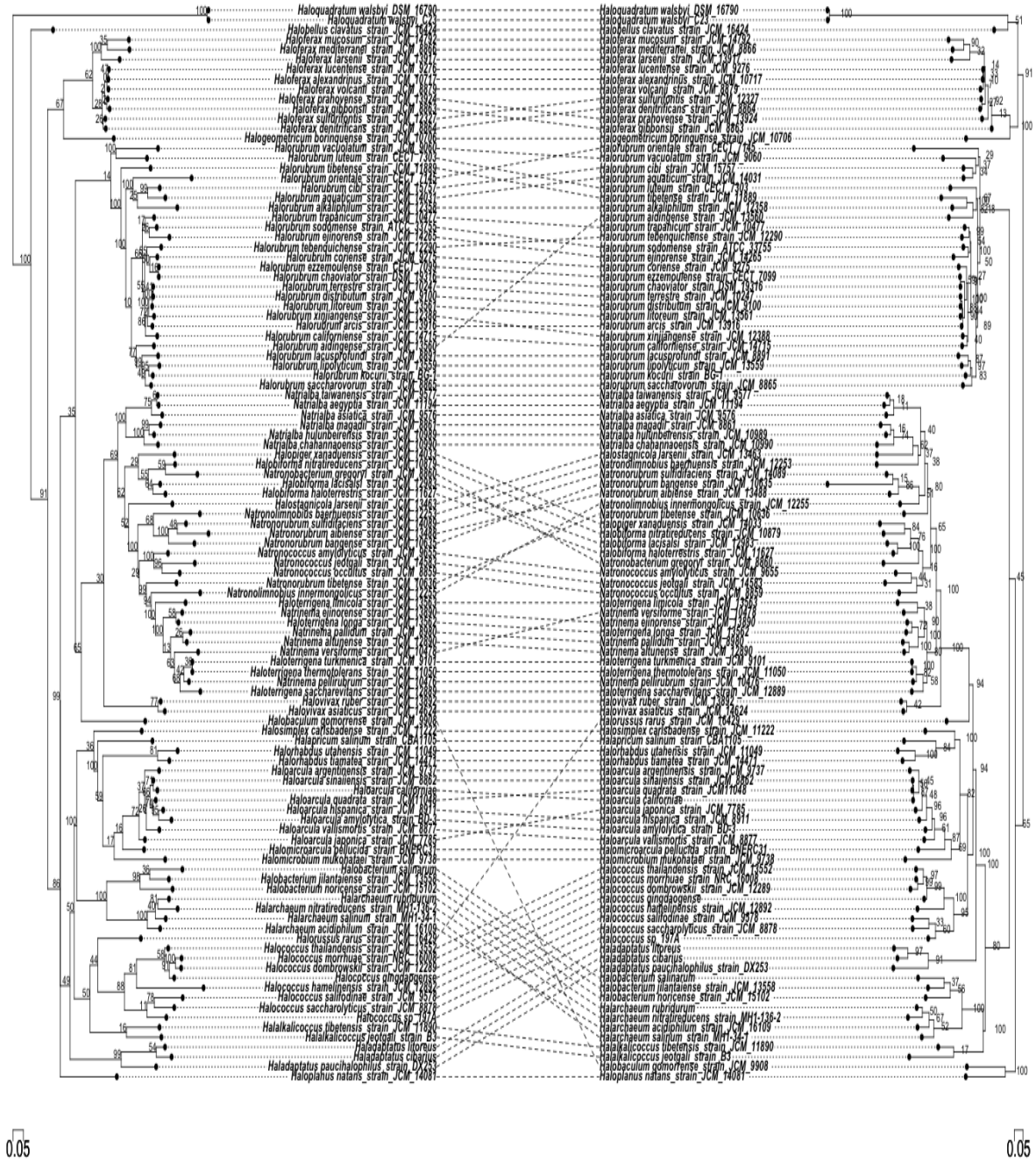


Figure 6-3. ML trees constructed from partial rpoB gene fragments *rpoB* _a and *rpoB* _b. ML tree (left) constructed with the first half of the gene from bases 1 to 915. ML tree (right) constructed from second half of the gene from bases 916 to 1830. Dashed lines in the center connect the same taxa on the two trees. Evidence for within and between clade clustering of taxa on the two trees.

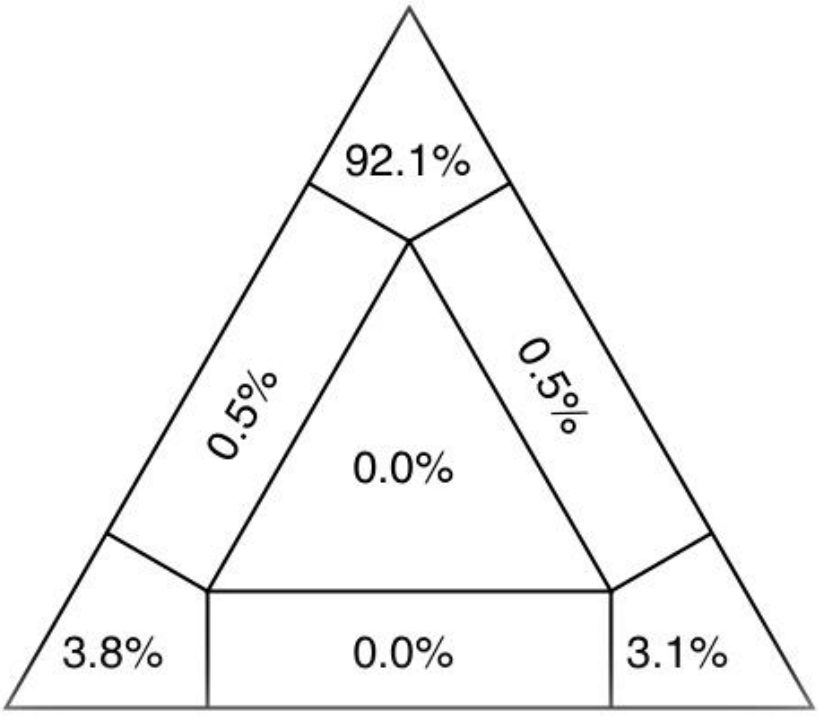


Figure 6-4. ML map for the 16S rRNA gene resulting from quartet puzzling. Numbers in the vertices of the triangles represent the percentage of quartets out of 384930 that support that particular configuration of the four taxon groups classified based on Figure 1.

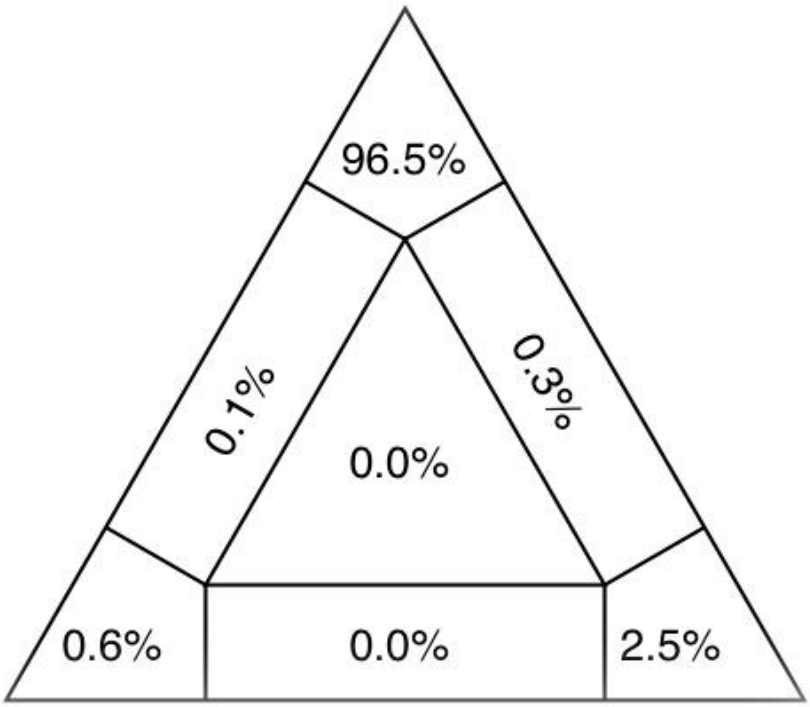


Figure 6-5. ML map for the *rpoB* gene resulting from quartet puzzling. Numbers in the vertices of the triangles represent the percentage of quartets out of 384930 that support that particular configuration of the four taxon groups classified based on Figure 1.

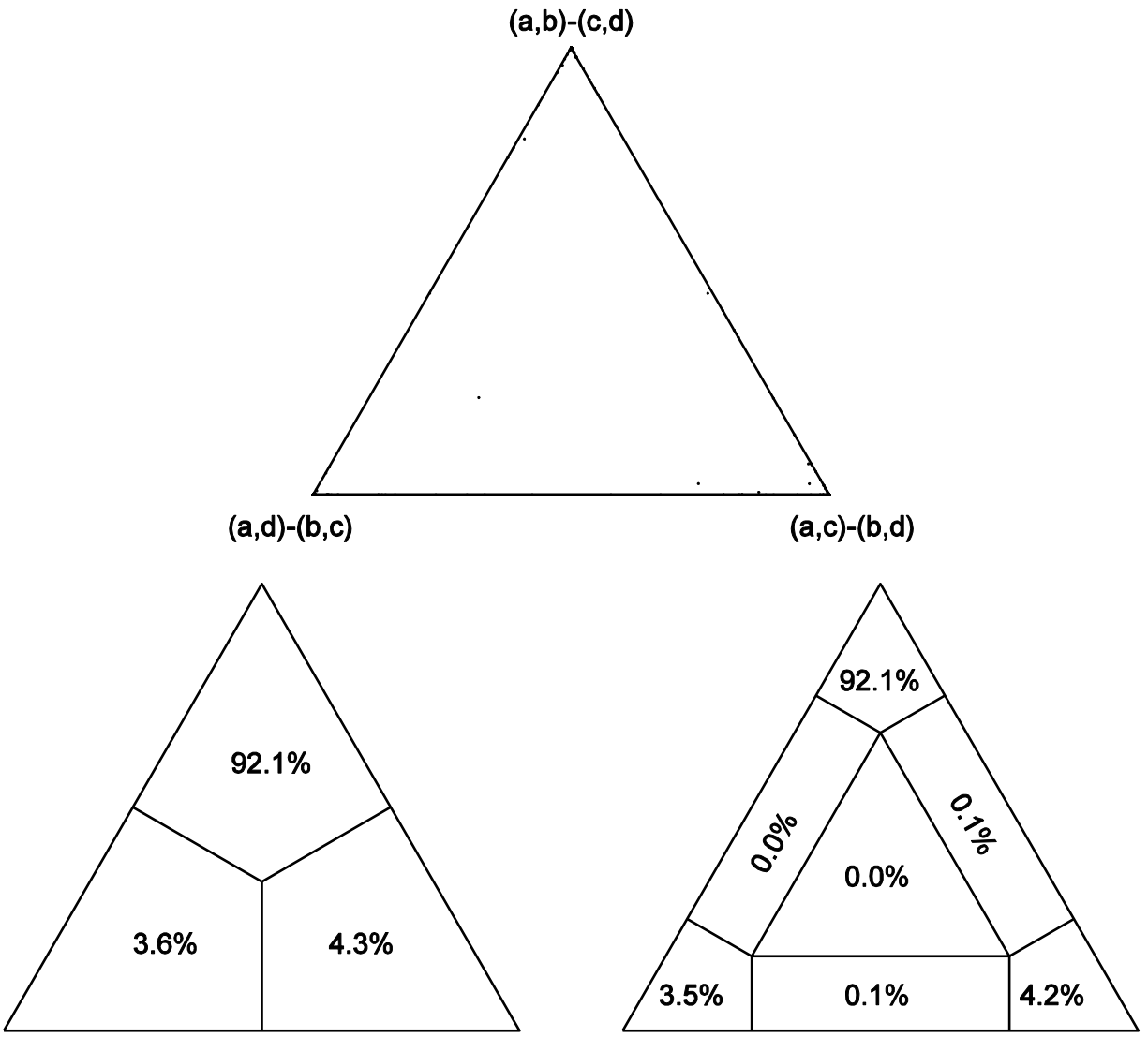


Figure 6-6. ML mapping of quartet puzzling within group A of the 16S rRNA phylogeny. Vertices of the triangle represent the percentage of quartets supporting each particular configuration of the sub groups within group A.

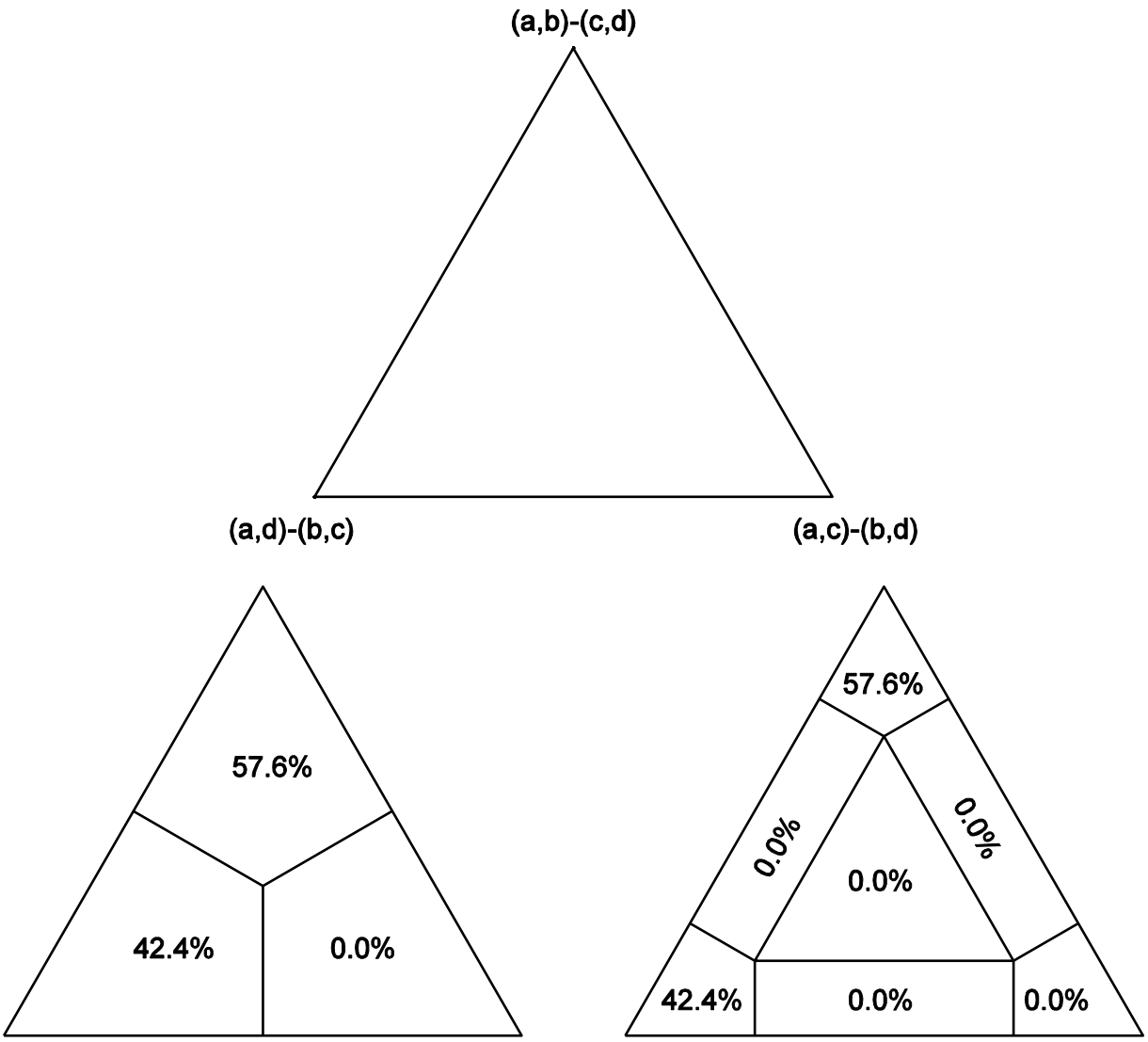


Figure 6-7. ML mapping of quartet puzzling within group B of the 16S rRNA phylogeny. Vertices of the triangle represent the percentage of quartets supporting each particular configuration of the sub groups within group B.

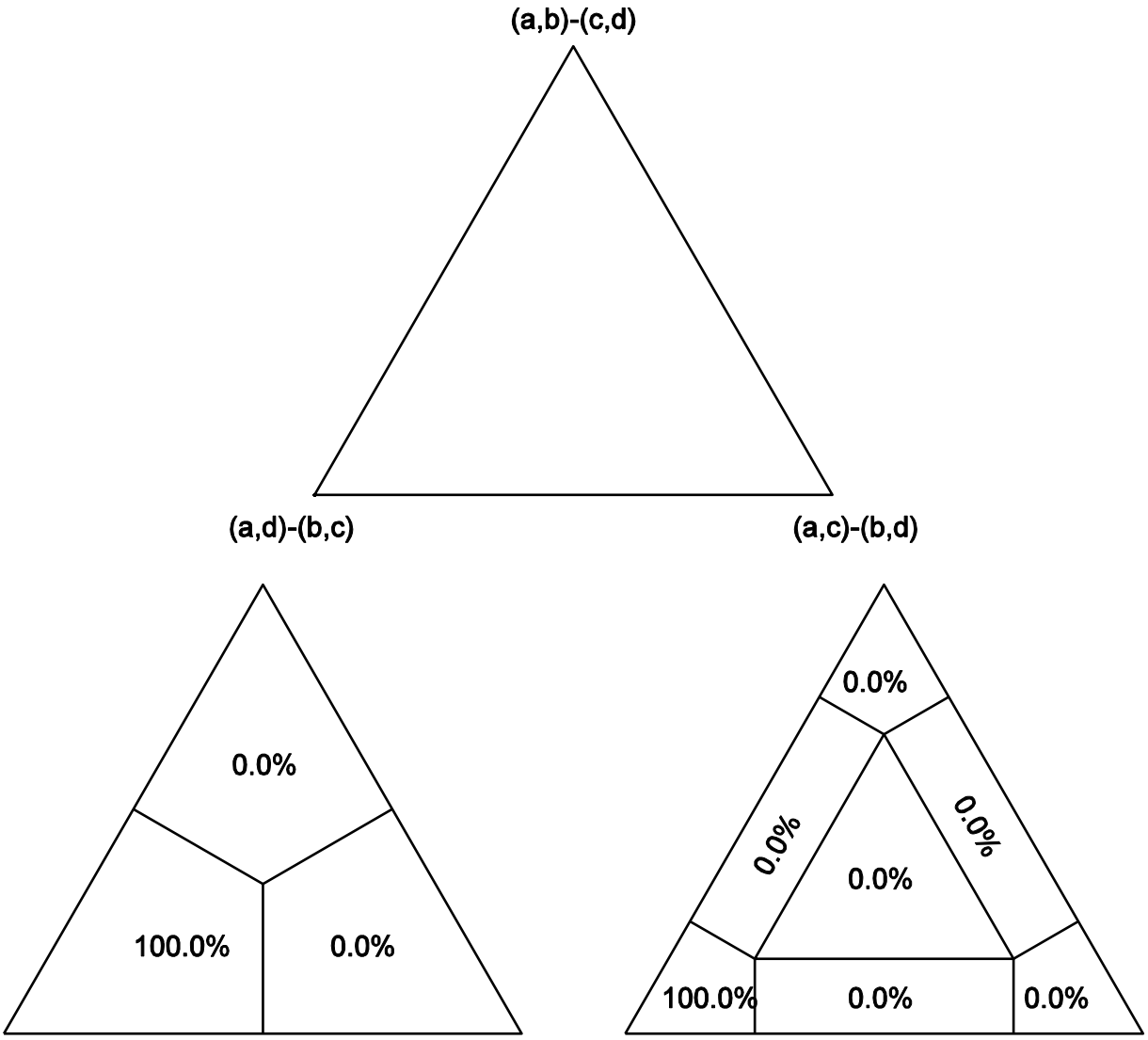


Figure 6-8. ML mapping of quartet puzzling within group C of the 16S rRNA phylogeny. Vertices of the triangle represent the percentage of quartets supporting each particular configuration of the sub groups within group C.

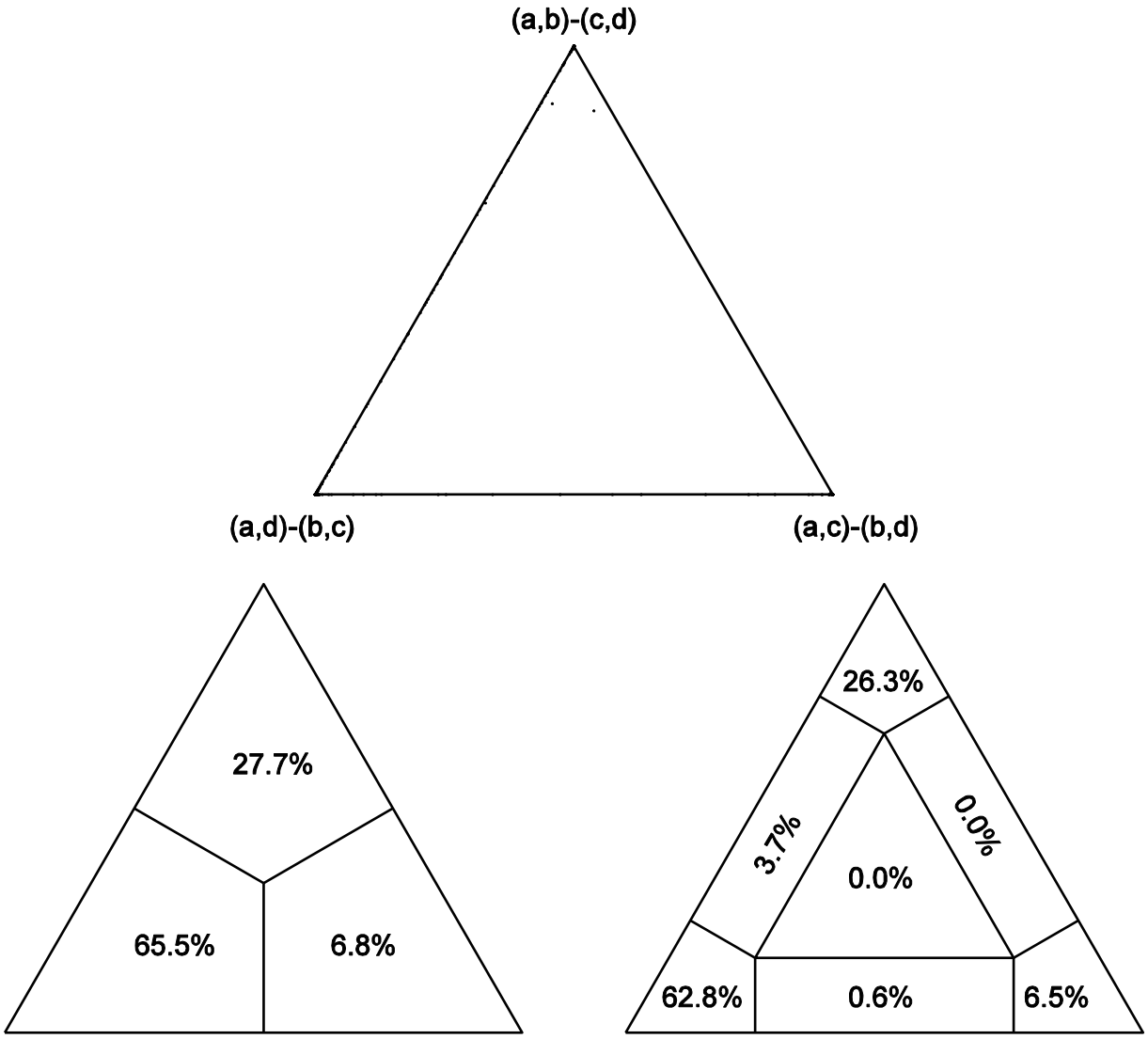


Figure 6-9. ML mapping of quartet puzzling within group D of the 16S rRNA phylogeny. Vertices of the triangle represent the percentage of quartets supporting each particular configuration of the sub groups within group D.

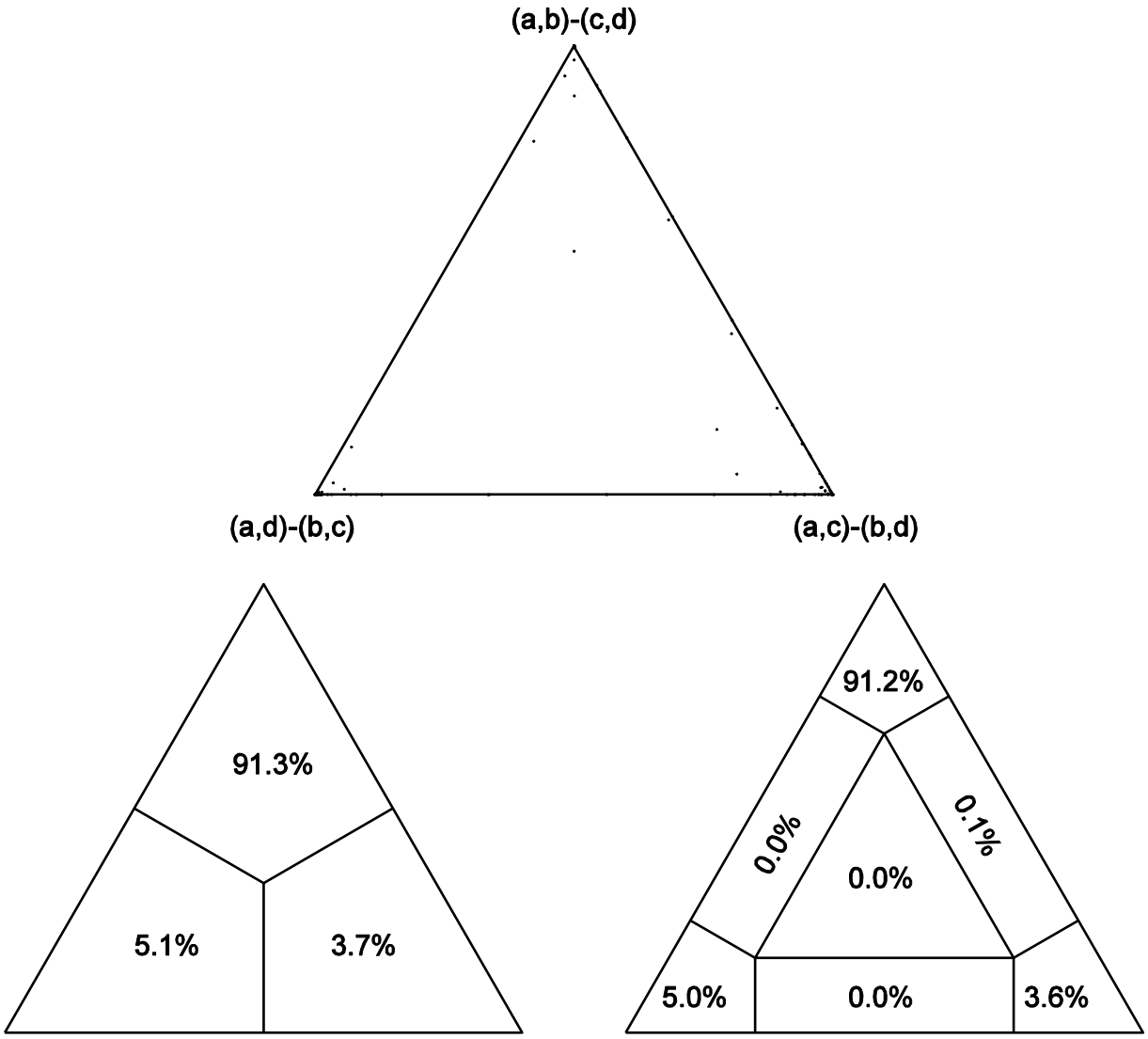


Figure 6-10. ML mapping of quartet puzzling within group A of the *rpoB* phylogeny. Vertices of the triangle represent the percentage of quartets supporting each particular configuration of the sub groups within group A.

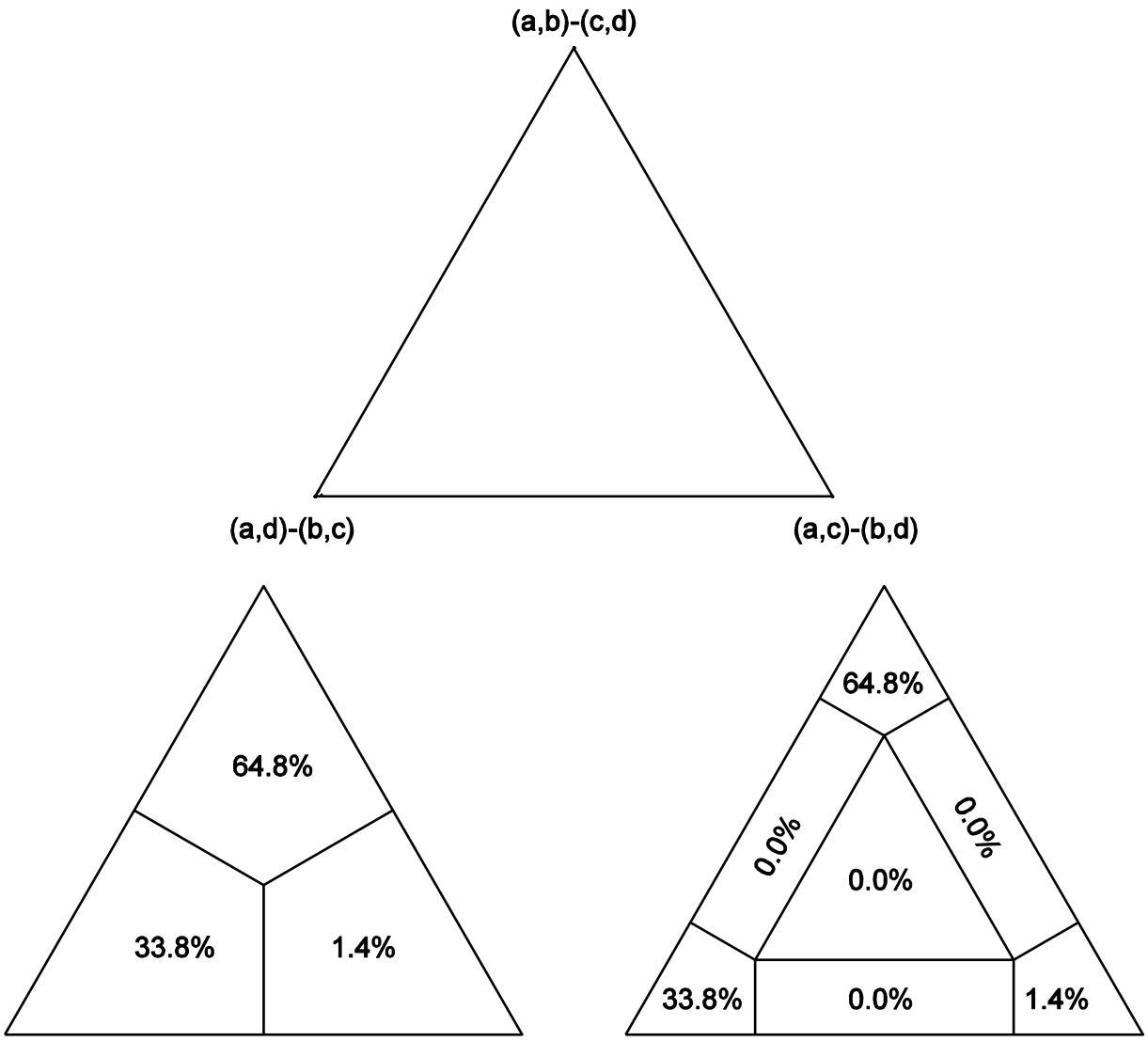


Figure 6-11. ML mapping of quartet puzzling within group B of the *rpoB* phylogeny. Vertices of the triangle represent the percentage of quartets supporting each particular configuration of the sub groups within group B.

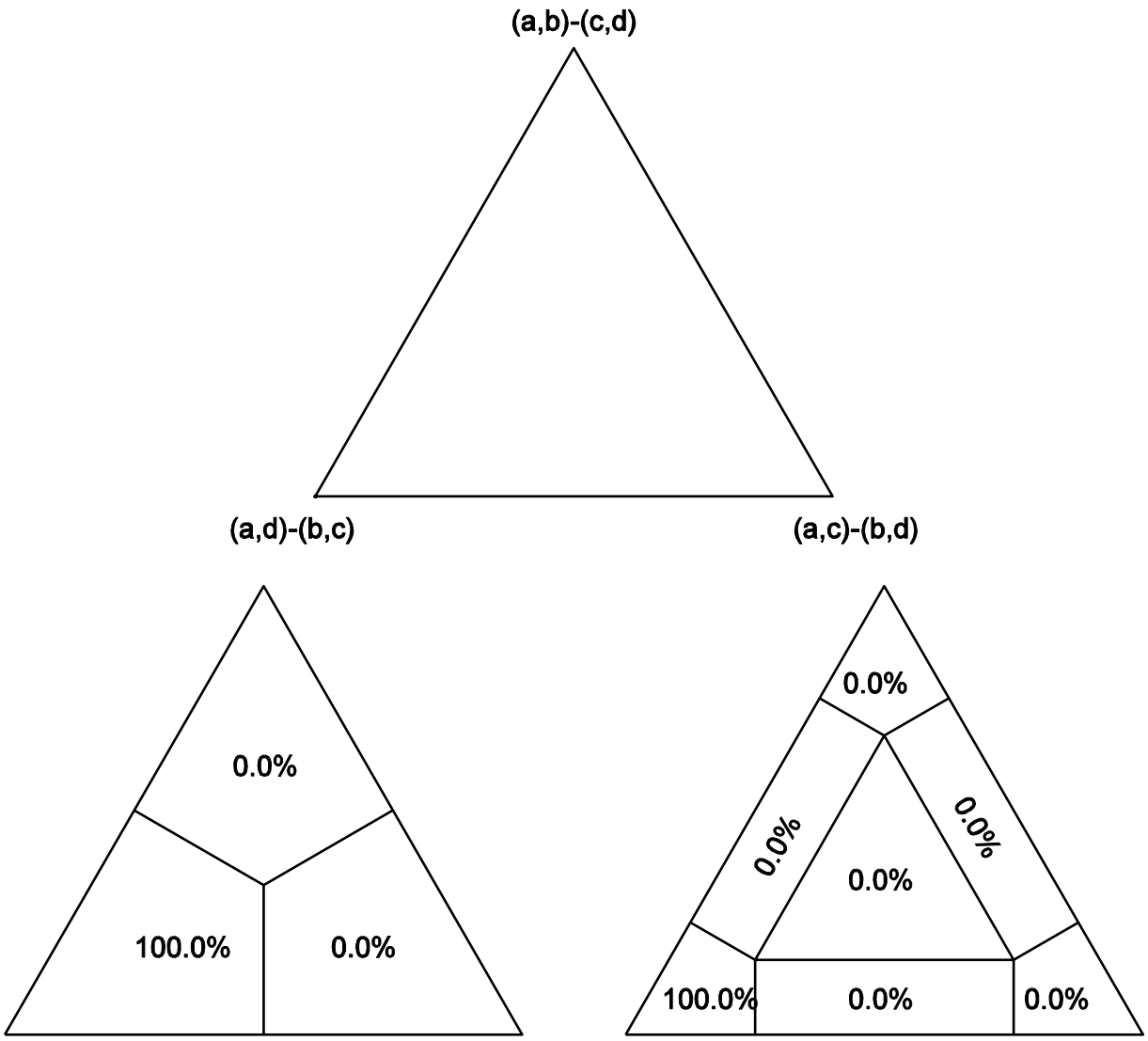


Figure 6-12. ML mapping of quartet puzzling within group C of the *rpoB* phylogeny. Vertices of the triangle represent the percentage of quartets supporting each particular configuration of the sub groups within group C.

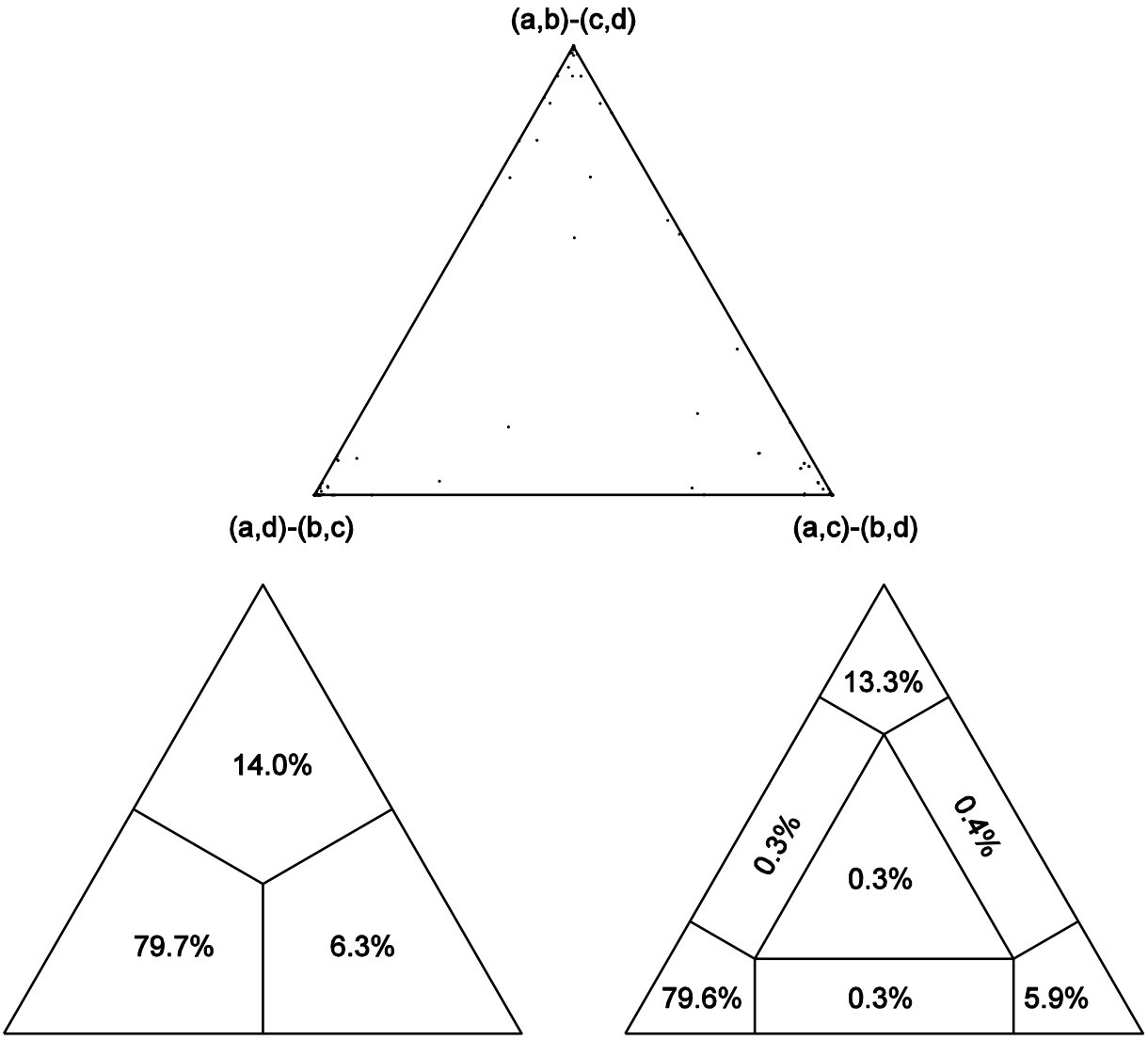


Figure 6-13. ML mapping of quartet puzzling within group D of the *rpoB* phylogeny. Vertices of the triangle represent the percentage of quartets supporting each particular configuration of the sub groups within group D.

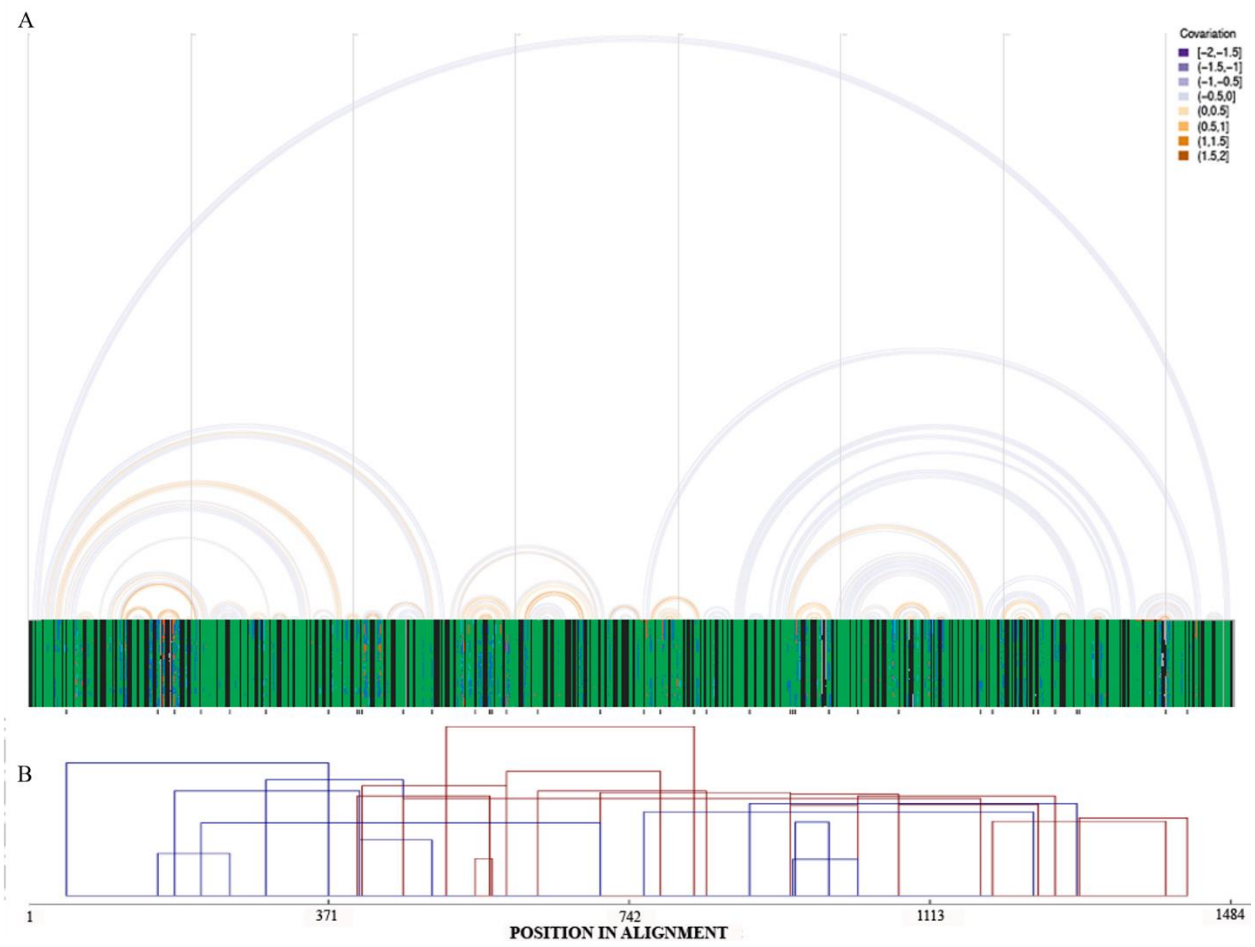
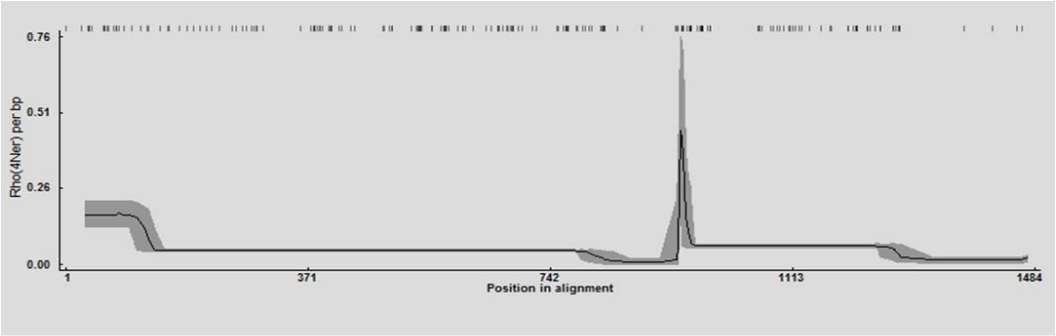


Figure 6-14. **A**. Consensus secondary structure of the Haloarchaeal 16S rRNA predicted by the RNAalifold webserver [228] and displayed along with the multiple sequence alignment using the R-chie webserver [229]. Arcs represent paired bases in the structure and colored based on the covariation of the bases. The bases in the multiple sequence alignment are colored based on base pair status. **B.** Map of the recombination events detected by RDP4 for the 16S rRNA. Each box plotted denotes a recombination event, the position along the 16S rRNA gene, and the size of the recombinant. Events are colored based on distance between species recombining – red for more closely related and blue for distantly related.

A



B

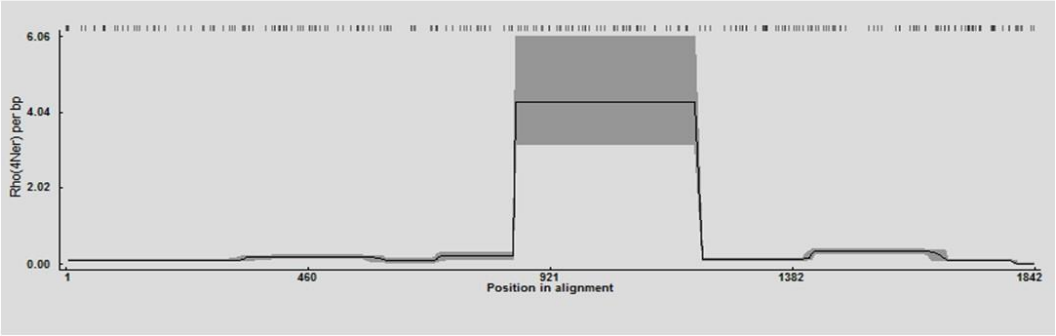


Figure 6-15. Recombination rate plots across the length of the **A.** 16S rRNA gene and **B.** *rpoB* gene. Rate of recombination was determined by running 10000000 MCMC updates on RDP4.

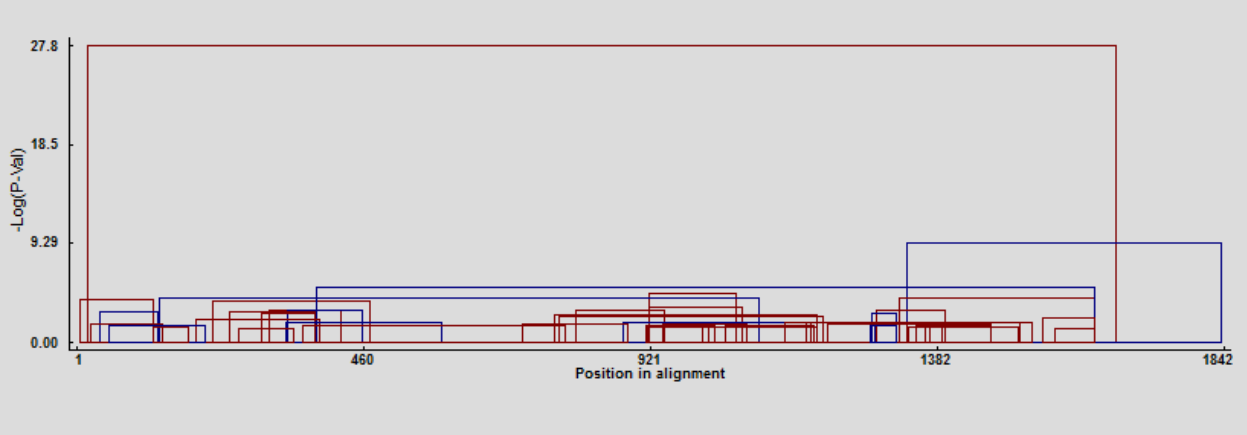


Figure 6-16. Map of the recombination events detected by RDP4 for the *rpoB*. Each box plotted denotes a recombination event, the position along the 16S rRNA gene, and the size of the recombinant. Events are colored based on distance between species recombining – red for more closely related and blue for distantly related.

Chapter 7 : Summary of Conclusions and Implications

7.1 Biogeography of Haloarchaea

The haloarchaea in the twelve different hypersaline environments form communities that are rich and diverse in species. Though this is the case in every site tested, the trend in the data obtained suggests that the salterns might be more rich and diverse in haloarchaea than the naturally occurring lakes. These communities formed are unique to each site studied both in the composition detected at the genus level as well as in the presence or absence of OTUs defined at the liberal cutoff of 95% sequence similarity. This endemism in haloarchaeal communities does seem to be a factor of geographical isolation. Unlike the eukaryotes however, the haloarchaea do not posit a distance-decay relationship in the similarity between communities with respect to the physical separation. This holds true with regards to the type of hypersaline environment sampled as well. The salterns are not more similar to one another than the lakes are or vice versa. Though the communities are unique, there is evidence for a small fraction of shared OTUs between sites that are geographically distant. This sharing is more likely to have occurred because of dispersal between sites rather than mutations in the sequences. Hence, geographical separation does not seem to act as a barrier to dispersal and yet endemism is apparent. This must be due to two reasons – the rate of dispersal is significantly slower than the rate of evolution; and not every haloarchaeal species that is dispersed can successfully invade and survive in an existing community.

7.2 Temporal analysis of one haloarchaeal community

The haloarchaeal community in the Eilat saltern in Israel surveyed through three years at five different time points described the existence of a core group of OTUs that represented 52% of all the sequences obtained throughout. These belonged to *Halorubrum, Haloarcula,* and unknown haloarchaea. 89% of the sequences were found in two or more time points suggesting that the community in Eilat is stable through time. There are fluctuations in the relative abundances of the members with changes in salinity, however the taxa present in Eilat remain consistent through time. Haloarchaeal community stability is not unique to the Eilat saltern. Stability at the genus level seems to be common to the haloarchaeal communities in hypersaline [59, 143, 165, 186]. This stability must drive the maintenance of endemic populations throughout the world.

7.3 Dynamics of Individual Haloarchaea in a population

Analysis of 43 isolates from the Aran-Bidgol Lake in Iran using MLSA with *atpB*, *ef-2*, *glnA*, *ppsA* and *rpoB* genes revealed that these isolates belonged to the *Halorubrum* and *Haloarcula* genera. Phylogenetic reconstruction of the concatenated sequence data identified three polytomous groups, two of which belonged to the *Halorubrum* cluster while the third group belonged to *Haloarcula.* Isolates within each of these polytomous groups did not differ by more than 10 nucleotides across the 2500 nucleotides tested when compared with the other isolates clustering in the same phylogroups. According to MLSA, these isolates were identical. However, comparing the whole genome fingerprints for these isolates derived by RAPD showed distinction in the patterning. Every isolate had a distinct whole genome fingerprint. Discrepancies in the gene level similarity and the whole genome structure led to the hypothesis that in natural populations, the rate of accumulation of genetic variation through HGT and recombination is much greater than the rate of accruement of third codon substitutions.

7.4 Assay the extent of recombination in the Haloarchaea

Analyzing the 16S rRNA and *rpoB* genes from 109 haloarchaeal genomes exposed the extent and effect of recombination in two highly conserved genes that are most commonly used molecular markers. Comparing phylogenies derived for each gene showed many discrepancies. Interestingly, partial gene fragments of each gene did not portray congruent evolutionary histories. Each half of the 16S rRNA and *rpoB* genes has a different evolutionary history when compared to the other half and the full length gene phylogeny. Both full length genes provide an average of the phylogeny for the two halves and though neither gene leaves any quartet unresolved, they display evidence for recombination between deep branching groups which results in discordant quartets during puzzling and ML mapping. The number of discordant quartets increase when reforming the analysis with only members within each group. There is more recombination within each group than between. Recombination is extensive in both gene datasets with highly supported evidence for many recombination events in each gene across the different haloarchaea. Many species show evidence of recombination and no genera are absent of it. In some species, there is evidence for multiple putative recombination events within one gene sequence. Recombination is so rampant in the haloarchaea that the predicted rate of recombination per site in the 16S rRNA and *rpoB* genes is considerably greater than the rate of mutation per site predicted.

7.5 Working model for the genesis of new haloarchaeal species

A model for the genesis of new species in the haloarchaea can be derived based on all the evidence described so far (see figure 7-1). Given that the haloarchaea are highly promiscuous through frequent gene transfer and recombination [80-91, 100], a barrier to this homogenizing force is a prerequisite to drive variation. This work describes the effect of geographic separation on the communities formed by the haloarchaea. Though leaky, geographic isolation does act as a barrier to recombination, although, not by hindering dispersal of the haloarchaea. Unlike in *Vibrio* [29], *Sulfolobus* [30], or some of the other systems studied [12, 17, 31-36], the spatial distribution observed in the haloarchaeal communities would suggest allopatric speciation as a driving force in this class of archaea, similar to the eukaryotes, but the inherent mechanism does not fit within either vicariance or dispersal since dispersal isn't completely inhibited. Modes of dispersal of haloarchaea aren't well known, there is one example of *Halococcus spp.* being found in the nostrils salt gland of *Calonectris diomedea* [168] that are known to migrate long distances for foraging and breeding, but dispersal does occur far and yet endemic populations are observed.

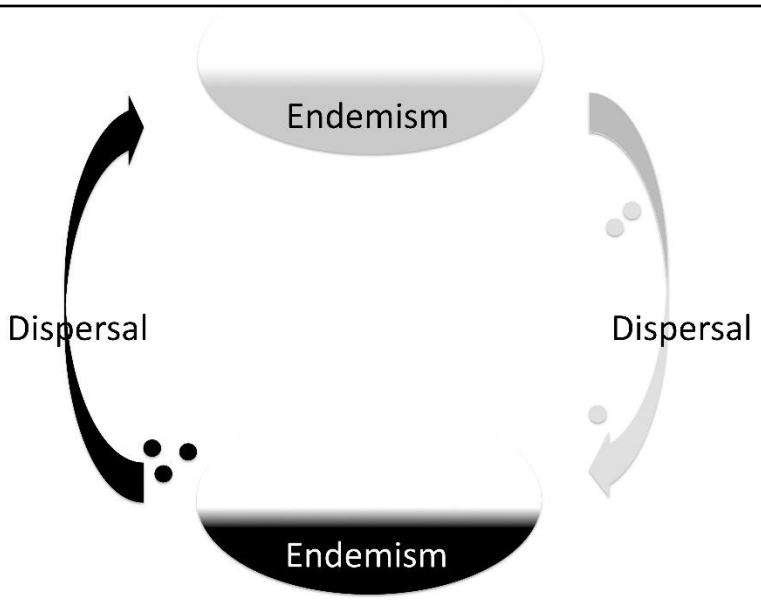


Figure 7-1. Working model for the genesis of new haloarchaeal species.

Endemic populations of haloarchaea are maintained by the stability in their communities. Unlike many niches that are subject to major fluctuations in environmental conditions [47-54], hypersaline lakes and salterns provide a more stable surrounding for the haloarchaea to thrive and maintain their diversity. This stability in the haloarchaeal community must inhibit dispersed incoming species from taking a foothold within the existing community. Stable, species rich, and diverse communities are often impenetrable [37-41]. As an additional defense mechanism, stable communities 'mount a territorial defense' through antimicrobial agents and complex resistant pathways [42, 43]. The haloarchaea produce their own diverse group of antimicrobial agents called haocins [230-234]. Communities also often employ various mechanisms to circumvent environmental volatility and maintain stability. Stable genetic heterogeneity in a community is achieved by maintaining high species diversity, and this is most frequently managed by phage predation of the dominant or abundant species in the community [45, 46]. The haloarchaea are predated by many viruses [235-239]. Despite not knowing the exact mechanisms of resistance to colonization and maintenance of diversity in the haloarchaeal communities, it can be concluded that the haloarchaeal communities are in fact stable and endemic in different hypersaline environments owing only in part to environmental stability.

Within each of these endemic haloarchaeal communities, the members undergo recombination profusely, thereby homogenizing the population at the gene level [80-86]. Recombination occurs at various scales – exchange of large chunks of genomic DNA [81, 82]; transfer of individual genes [87-90]; and even smaller fragments of varying sizes between genes as seen in the 16S rRNA and *rpoB*. This process happens rapidly, possibly during every replication event [100], at rates much greater than that of mutation even in the most conserved genes. Such extensive recombination within each population works towards ensuring the diversity between

populations and hence with geographic separation leads to divergent communities. Within populations however, apart from acting as a homogenizing force, recombination also gives rise to ‘hopeful monsters’ [240] or stable chimaeras that end up being divergent from the recombining parental strains thereby forming new species.

References

1. Avise, J.C., *Phylogeography: the history and formation of species*. 2000: Harvard university press.
2. Naor, A. and U. Gophna, *Cell fusion and hybrids in Archaea: prospects for genome shuffling and accelerated strain development for biotechnology.* Bioengineered, 2013. **4**(3): p. 126-9.
3. Howard, D.J. and S.H. Berlocher, *Endless Forms: Species and Speciation*. 1998: Oxford University Press.
4. Ronquist, F., *Dispersal-vicariance analysis: a new approach to the quantificationof histroical biogeography.* Systematic Biology, 1997. **46**: p. 195-203.
5. Darwin, C., *The Voyage of the Beagle*. 1909: P.F. Collier.
6. Wallace, A.R., *The geographical distribution of animals, with a study of the relations of living and extinct faunas as elucidating the past changes of the earth's surface. By Alfred Russel Wallace*. Vol. 1. 1876, London: Macmillan and Co.
7. Mayr, E., *Systematics and the origin of species from the viewpoint of a zoologist*. 1942, New York: Columbia University Press.
8. Martiny, J.B., et al., *Microbial biogeography: putting microorganisms on the map.* Nat Rev Microbiol, 2006. **4**(2): p. 102-12.
9. Papke, R.T. and D.M. Ward, *The importance of physical isolation to microbial diversification.* FEMS Microbiol Ecol, 2004. **48**: p. 293-303.
10. Pointing, S.B. and J. Belnap, *Microbial colonization and controls in dryland systems.* Nat Rev Microbiol, 2012.
11. Whitaker, R.J., *Allopatric origins of microbial species.* Philos Trans R Soc Lond B Biol Sci, 2006. **361**(1475): p. 1975-84.
12. Zhaxybayeva, O., et al., *Cell sorting analysis of geographically separated hypersaline environments.* Extremophiles, 2013. **17**(2): p. 265-75.

13. Dillon, J.G., et al., *Patterns of microbial diversity along a salinity gradient in the Guerrero Negro solar saltern, Baja CA Sur, Mexico.* Front Microbiol, 2013. **4**: p. 399.
14. Oh, D., et al., *Diversity of Haloquadratum and other haloarchaea in three, geographically distant, Australian saltern crystallizer ponds.* Extremophiles, 2010. **14**(2): p. 161-9.
15. Ragon, M., et al., *Different biogeographic patterns of prokaryotes and microbial eukaryotes in epilithic biofilms.* Mol Ecol, 2012. **21**(15): p. 3852-68.
16. Youssef, N.H., K.N. Ashlock-Savage, and M.S. Elshahed, *Phylogenetic diversities and community structure of members of the extremely halophilic Archaea (order Halobacteriales) in multiple saline sediment habitats.* Appl Environ Microbiol, 2012. **78**(5): p. 1332-44.
17. Salazar, G., et al., *Global diversity and biogeography of deep-sea pelagic prokaryotes.* ISME J, 2015.
18. Chong, C.W., D.A. Pearce, and P. Convey, *Emerging spatial patterns in Antarctic prokaryotes.* Front Microbiol, 2015. **6**: p. 1058.
19. Zhang, Y., et al., *Drivers shaping the diversity and biogeography of total and active bacterial communities in the South China Sea.* Mol Ecol, 2014. **23**(9): p. 2260-74.
20. Wilkins, D., et al., *Biogeographic partitioning of Southern Ocean microorganisms revealed by metagenomics.* Environ Microbiol, 2013. **15**(5): p. 1318-33.
21. Gaisin, V.A., et al., *Biogeography of thermophilic phototrophic bacteria belonging to Roseiflexus genus.* FEMS Microbiol Ecol, 2016.
22. van der Gast, C.J., *Microbial biogeography: the end of the ubiquitous dispersal hypothesis?* Environ Microbiol, 2015. **17**(3): p. 544-6.
23. Lanzen, A., et al., *Surprising prokaryotic and eukaryotic diversity, community structure and biogeography of Ethiopian soda lakes.* PLoS One, 2013. **8**(8): p. e72577.
24. Polme, S., et al., *Global biogeography of Alnus-associated Frankia actinobacteria.* New Phytol, 2014. **204**(4): p. 979-88.

25. Barreto, D.P., et al., *Distance-decay and taxa-area relationships for bacteria, archaea and methanogenic archaea in a tropical lake sediment.* PLoS One, 2014. **9**(10): p. e110128.
26. Kellogg, C.A. and D.W. Griffin, *Aerobiology and the global transport of desert dust.* Trends Ecol Evol, 2006. **21**(11): p. 638-44.
27. Finlay, B.J. and K.J. Clark, *Ubiquitous dispersal of microbial species.* Nature, 1999. **400**: p. 1061-1063.
28. Finlay, B.J., G.F. Esteban, and T. Fenchel, *Global diversity and body size.* Nature, 1996. **383**: p. 132-133.
29. Shapiro, B.J., et al., *Population genomics of early events in the ecological differentiation of bacteria.* Science, 2012. **336**(6077): p. 48-51.
30. Cadillo-Quiroz, H., et al., *Patterns of gene flow define species of thermophilic Archaea.* PLoS Biol, 2012. **10**(2): p. e1001265.
31. Bahl, J., et al., *Ancient origins determine global biogeography of hot and cold desert cyanobacteria.* Nat Commun, 2011. **2**: p. 163.
32. Darling, K.F., et al., *Molecular evidence for genetic mixing of Arctic and Antarctic subpolar populations of planktonic foraminifers.* Nature, 2000. **405**(6782): p. 43-7.
33. Finlay, B.J., *Global dispersal of free-living microbial eukaryote species.* Science, 2002. **296**(5570): p. 1061-3.
34. Ionescu, D., et al., *Biogeography of thermophilic cyanobacteria: insights from the Zerka Ma'in hot springs (Jordan).* FEMS Microbiol Ecol, 2010. **72**(1): p. 103-13.
35. Papke, R.T., et al., *Geographical isolation in hot spring cyanobacteria.* Environ Microbiol, 2003. **5**(8): p. 650-9.
36. Staley, J.T. and J.J. Gosink, *Poles apart: biodiversity and biogeography of sea ice bacteria.* Annu Rev Microbiol, 1999. **53**: p. 189-215.
37. van Elsas, J.D., et al., *Survival of genetically marked Escherichia coli O157:H7 in soil as affected by soil microbial community shifts.* ISME J, 2007. **1**(3): p. 204-14.

38. van Elsas, J.D., et al., *Survival of Escherichia coli in the environment: fundamental and public health aspects.* ISME J, 2011. **5**(2): p. 173-83.
39. Ma, J., et al., *Influence of bacterial communities based on 454-pyrosequencing on the survival of Escherichia coli O157:H7 in soils.* FEMS Microbiol Ecol, 2013. **84**(3): p. 542-54.
40. van Elsas, J.D., et al., *Microbial diversity determines the invasion of soil by a bacterial pathogen.* Proceedings of the National Academy of Sciences, 2012. **109**(4): p. 1159-1164.
41. Yao, Z., et al., *Interaction between the microbial community and invading Escherichia coli O157:H7 in soils from vegetable fields.* Appl Environ Microbiol, 2014. **80**(1): p. 70-6.
42. Juhasz, J., et al., *Emergence of collective territorial defense in bacterial communities: horizontal gene transfer can stabilize microbiomes.* PLoS One, 2014. **9**(4): p. e95511.
43. He, X., et al., *The social structure of microbial community involved in colonization resistance.* ISME J, 2014. **8**(3): p. 564-74.
44. Erkus, O., et al., *Multifactorial diversity sustains microbial community stability.* ISME J, 2013. **7**(11): p. 2126-36.
45. Rodriguez-Valera, F., et al., *Explaining microbial population genomics through phage predation.* Nat Rev Microbiol, 2009. **7**(11): p. 828-36.
46. Thingstad, T.F., *Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems.* Limnology and Oceanography, 2000. **45**(6): p. 1320-1328.
47. Rogers, B.F. and R.L. Tate, *Temporal analysis of the soil microbial community along a toposequence in Pineland soils.* Soil Biol Biochem, 2001. **33**(10): p. 1389-1401.
48. Liu, L., J. Yang, and Y. Zhang, *Genetic diversity patterns of microbial communities in a subtropical riverine ecosystem (Jiulong River, southeast China).* Hydrobiologia, 2011. **678**(1): p. 113-125.

49. Winter, C., et al., *Longitudinal changes in the bacterial community composition of the Danube River: a whole-river approach.* Appl Environ Microbiol, 2007. **73**(2): p. 421-31.
50. Wu, N., B. Schmalz, and N. Fohrer, *Distribution of phytoplankton in a German lowland river in relation to environmental factors.* J Plankton Res, 2011. **33**(5): p. 807-820.
51. Liu, L., et al., *Patterns in the composition of microbial communities from a subtropical river: effects of environmental, spatial and temporal factors.* PLoS One, 2013. **8**(11): p. e81232.
52. Dickerson, T.L. and H.N. Williams, *Functional diversity of bacterioplankton in three North Florida freshwater lakes over an annual cycle.* Microb Ecol, 2014. **67**(1): p. 34-44.
53. Zaccone, R., et al., *Seasonal dynamics of prokaryotic abundance and activities in relation to environmental parameters in a transitional aquatic ecosystem (Cape Peloro, Italy).* Microb Ecol, 2014. **67**(1): p. 45-56.
54. Dillon, J., L. McMath, and A. Trout, *Seasonal changes in bacterial diversity in the Salton Sea.* Hydrobiologia, 2009. **632**(1): p. 49-64.
55. Ferris, M.J. and D.M. Ward, *Seasonal distributions of dominant 16S rRNA-defined populations in a hot spring microbial mat examined by denaturing gradient gel electrophoresis.* Appl Environ Microbiol, 1997. **63**(4): p. 1375-81.
56. Rodriguez, R.L., et al., *Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill.* ISME J, 2015. **9**(9): p. 1928-40.
57. Oren, A., *Life at High Salt Concentrations.*, in *The Prokaryotes*, M. Dworkin, et al., Editors. 2006, Springer New York. p. 263-282.
58. Mutlu, M.B., et al., *Prokaryotic diversity in Tuz Lake, a hypersaline environment in Inland Turkey.* FEMS Microbiol Ecol, 2008. **65**(3): p. 474-83.
59. Podell, S., et al., *Seasonal fluctuations in ionic concentrations drive microbial succession in a hypersaline lake community.* ISME J, 2014. **8**(5): p. 979-90.

60. Litchfield, C.D., *Survival strategies for microorganisms in hypersaline environments and their relevance to life on early Mars.* Meteorit Planet Sci, 1998. **33**(4): p. 813-9.
61. Oren, A., *Molecular ecology of extremely halophilic Archaea and Bacteria.* FEMS Microbiol Ecol, 2002. **39**(1): p. 1-7.
62. Antón, J., et al., *Fluorescence in situ hybridization analysis of the prokaryotic community inhabiting crystallizer ponds.* Environ Microbiol, 1999. **1**(6): p. 517-23.
63. Ghai, R., et al., *New abundant microbial groups in aquatic hypersaline environments.* Sci Rep, 2011. **1**: p. 135.
64. Ochsenreiter, T., F. Pfeifer, and C. Schleper, *Diversity of Archaea in hypersaline environments characterized by molecular-phylogenetic and cultivation studies.* Extremophiles, 2002. **6**(4): p. 267-74.
65. Benlloch, S., et al., *Archaeal biodiversity in crystallizerpPonds from a solar saltern: culture versus PCR.* Microb Ecol, 2001. **41**(1): p. 12-19.
66. Benlloch, S., et al., *Prokaryotic genetic diversity throughout the salinity gradient of a coastal solar saltern.* Environ Microbiol, 2002. **4**(6): p. 349-60.
67. Martínez-Murcia, A.J., S.G. Acinas, and F. Rodríguez-Valera, *Evaluation of prokaryotic diversity by restrictase digestion of 16S rDNA directly amplified from hypersaline environments.* FEMS Microbiology Ecology, 1995. **17**(4): p. 247-255.
68. Maturrano, L., et al., *Microbial diversity in Maras salterns, a hypersaline environment in the Peruvian Andes.* Appl Environ Microbiol, 2006. **72**(6): p. 3887-95.
69. Litchfield, C.D. and P.M. Gillevet, *Microbial diversity and complexity in hypersaline environments: a preliminary assessment.* J Ind Microbiol Biotechnol, 2002. **28**(1): p. 48-55.
70. Øvreås, L., et al., *Characterization of microbial diversity in hypersaline environments by melting profiles and reassociation kinetics in combination with terminal restriction fragment length polymorphism (T-RFLP).* Microb Ecol, 2003. **46**(3): p. 291-301.

71. Walsh, D.A., R.T. Papke, and W.F. Doolittle, *Archaeal diversity along a soil salinity gradient prone to disturbance.* Environ Microbiol, 2005. **7**(10): p. 1655-66.
72. Grant, S., et al., *Novel archaeal phylotypes from an East African alkaline saltern.* Extremophiles, 1999. **3**(2): p. 139-45.
73. Sabet, S., et al., *Characterization of halophiles isolated from solar salterns in Baja California, Mexico.* Extremophiles, 2009. **13**(4): p. 643-56.
74. Lozier, R.H., R.A. Bogomolni, and W. Stoeckenius, *Bacteriorhodopsin: a light-driven proton pump in Halobacterium Halobium.* Biophys J, 1975. **15**(9): p. 955-62.
75. Jones, J.G., D.C. Young, and S. DasSarma, *Structure and organization of the gas vesicle gene cluster on the Halobacterium halobium plasmid pNRC100.* Gene, 1991. **102**(1): p. 117-22.
76. Benlloch, S., A.J. Martínez-Murcia, and F. Rodríguez-Valera, *Sequencing of bacterial and archaeal 16S rRNA genes directly amplified from a hypersaline environment.* Syst Appl Microbiol, 1995. **18**(4): p. 574-581.
77. Legault, B.A., et al., *Environmental genomics of "Haloquadratum walsbyi" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species.* BMC Genomics, 2006. **7**: p. 171.
78. Pašic, L., et al., *Diversity of halophilic archaea in the crystallizers of an Adriatic solar saltern.* FEMS Microbiol Ecol, 2005. **54**(3): p. 491-8.
79. Bidle, K., et al., *Research Article: A phylogenetic analysis of haloarchaea found in a solar saltern.* BIOS, 2005. **76**(2): p. 89-96.
80. Cuadros-Orellana, S., et al., *Genomic plasticity in prokaryotes: the case of the square haloarchaeon.* ISME J, 2007. **1**(3): p. 235-45.
81. DeMaere, M.Z., et al., *High level of intergenera gene exchange shapes the evolution of haloarchaea in an isolated Antarctic lake.* Proc Natl Acad Sci U S A, 2013. **110**(42): p. 16939-44.

82. Naor, A., et al., *Low species barriers in halophilic archaea and the formation of recombinant hybrids.* Curr Biol, 2012. **22**(15): p. 1444-8.
83. Papke, R.T., et al., *Frequent recombination in a saltern population of Halorubrum.* Science, 2004. **306**(5703): p. 1928-9.
84. Papke, R.T., et al., *Searching for species in haloarchaea.* Proc Natl Acad Sci U S A, 2007. **104**(35): p. 14092-7.
85. Williams, D., J.P. Gogarten, and R.T. Papke, *Quantifying homologous replacement of loci between haloarchaeal species.* Genome Biol Evol, 2012. **4**(12): p. 1223-44.
86. Podell, S., et al., *Assembly-driven community genomics of a hypersaline microbial ecosystem.* PLoS One, 2013. **8**(4): p. e61692.
87. Ng, W.V., et al., *Genome sequence of Halobacterium species NRC-1.* Proc Natl Acad Sci U S A, 2000. **97**(22): p. 12176-81.
88. Sharma, A.K., et al., *Evolution of rhodopsin ion pumps in haloarchaea.* BMC Evol Biol, 2007. **7**: p. 79.
89. Boucher, Y., et al., *Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes.* J Bacteriol, 2004. **186**(12): p. 3980-90.
90. Andam, C.P., et al., *Ancient origin of the divergent forms of leucyl-tRNA synthetases in the Halobacteriales.* BMC Evol Biol, 2012. **12**: p. 85.
91. Nelson-Sathi, S., et al., *Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea.* Proc Natl Acad Sci U S A, 2012. **109**(50): p. 20537-42.
92. Rosenshine, I., R. Tchelet, and M. Mevarech, *The mechanism of DNA transfer in the mating system of an archaebacterium.* Science, 1989. **245**(4924): p. 1387-9.
93. Maiden, M.C., et al., *Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.* Proc Natl Acad Sci U S A, 1998. **95**(6): p. 3140-5.

94. de la Haba, R.R., et al., *Multilocus sequence analysis of the family Halomonadaceae.* Int J Syst Evol Microbiol, 2012. **62**(Pt 3): p. 520-38.
95. Feil, E.J., M.C. Enright, and B.G. Spratt, *Estimating the relative contributions of mutation and recombination to clonal diversification: a comparison between Neisseria meningitidis and Streptococcus pneumoniae.* Res Microbiol, 2000. **151**(6): p. 465-9.
96. Thompson, J.R., et al., *Genotypic diversity within a natural coastal bacterioplankton population.* Science, 2005. **307**(5713): p. 1311-3.
97. Tettelin, H., et al., *Comparative genomics: the bacterial pan-genome.* Curr Opin Microbiol, 2008. **11**(5): p. 472-7.
98. Lapierre, P. and J.P. Gogarten, *Estimating the size of the bacterial pan-genome.* Trends Genet, 2009. **25**(3): p. 107-10.
99. Lynch, E.A., et al., *Sequencing of seven haloarchaeal genomes reveals patterns of genomic flux.* PLoS One, 2012. **7**(7): p. e41389.
100. Ram Mohan, N., et al., *Evidence from phylogenetic and genome fingerprinting analyses suggests rapidly changing variation in Halorubrum and Haloarcula populations.* Front Microbiol, 2014. **5**: p. 143.
101. Gophna, U., et al., *No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales.* ISME J, 2015. **9**(9): p. 2021-7.
102. Woese, C.R. and G.E. Fox, *Phylogenetic structure of the prokaryotic domain: the primary kingdoms.* Proc Natl Acad Sci U S A, 1977. **74**(11): p. 5088-90.
103. Case, R.J., et al., *Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies.* Appl Environ Microbiol, 2007. **73**(1): p. 278-88.
104. Pei, A.Y., et al., *Diversity of 16S rRNA genes within individual prokaryotic genomes.* Appl Environ Microbiol, 2010. **76**(12): p. 3886-97.

105. Větrovský, T. and P. Baldrian, *The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses.* PLoS One, 2013. **8**(2): p. e57923.
106. Bodilis, J., et al., *Variable copy number, intra-genomic heterogeneities and lateral transfers of the 16S rRNA gene in Pseudomonas.* PLoS One, 2012. **7**(4): p. e35647.
107. Coenye, T. and P. Vandamme, *Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes.* FEMS Microbiol Lett, 2003. **228**(1): p. 45-9.
108. Acinas, S.G., et al., *Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons.* J Bacteriol, 2004. **186**(9): p. 2629-35.
109. Sanz, J.L., et al., *Variable rRNA gene copies in extreme halobacteria.* Nucleic Acids Res, 1988. **16**(16): p. 7827-32.
110. Hofman, J.D., R.H. Lau, and W.F. Doolittle, *The number, physical organization and transcription of ribosomal RNA cistrons in an archaebacterium: Halobacterium halobium.* Nucleic Acids Res, 1979. **7**(5): p. 1321-33.
111. Hui, I. and P.P. Dennis, *Characterization of the ribosomal RNA gene clusters in Halobacterium cutirubrum.* J Biol Chem, 1985. **260**(2): p. 899-906.
112. Mylvaganam, S. and P.P. Dennis, *Sequence heterogeneity between the two genes encoding 16S rRNA from the halophilic archaebacterium Haloarcula marismortui.* Genetics, 1992. **130**(3): p. 399-410.
113. Mevarech, M., et al., *Isolation and characterization of the rRNA gene clusters of Halobacterium marismortui.* J Bacteriol, 1989. **171**(6): p. 3479-85.
114. Baliga, N.S., et al., *Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea.* Genome Res, 2004. **14**(11): p. 2221-34.
115. Vreeland, R.H., et al., *Halosimplex carlsbadense gen. nov., sp. nov., a unique halophilic archaeon, with three 16S rRNA genes, that grows*

only in defined medium with glycerol and acetate or pyruvate. Extremophiles, 2002. **6**(6): p. 445-52.
116. Cui, H.L., et al., *Intraspecific polymorphism of 16S rRNA genes in two halophilic archaeal genera, Haloarcula and Halomicrobium.* Extremophiles, 2009. **13**(1): p. 31-7.
117. Dahllof, I., H. Baillie, and S. Kjelleberg, *rpoB-based microbial community analysis avoids limitations inherent in 16S rRNA gene intraspecies heterogeneity.* Appl Environ Microbiol, 2000. **66**(8): p. 3376-80.
118. Peixoto, R.S., et al., *Use of rpoB and 16S rRNA genes to analyse bacterial diversity of a tropical soil using PCR and DGGE.* Lett Appl Microbiol, 2002. **35**(4): p. 316-20.
119. Vos, M., et al., *A comparison of rpoB and 16S rRNA as markers in pyrosequencing studies of bacterial diversity.* PLoS One, 2012. **7**(2): p. e30600.
120. Roux, S., et al., *Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (Bacteria and Archaea) in ecosystems.* FEMS Microbiol Ecol, 2011. **78**(3): p. 617-28.
121. Khamis, A., D. Raoult, and B. La Scola, *Comparison between rpoB and 16S rRNA gene sequencing for molecular identification of 168 clinical isolates of Corynebacterium.* J Clin Microbiol, 2005. **43**(4): p. 1934-6.
122. Mollet, C., M. Drancourt, and D. Raoult, *rpoB sequence analysis as a novel basis for bacterial identification.* Mol Microbiol, 1997. **26**(5): p. 1005-11.
123. Ki, J.S., W. Zhang, and P.Y. Qian, *Discovery of marine Bacillus species by 16S rRNA and rpoB comparisons and their usefulness for species identification.* J Microbiol Methods, 2009. **77**(1): p. 48-57.
124. Khamis, A., D. Raoult, and B. La Scola, *rpoB gene sequencing for identification of Corynebacterium species.* J Clin Microbiol, 2004. **42**(9): p. 3925-31.

125. da Mota, F.F., et al., *Use of rpoB gene analysis for identification of nitrogen-fixing Paenibacillus species as an alternative to the 16S rRNA gene.* Lett Appl Microbiol, 2004. **39**(1): p. 34-40.
126. Lee, M.J., et al., *Comparison of rpoB gene sequencing, 16S rRNA gene sequencing, gyrB multiplex PCR, and the VITEK2 system for identification of Acinetobacter clinical isolates.* Diagn Microbiol Infect Dis, 2014. **78**(1): p. 29-34.
127. Adekambi, T. and M. Drancourt, *Dissection of phylogenetic relationships among 19 rapidly growing Mycobacterium species by 16S rRNA, hsp65, sodA, recA and rpoB gene sequencing.* Int J Syst Evol Microbiol, 2004. **54**(Pt 6): p. 2095-105.
128. Enache, M., et al., *Phylogenetic relationships within the family Halobacteriaceae inferred from rpoB' gene and protein sequences.* Int J Syst Evol Microbiol, 2007. **57**(Pt 10): p. 2289-95.
129. Minegishi, H., et al., *Further refinement of the phylogeny of the Halobacteriaceae based on the full-length RNA polymerase subunit B' (rpoB') gene.* Int J Syst Evol Microbiol, 2010. **60**(Pt 10): p. 2398-408.
130. Soppa, J., J. Duschl, and D. Oesterhelt, *Bacterioopsin, haloopsin, and Sensory Opsin I of the halobacterial isolate Halobacterium sp. Strain SG1: three new members of a growing family.* J Bacteriol, 1993. **175**: p. 2720-2726.
131. Drachev, L., et al., *Direct measurement of electric current generation by cytochrome oxidase, H+-ATPase and bacteriorhodopsin.* Nature, 1974. **249**(455): p. 321-324.
132. Skulachev, V., *Conversion of light energy into electric energy by bacteriorhodopsin.* FEBS letters, 1976. **64**(1): p. 23-25.
133. Javor, B.J., *Planktonic standing crop and nutrients in a saltern ecosystem.* Limnol Oceanogr, 1983. **28**: p. 153-159.
134. Oren, A. and M. Shilo, *Bacteriorhodopsin in a bloom of Halobacteria in the Dead Sea.* Arch Microbiol, 1981. **130**(185-187).

135. Papke, R.T., et al., *Diversity of bacteriorhodopsins in different hypersaline waters from a single Spanish saltern.* Environ Microbiol, 2003. **5**(11): p. 1039-45.
136. Pašic, L., et al., *Haloarchaeal communities in the crystallizers of two adriatic solar salterns.* Can J Microbiol, 2007. **53**(1): p. 8-18.
137. Litchfield, C.D., et al., *Temporal and salinity impacts on the microbial diversity at the Eilat, Israel solar salt plant.* Global NEST Journal 2009. **11**(1): p. 86-90.
138. Schloss, P.D., et al., *Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities.* Appl Environ Microbiol, 2009. **75**(23): p. 7537-41.
139. Martin, A.P., *Phylogenetic approaches for describing and comparing the diversity of microbial communities.* Appl Environ Microbiol, 2002. **68**(8): p. 3673-82.
140. Lozupone, C., M. Hamady, and R. Knight, *UniFrac-an online tool for comparing microbial community diversity in a phylogenetic context.* BMC Bioinformatics, 2006. **7**: p. 371.
141. Trigui, H., et al., *Characterization of heterotrophic prokaryote subgroups in the Sfax coastal solar salterns by combining flow cytometry cell sorting and phylogenetic analysis.* Extremophiles, 2011. **15**(3): p. 347-58.
142. Dyall-Smith, M., ed. *The Halohandbook: Protocols for haloarchaeal genetics*. 7 ed. 2008: http://www.haloarchaea.com/resources/halohandbook/index.html .
143. Gomariz, M., et al., *Retinal-binding proteins mirror prokaryotic dynamics in multipond solar salterns.* Environ Microbiol, 2015. **17**(2): p. 514-26.
144. Fernandez, A.B., et al., *Prokaryotic taxonomic and metabolic diversity of an intermediate salinity hypersaline habitat assessed by metagenomics.* FEMS Microbiol Ecol, 2014. **88**(3): p. 623-35.

145. Fernandez, A.B., et al., *Metagenome sequencing of prokaryotic microbiota from two hypersaline ponds of a marine saltern in santa pola, Spain.* Genome Announc, 2013. **1**(6).
146. Plominsky, A.M., et al., *Metagenome sequencing of the microbial community of a solar saltern crystallizer pond at cahuil lagoon, chile.* Genome Announc, 2014. **2**(6).
147. Rodriguez-Brito, B., et al., *Viral and microbial community dynamics in four aquatic environments.* ISME J, 2010. **4**(6): p. 739-51.
148. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
149. Maddison, D. and W. Maddison, *MacClade*. 2003, Sinauer Associates: Sunderland, MA.
150. McMurdie, P.J. and S. Holmes, *phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.* PLoS One, 2013. **8**(4): p. e61217.
151. Charlop-Powers, Z. and S.F. Brady, *phylogeo: an R package for geographic analysis and visualization of microbiome data.* Bioinformatics, 2015. **31**(17): p. 2909-11.
152. Chao, A., *Nonparametric estimation of the number of classes in a population.* Scand. J. Statist., 1984. **11**: p. 265-270.
153. Chao, A. and S.-M. Lee, *Estimating the number of classes via sample coverage.* J. Amer. Statist. Assoc., 1992. **87**: p. 210-217.
154. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
155. Singleton, D.R., et al., *Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples.* Appl Environ Microbiol, 2001. **67**(9): p. 4374-6.
156. Anderson, T.W., *On the distribution of the two-sample Cramer-von Mises Criterion.* 1962: p. 1148-1159.
157. Jaccard, P., *Étude comparative de la distribution florale dans une portion des Alpes et des Jura.* Bulletin del la Société Vaudoise des Sciences Naturelles, 1901. **37**: p. 547-579.

158. Lance, G.N. and W.T. Williams, *Computer Programs for Hierarchical Polythetic Classification ("Similarity Analyses").* The Computer Journal, 1966. **9**(1): p. 60-64.
159. Beerli, P. and P. Beerli, *How to use MIGRATE or why are Markov chain Monte Carlo programs difficult to use?*

Population Genetics for Animal Conservation. 2009: Cambridge University Press.
160. Grant, W.D., *Introductory chapter: half a lifetime in soda lakes*, in *Halophilic Microorganisms*, A. Ventosa, Editor. 2004, Springer-Verlag: Heidelberg.
161. Zafrilla, B., et al., *Biodiversity of Archaea and floral of two inland saltern ecosystems in the Alto Vinalopo Valley, Spain.* Saline Systems, 2010. **6**: p. 10.
162. Bowman, J.P., et al., *The microbial composition of three limnologically disparate hypersaline Antarctic lakes.* FEMS Microbiol Lett, 2000. **183**(1): p. 81-8.
163. Liu, W.T., et al., *Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA.* Appl Environ Microbiol, 1997. **63**(11): p. 4516-22.
164. Makhdoumi-Kakhki, A., et al., *Prokaryotic diversity in Aran-Bidgol salt lake, the largest hypersaline playa in Iran.* Microbes Environ, 2012. **27**(1): p. 87-93.
165. Boujelben, I., et al., *Spatial and seasonal prokaryotic community dynamics in ponds of increasing salinity of Sfax solar saltern in Tunisia.* A Van Leeuw J Microb, 2012.
166. Fernández, A.B., et al., *Comparison of prokaryotic community structure from Mediterranean and Atlantic saltern concentrator ponds by a metagenomic approach.* Front Microbiol, 2014. **5**: p. 196.
167. Litchfield, C.D., et al., *Comparisons of the polar lipid and pigment profiles of two solar salterns located in Newark, California, USA, and Eilat, Israel.* Extremophiles, 2000. **4**(5): p. 259-65.

168. Brito-Echeverria, J., et al., *Occurrence of Halococcus spp. in the nostrils salt glands of the seabird Calonectris diomedea.* Extremophiles, 2009. **13**(3): p. 557-65.
169. Papke, R.T., et al., *Diversity of bacteriorhodopsins in different hypersaline waters from a single Spanish saltern.* Environ Microbiol, 2003. **5**(11): p. 1039-45.
170. Rosner, B., *Percentage Points for a Generalized ESD Many-Outlier Procedure.* Technometrics, 1983. **25**(2): p. 165-172.
171. Hampel, F.R., *A General Qualitative Definition of Robustness.* The Annals of Mathematical Statistics, 1971. **42**(6): p. 1887-1896.
172. Tukey, J.W., *Exploratory Data Analysis*. 1977: Addison-Wesley Publishing Company.
173. Tavaré, S., *Some probablilistic and statistical problems in the anlysis of DNA sequences.*, in *Some mathematical questions in biology*, R.M. Miura, Editor. 1986, American Mathematical Society: Providence, RI.
174. Guindon, S., et al., *New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.* Syst Biol, 2010. **59**(3): p. 307-21.
175. Spearman, C., *The proof and measurement of association between two things.* Am J Psychol, 1904. **15**(1): p. 72-101.
176. Gomariz, M., et al., *From community approaches to single-cell genomics: the discovery of ubiquitous hyperhalophilic Bacteroidetes generalists.* ISME J, 2015. **9**(1): p. 16-31.
177. Rodriguez-Brito, B., et al., *Viral and microbial community dynamics in four aquatic environments.* ISME J, 2010. **4**(6): p. 739-51.
178. Oesterhelt, D. and W. Stoeckenius, *Functions of a new photoreceptor membrane.* Proc Natl Acad Sci U S A, 1973. **70**(10): p. 2853-7.
179. Stoeckenius, W., D. Bivin, and K. McGinnis, *Photoactive pigments in Halobacteria from the Gavish Sabkha.*, in *Hypersaline Ecosystems*, G. Friedman and W. Krumbein, Editors. 1985, Springer Berlin Heidelberg. p. 288-295.

180. Wright, A.D., *Phylogenetic relationships within the order Halobacteriales inferred from 16S rRNA gene sequences.* Int J Syst Evol Microbiol, 2006. **56**(Pt 6): p. 1223-7.
181. Oren, A., *Diversity of halophilic microorganisms: environments, phylogeny, physiology, and applications.* J Ind Microbiol Biotechnol, 2002. **28**(1): p. 56-63.
182. Oren, A., S. Duker, and S. Ritter, *The polar lipid composition of Walsby's square bacterium.* FEMS Microbiology Letters, 1996. **138**(2–3): p. 135-140.
183. Suzuki, M.T. and S.J. Giovannoni, *Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR.* Appl Environ Microbiol, 1996. **62**(2): p. 625-30.
184. Bardgett, R.D., et al., *Seasonal changes in soil microbial communities along a fertility gradient of temperate grasslands.* Soil Biol Biochem, 1999. **31**(7): p. 1021-1030.
185. Väätänen, P., *Effects of environmental factors on microbial populations in brackish waters off the southern coast of Finland.* Appl Environ Microbiol, 1980. **40**(1): p. 48-54.
186. Gasol, J., et al., *Control of heterotrophic prokaryotic abundance and growth rate in hypersaline planktonic environments.* Aquat Microb Ecol, 2004. **34**(2): p. 193-206.
187. Narasingarao, P., et al., *De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities.* ISME J, 2012. **6**(1): p. 81-93.
188. Allers, T., et al., *Development of additional selectable markers for the halophilic archaeon Haloferax volcanii based on the leuB and trpA genes.* Appl Environ Microbiol, 2004. **70**(2): p. 943-53.
189. Darriba, D., et al., *jModelTest 2: more models, new heuristics and parallel computing.* Nat Methods, 2012. **9**(8): p. 772.
190. Akaike, H., *A new look at the statistical model* IEEE Trans. Automatic Control, 1974. **AC-19**: p. 716 - 723.

191. Tamura, K., et al., *MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.* Mol Biol Evol, 2011. **28**(10): p. 2731-9.
192. Shivu, M.M., et al., *Molecular characterization of Vibrio harveyi bacteriophages isolated from aquaculture environments along the coast of India.* Environ Microbiol, 2007. **9**(2): p. 322-31.
193. Winget, D.M. and K.E. Wommack, *Randomly amplified polymorphic DNA PCR as a tool for assessment of marine viral richness.* Appl Environ Microbiol, 2008. **74**(9): p. 2612-8.
194. Barrangou, R., et al., *Characterization of six Leuconostoc fallax bacteriophages isolated from an industrial sauerkraut fermentation.* Appl Environ Microbiol, 2002. **68**(11): p. 5452-8.
195. Dice, L.R., *Measures of the Amount of Ecologic Association Between Species.* Ecology, 1945. **26**(3): p. 6.
196. Sokal, R.R. and F.J. Rohlf, *The Comparison of Dendrograms by Objective Methods.* Taxon, 1962. **11**(2): p. 8.
197. Martinez-Murcia, A.J. and F. Rodriquez-Valera, *The use of arbitrarily primed PCR (AP-PCR) to develop taxa specific DNA probes of known sequence.* FEMS Microbiology Letters, 1994. **124**: p. 265-270.
198. Bolhuis, H., et al., *The genome of the square archaeon Haloquadratum walsbyi : life at the limits of water activity.* BMC Genomics, 2006. **7**: p. 169.
199. Kennedy, S.P., et al., *Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence.* Genome Res, 2001. **11**(10): p. 1641-50.
200. DasSarma, S., U.L. RajBhandary, and H.G. Khorana, *High-frequency spontaneous mutation in the bacterio-opsin gene in Halobacterium halobium is mediated by transposable elements.* Proc Natl Acad Sci U S A, 1983. **80**(8): p. 2201-5.
201. Fullmer, M.S., et al., *Population and genomic analysis of the genus Halorubrum.* Front. Microbiol. - Extreme Microbiology, 2014.

202. Mackwan, R.R., et al., *An unusual pattern of spontaneous mutations recovered in the halophilic archaeon Haloferax volcanii.* Genetics, 2007. **176**(1): p. 697-702.
203. Kottemann, M., et al., *Physiological responses of the halophilic archaeon Halobacterium sp. strain NRC1 to desiccation and gamma irradiation.* Extremophiles, 2005. **9**(3): p. 219-27.
204. McCready, S., *The repair of ultraviolet light-induced DNA damage in the halophilic archaebacteria, Halobacterium cutirubrum, Halobacterium halobium and Haloferax volcanii.* Mutat Res, 1996. **364**(1): p. 25-32.
205. Lange, C., et al., *Gene conversion results in the equalization of genome copies in the polyploid haloarchaeon Haloferax volcanii.* Mol Microbiol, 2011. **80**(3): p. 666-77.
206. Ehrlich, G.D., et al., *The distributed genome hypothesis as a rubric for understanding evolution in situ during chronic bacterial biofilm infectious processes.* FEMS Immunol Med Microbiol, 2010. **59**(3): p. 269-79.
207. Gogarten, J.P. and J.P. Townsend, *Horizontal gene transfer, genome innovation and evolution.* Nat Rev Microbiol, 2005. **3**(9): p. 679-87.
208. Darling, A.E., et al., *Mauve assembly metrics.* Bioinformatics, 2011. **27**(19): p. 2756-7.
209. Darling, A.E., B. Mau, and N.T. Perna, *progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement.* PLoS One, 2010. **5**(6): p. e11147.
210. Tavare, S., *Some probablilistic and statistical problems in the anlysis of DNA sequences*, in *Some mathematical questions in biology*, R.M. Miura, Editor. 1986, American Mathematical Society: Providence, RI.
211. Paradis, E., J. Claude, and K. Strimmer, *APE: Analyses of Phylogenetics and Evolution in R language.* Bioinformatics, 2004. **20**(2): p. 289-90.
212. Chakerian, J. and S. Holmes, *distory: Distance Between Phylogenetic Histories*. 2013: http://CRAN.R-project.org/package=distory.

213. Kuhner, M.K. and J. Felsenstein, *A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates.* Mol Biol Evol, 1994. **11**(3): p. 459-68.
214. Strimmer, K. and A. von Haeseler, *Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies.* Molecular Biology and Evolution, 1996. **13**(7): p. 964.
215. Schmidt, H.A., et al., *TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.* Bioinformatics, 2002. **18**(3): p. 502-4.
216. Hasegawa, M., H. Kishino, and T. Yano, *Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.* J Mol Evol, 1985. **22**(2): p. 160-74.
217. Bruen, T.C., H. Philippe, and D. Bryant, *A simple and robust statistical test for detecting the presence of recombination.* Genetics, 2006. **172**(4): p. 2665-81.
218. Huson, D.H. and D. Bryant, *Application of phylogenetic networks in evolutionary studies.* Mol Biol Evol, 2006. **23**(2): p. 254-67.
219. Martin, D.P., et al., *RDP4: Detection and analysis of recombination patterns in virus genomes.* Virus Evolution, 2015. **1**(1).
220. Martin, D. and E. Rybicki, *RDP: detection of recombination amongst aligned sequences.* Bioinformatics, 2000. **16**(6): p. 562-3.
221. Padidam, M., S. Sawyer, and C.M. Fauquet, *Possible emergence of new geminiviruses by frequent recombination.* Virology, 1999. **265**(2): p. 218-25.
222. Martin, D.P., et al., *A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints.* AIDS Res Hum Retroviruses, 2005. **21**(1): p. 98-102.
223. Smith, J.M., *Analyzing the mosaic structure of genes.* J Mol Evol, 1992. **34**(2): p. 126-9.
224. Posada, D. and K.A. Crandall, *Evaluation of methods for detecting recombination from DNA sequences: computer simulations.* Proc Natl Acad Sci U S A, 2001. **98**(24): p. 13757-62.

225. Gibbs, M.J., J.S. Armstrong, and A.J. Gibbs, *Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences.* Bioinformatics, 2000. **16**(7): p. 573-82.
226. Boni, M.F., D. Posada, and M.W. Feldman, *An exact nonparametric method for inferring mosaic structure in sequence triplets.* Genetics, 2007. **176**(2): p. 1035-47.
227. Drummond, D.A., et al., *On the conservative nature of intragenic recombination.* Proc Natl Acad Sci U S A, 2005. **102**(15): p. 5380-5.
228. Bernhart, S.H., et al., *RNAalifold: improved consensus structure prediction for RNA alignments.* BMC Bioinformatics, 2008. **9**(1): p. 1-13.
229. Lai, D., et al., *R-CHIE: a web server and R package for visualizing RNA secondary structures.* Nucleic Acids Res, 2012. **40**(12): p. e95.
230. Rodriguez-Valera, F., G. Juez, and D.J. Kushner, *Halocins: salt-dependent bacteriocins produced by extremely halophilic rods.* Canadian Journal of Microbiology, 1982. **28**(1): p. 151-154.
231. Cheung, J., et al., *Isolation, sequence, and expression of the gene encoding halocin H4, a bacteriocin from the halophilic archaeon Haloferax mediterranei R4.* Journal of bacteriology, 1997. **179**(2): p. 548-551.
232. Torreblanca, M., I. Meseguer, and A. Ventosa, *Production of halocin is a practically universal feature of archaeal halophilic rods.* Letters in applied microbiology, 1994. **19**(4): p. 201-205.
233. Meseguer, I. and F. Rodriguez-Valera, *Production and purification of halocin H4.* FEMS microbiology letters, 1985. **28**(2): p. 177-182.
234. Torreblanca, M., I. Meseguer, and F. Rodríguez-Valera, *Halocin H6, a bacteriocin from Haloferax gibbonsii.* Microbiology, 1989. **135**(10): p. 2655-2661.
235. Dyall-Smith, M., S.-L. Tang, and C. Bath, *Haloarchaeal viruses: how diverse are they?* Research in microbiology, 2003. **154**(4): p. 309-313.

236. Tang, S.L., et al., *HF2: a double-stranded DNA tailed haloarchaeal virus with a mosaic genome.* Molecular microbiology, 2002. **44**(1): p. 283-296.
237. Roine, E., et al., *New, closely related haloarchaeal viral elements with different nucleic acid types.* Journal of virology, 2010. **84**(7): p. 3682-3689.
238. Sencilo, A. and E. Roine. *A glimpse of the genomic diversity of haloarchaeal tailed viruses*. in *The Proceedings from Halophiles 2013, the International Congress on Halophilic Microorganisms*. 2015. Frontiers Media SA.
239. Prangishvili, D., P. Forterre, and R.A. Garrett, *Viruses of the Archaea: a unifying view.* Nature Reviews Microbiology, 2006. **4**(11): p. 837-848.
240. Goldschmidt, R., *The material basis of evolution*. Vol. 28. 1940: Yale University Press.