Analysis of the Metabolic and Microbial Diversity of the Solar Salterns in Cabo Rojo using Metagenomics

By:

Ricardo L. Couto-Rodríguez

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

MASTER OF SCIENCE
In
Biology

University of Puerto Rico Mayagüez Campus
2018

Approved by:

_____	_____
Carlos Rodríguez-Minguela, PhD Member, Graduate Comittee	Date
_____	_____
Carlos J. Santos-Flores, PhD Member, Graduate Committee	Date
_____	_____
Rafael Montalvo-Rodríguez, PhD President, Graduate Committee	Date
_____	_____
Kurt Allen Grove, PhD. Representative of Graduate Studies	Date
_____	_____
Ana V. Vélez-Díaz, M.S. Chairperson of the Department	Date

Abstract

The Cabo Rojo solar salterns is a hypersaline environment located in a tropical climate. Conditions in these environments remain stable throughout the year and therefore allow for the establishment of steady microbial communities. In this study, the aim was to describe the microbial community in terms of composition and metabolic processes across time using metagenomics annotation and binning techniques which provide a more comprehensive approach to assess microbial taxonomic and functional diversity. Furthermore, access to functional gene composition can give us insight into microbial processes being undertaken in these tropical hypersaline environments. Sampling was carried out in December 2014, March 2016 and July 2016; samples of 50L each were filtered through a Millipore pressurized filtering system consisting of two nitrocellulose membranes of pore sizes of 5 μm and 0.22 μm respectively. DNA was extracted from the material collected on the 0.22 μm membrane using physical-chemical methods and sequenced using paired end Illumina technologies. The sequencing effort produced 3 paired end libraries that averaged 32 million reads that were subsequently assembled into 3 metagenomes. The microbial diversity was dominated in all three samples by the phylum *Euryarchaeota,* followed by *Bacteroidetes* and *Proteobacteria.* However, assessment at the genus level revealed a change in predominance across all three samples with *Salinibacter* predominating in the first sample whereas *Halorubrum* and *Halogeometricum* predominated in the second and third samples respectively. Possible factors influencing this dynamic community could be unusual rain events as well as anthropogenic impact. Functional gene composition revealed processes related to nitrogen reduction, carbon fixation and sulfur metabolism. Binning efforts returned 6 bins that were further analyzed taxonomically. Five of these genomes were related to the halophilic archaeal genera *Natronomonas, Haloferax, Haloquadratum* and *Halomicrobium* and based on Amino Acid Identity (AAI) are most likely novel organisms. The last bin was related to the *Bacteroidetes* and could represent a novel genus within this phylum based on AAI. These results show a microbial community different to that encountered in other hypersaline environments worldwide and also indicates the presence of putative novel organisms.

Resumen

Las Salinas de Cabo Rojo son un ambiente hipersalino localizado en un clima tropical. Las condiciones en estos ambientes se mantienen estables a través del año y facilita interacciones microbianas más estables. En este estudio, el fue describir la comunidad microbiana a través del tiempo de las salinas de Cabo Rojo en términos de composición microbiana y diversidad de genes funcionales utilizando técnicas metagenómicas que nos proveen un método más comprensivo para medir diversidad microbiana. Adicionalmente, el acceso a la composición genética nos puede proveer información acerca de procesos microbianos ocurriendo en estos ambientes. Se llevaron a cabo tres muestreos en diciembre 2014, marzo 2016 y junlio 2016. Se filtraron muestras de 50L a través de un sistema de filtración Millipore presurizado que consistía de dos membranas de nitrocelulosa con un tamaño de poro de 5μm y 0.22 μm respectivamente. Se realizó extracción de ADN en la membrana de 0.22 μm usando métodos físico-químicos y el producto fue secuenciado utilizando Illumina. La secuenciación produjo 3 bibliotecas pareadas con un promedio de 32 millones de secuencias que subsiguientemente fueron ensambladas en 3 metagenomas. La diversidad microbiana fue predominada en las tres muestras por el filo *Euryarchaeota,* seguido por *Bacteroidetes* y *Proteobacteria.* La determinación a nivel de género reveló que el género predominante cambiaba en las 3 muestras donde *Salinibacter* predominó en el primer metagenoma mientras que *Halorubrum* y *Halogeometricum* predominaron en el segundo y tercer metagenoma respectivamente. Posibles factores que puede influenciar incluyen eventos de alta pluvosidad y actividad antropogénica. En la composición genética se observó procesos relacionados a reducción de nitrógeno, producción primaria y genes relacionados al ciclo de azufre. “Binning” para genomas putativos nos devolvió 6 bins que fueron analizados taxonómicamente. Cinco de estos genomas se relacionaron a arqueas halofílicas *Natronomonas, Haloferax, Haloquadratum* and *Halomicrobium* y basado en Amino Acid Identity pueden representar nuevos organismos. El último bin estaba relacionado al filo *Bacteroidetes* y puede representar un género no descrito en el grupo. A través de este método metagenómico hemos descubierto una diversidad microbiana cambiante diferente a la que es encontrada en otros ambientes hipersalinos e incluye nuevos organismos. Muestreos a través del gradiente de salinidad usando metagenómica nos puede proveer información acerca de otros procesos microbianos y más organismos nuevos.

Dedication

I want to dedicate all of my work to the memory of my grandparents: José “Pepín” Couto, Tomasita Cordero, José “Rigo” Rodríguez, and María “Cuca” García. They have been inspirations throughout my whole life by teaching me to embrace where I come from always rejoicing at whatever achievement came my way. I hope they rejoice with this one.

I would also like to dedicate this work to my parents Luis Couto and Marisa Rodríguez who always pushed me to become a professional, my sister Mara Couto who has been my main source of inspiration throughout my undergraduate and graduate career and my brother Rafael Couto whose drive and determination I have always admired. Thank you, I would not have been who I am without your support and influence.

Acknowledgements

First of all, words cannot describe what an inspiration my advisor Dr. Montalvo-Rodríguez has been towards my development as a professional. Since taking General Microbiology in the fall of 2012, I fell in love with the unseen world of microbiology. Working as an undergraduate in his laboratory and later on as a graduate student, he has always offered me his guidance and encouragement. Most importantly, his number one lesson to me has always been that achievements are meaningless if I don't have the life experiences and lessons to accompany them. Thank you for making me grow as a professional and as an individual.

Special thanks also go to Dr. Carlos Rodríguez-Minguela and Dr. Carlos Santos-Flores for their advice on the development of this thesis project. Their insight was always supportive and helpful. I was always amazed during the committee meetings at their vast knowledge. Thank you for always being available to discuss results with me and occasional anecdotes.

Without Dr. Alex Van Dam, I would not have been able to learn what I did in bioinformatics. The Introduction to Scripting and Python for Bioinformatics course taught me everything I know including assembly of the metagenomes, annotation and binning. Thank you also for always providing advice and recommendations for my bioinformatics work.

Dr. Dimaris Acosta-Mercado for reinforcing my writing skills in the Biological Methods class. Thanks to her guidance and support, this work was manageable and enjoyable to write.

Dr. Rosa Buxeda-Pérez for her always welcoming advice while working together in BioTalents.

Lizbeth Dávila-Santiago was crucial in the analyses of my metagenomic bins. She introduced me to all the tools necessary in order to assess the quality of the bins as well as assigning taxonomy.

Special thanks to Patricia who was my main source of support throughout my three years in the development of this thesis project. Thanks to Jeysika, Sebastián, Dianiris, Gamalier and Santos for being indispensable partners throughout my bachelor's degree and specially throughout my master's degree. I am very lucky to be able to call these people more than friends, they are part of my extended family.

I would like to thank every member from the Extremophiles Laboratory. Particular mention to Nicole Feliberty and Daliana Campos for their assistance in the hundreds of DNA extractions for the metagenomes obtained. Without their help, this thesis would certainly have been a much taller order. To Valeria and Krismarie for offering me the chance to be their mentor and pass on what I have learned. Thanks to Eduardo, Coralis and Rubén for their welcome and assistance in the lab as well as offering me the training necessary to succeed.

Thanks also goes to Magaly Zapata for her humility, availability and friendship throughout my years as an undergraduate and graduate student.

Katherine Carrero and Ana Vélez for their availability, conversations and input into my project as well as for teaching me many things about being a TA and organized in lab work.

Special mention goes to my graduate colleagues, especially the cell physiology crew (Margarette, Kenneth and Neisha), Amelia, José, Ángel and Alicia for the occasional beer and conversations outside of lab work.

Thanks to the undergraduate students I have been able to impact in my 3 years as a teaching assistant. I have learned so much from my students as they have learned from me. I have also made valuable friends from students as a TA, special mention goes to Mariela, Luis Enrique, Javier, Andrea, and Ginna among others.

This research project was sponsored by funds from Howard Hughes Medical Institute (HHMI).

Thanks to the Biology Department for being my home for the past 8 years and for supporting my research.

Finally, thanks to all the people who have helped me one way or another. This thesis project is for you.

Table of Contents

List of Tables

List of Figures

1. Introduction

Hypersaline environments are known for showing a high NaCl content (~3M) (Oren, 2002). Due to their high salt concentration, they are considered extreme environments. The organisms that mostly predominate in these ecosystems are known as extreme halophiles, and thrive starting at NaCl concentrations of 20% (w/v) (Ventosa, 2006). Hypersaline habitats, particularly marine solar salterns, have been studied in temperate locations worldwide. Spain's salterns have been the most extensively studied (Ghai et al., 2011, 2012; Ventosa, 2014), followed by Turkey's where the microbial diversity was determined from six hypersaline lakes across the country (Ozcan et al., 2007), and the Dead Sea (Oren, 2015). The stable diversity due to extreme conditions of these environments makes them model ecosystems for understanding microbial community dynamics (Rodríguez-Valera et al., 2009). Few studies have been performed in tropical environments where conditions remain relatively stable throughout the year with a temperature of approximately 40°C, as well as rain seasons and temperature seasonal changes.

One of the few studies was carried out in Salt Pond, San Salvador, Bahamas which is a hypersaline lake separated from the Atlantic Ocean by a lone Dune Ridge and a coastal road (Puckett et al., 2011). Yannarell et al., (2006) described community dynamics in a site impacted by hurricane Frances involving nitrogen-fixing microbes as well as cyanobacteria. On the other hand, diversity studies have been carried out in Cabo Rojo, Puerto Rico where novel microbes have been isolated and described. *Halogeometricum borinquense* was first isolated from these salterns (Montalvo-Rodríguez et al., 1998) and subsequent diversity surveys yielded novel organisms including *Haloterrigena thermotolerans, Halomonas avicenneae (*later *Kushneria avicenneae), Halobacillus mangrovii,* as well as two recent novel isolates proposed as "*Haloarcula rubripromontorii"* and "*Halorubrum tropicale"*(Montalvo-Rodríguez et al., 2000; Soto-Ramírez et al., 2007, 2008, Sánchez-Nieves et al., 2016a, 2016b). Therefore, microbial diversity

reported in these environments is different from other hypersaline environments worldwide.

However, the aforementioned studies were performed using culture-dependent methods. Culture-dependent methods are limited due to culture media bias and as a result only 0.1% of the diversity can be isolated in pure culture (Thomas et al., 2012). The science of metagenomics has emerged as an answer to the limitations produced by culture media bias and has been applied for studies in these high salt habitats (Ghai et al., 2011, 2012). Therefore, in this project the aim was to obtain a comprehensive assessment of microbial as well as functional gene diversity in the crystallizer ponds of the solar salterns in Cabo Rojo using metagenomics across time. Microbial diversity in the Cabo Rojo salterns by means of culture independent methods was determined by using pyrosequencing of partial 16S rRNA genes in a previous study (Rodríguez-García, 2016). However, to our knowledge, a full scale metagenomic approach to assess gene diversity has not been performed yet in an extreme environment in the Caribbean. Therefore, this project is focused on the study of the first full scale metagenome from an extreme environment in the Caribbean.

2. Literature Review

2.1 Hypersaline Environments and Applications

Hypersaline environments are classified as extreme due to their unusually high salt concentrations. These conditions are a limiting factor when it comes to creating a habitat for organisms to live in. These ecosystems also exhibit other factors such as alkaline pH, high temperatures, low nutrient availability, high concentration of heavy metals and other toxic compounds (Ventosa et al., 2014) that also limit life conditions. Due to these factors, the biota in these environments is reduced to well adapted eukaryotes, prokaryotes and phages (Oren, 2002). Most of these inhabitants with a few notable exceptions, are classified as "halophiles". Two main groups of microorganisms predominate these hypersaline habitats: moderately halophilic bacteria and halophilic archaea (Ventosa, 2006). Due to the special qualities of these microorganisms, it was originally believed that very few diversity was present in these ecosystems. However, with the dawn of genomics and molecular methods, a large diversity of halophiles has been uncovered. Benlloch et al. (2002) sampled saltern ponds of 8%, 22% and 30% (w/v) salinities and employed a denaturing gradient gel electrophoresis (DGGE) method. Their results unveiled a diverse representation of microorganisms along this salinity gradient including organisms from the alpha and gamma *Proteobacteria*, cyanobacteria as well as various representatives from the Archaea domain. In another study Ghai et al., (2011) described the microbiota of two different hypersaline ponds using culture-dependent methods and uncovered novel microbial groups including a group of low GC (Guanine-Cytosine) Actinobacteria, low GC Euryarchaea and a high GC Euryarchaeon. Halophiles possess representatives in all three domains of life and most modes of energy generation known in non-halophiles are also used by halophilic counterparts (Oren, 2002). Consequently, a great metabolic diversity of halophilic and halotolerant microorganisms has been found, many of which have biotechnological potential (Caton et al., 2004; Ventosa, 2008). Different enzymes have been described that can be used for harsh processes such as food

processing, biosynthetic processes, and washing, including hydrolases such as amylases, lipases and proteases (Moreno et al., 2013; Sánchez-Porro et al., 2003; Ventosa et al., 2005). Enzymes isolated from haloarchaea can withstand conditions such as extreme temperatures and pH, making them very favorable for industrial processes. However, their potential is still underdeveloped and very few biocatalysts from haloarchaea have been characterized (Amoozegar et al., 2017). Likewise, different metabolites produced by these organisms, such as ectoine, employed as a cosmetic to combat UV-induced and accelerated skin aging, have been adapted for industrial use (Oren, 2010). Finally, it has also been found that halophiles can be used in bioremediation to treat different waste products found in water and oilfields (Moreno et al., 2013; Oren 2010; Ventosa, 2006, 2008). Therefore, the applications for halophiles range from biotechnological uses to industrial uses.

Halophiles have been isolated from a diverse variety of environments. Initial studies included strains that were isolated from salted foods such as meat or fish. On the other hand, strains were also obtained from fermented products (Ventosa et al., 1998). However, today most studies of hypersaline environments have been focused on aquatic environments such as saline lakes or marine solar salterns. A great number of studies have been overtaken in locations such as the Dead Sea, Great Salt Lake, Santa Pola, France, and Turkey, among others (Ventosa et al., 2014). According to Rodríguez-Brito et al. (2010), microbial community dynamics in these environments have been found to be relatively stable over time. As encountered by Gasol et al. (2004), even when subjected to environmental perturbations, hypersaline communities experience very little change. This simplicity provides the means to make these extreme environments to be model systems.

Applications for halophiles have also been found in the field of astrobiology. Several lander missions in Mars have been conducted and yielded new findings. The Phoenix Lander mission in particular reported direct observation of ice in the surface of martian soils. However the most important

discovery was the detection of perchlorate (ClO^-_4) using the MECA (Microscopy, Electrochemistry, and Conductivity Analyzer) instrument as described by Hecht et al. (2009) and Kounaves et al. (2009). Chevrier et al. (2009) tested the capabilities of two perchlorate solutions as possible liquid brines at the Phoenix landing site. The evaporation rates of these solutions were measured modeling Martian conditions, the results suggested that these solutions can possibly exist as liquid brines during the day in the Martian summer and exhibited low evaporation rates. Furthermore, perchlorate increases the possibility of the formation of stable brines (Rennó et al., 2009).

It is important to understand if these liquid brines combined with other conditions in Mars are enough to sustain life in them. Perchlorates are toxic and strong oxidants, suggesting that their presence may inhibit life. To evaluate the possibility of life of halophilic prokaryotes in Mars, Oren (2013) tested the adaptability of several halophilic prokaryotes in the presence of different concentrations of perchlorate. In this experiment, all archaeal strains that were used grew well in media containing up to 0.4 M perchlorate. Additionally, a strain of *Haloferax volcanii* exhibited growth in 0.6 M perchlorate. The most interesting result, however, was the fact that strains of the haloarchaea *Haloferax denitrificans, H. mediterranei, H. gibbonsi, Haloarcula marismortui* and *H. vallismortis* could use perchlorate as an electron acceptor in cellular respiration. Thus, some halophilic prokaryotes can tolerate and even thrive in the presence of perchlorate. Moreover, other adaptabilities in halophiles have been reported. Fendrihan et al. (2009) exposed the haloarchaeon *Halococcus dobrowskii* to ultraviolet radiation and encountered the archaeon to be resistant to high levels of UV radiation. In another study, Kotteman et al. (2005) described the ability of *Halobacterium* strain NRC11 to tolerate exposure to high levels of desiccation and gamma irradiation. Finally, a strain of *Halorubrum chaoviator* has been known as the traveler of the void due to its ability to tolerate high vacuum and exposure to high radiations (Mancinelli et al., 1998).

The resistance of these organisms to radiation is due to the presence of carotenoid pigments (Litchfield, 1998). Moreover, halophilic archaea possess several strategies to overcome adverse conditions. For example, these organisms are characterized for the presence of acidic aminoacids in their proteins as well as high intracellular K^+ concentration. This high concentration of K^+ allows these ions to interact with the negative aminoacids and form a hydration barrier protecting their cells from desiccation (Litchfield, 1998). With these findings, it has been established that halophilic archaea commonly found in hypersaline environments can be used as model organisms for astrobiology due to their ability to withstand adverse conditions.

In conclusion, halophiles possess a wide variety of applications from industrial and clinical use to being study subjects for modeling possible conditions of life on other planets. Due to these facts, a great number of studies have been carried out and many more are yet to be performed.

2.2 Metagenomics and Applications in hypersaline environments

One molecular technique that has possibly defined the last decade of molecular biology is known as metagenomics. As opposed to studies focused on the 16S rRNA gene, which assess diversity based on only a single gene, this method is defined as the direct analysis of all the genes contained within an environmental sample (Thomas, 2012). The main challenge when it comes to understanding microbial diversity is the fact that more than 99% of microorganisms present in any given sample cannot be studied using culture-dependent methods and therefore are not available for biotechnological uses or research. These numbers indicate that the great majority of microbial species has never been described and as such it will remain that way until new culture technologies are developed (Streit & Schmitz, 2004). Metagenomics partially solves these problems because it provides access to the functional gene diversity in any given sample. Essentially, with metagenomics you can obtain genetic information on potentially novel biocatalysts, phylogeny for uncultured organisms

and evolutionary profiles of community function and structure. The rapid decrease in cost of Next Generation Sequencing (NGS) technologies has accelerated the development of these methods. Auld et al. (2013) compared both traditional culture-dependent methods and direct environmental sequencing methods when characterizing the bacterial community in an acid mine drainage pond. In their comparison, the methods were complementary. In general, both techniques indicated a very similar community structure. However, general species abundance could only be determined from direct sequencing results and provided a more comprehensive analysis of the community.

Various metagenomic studies have been undertaken in hypersaline habitats. One of the first reports was performed by Legault et al. (2006) on a saturated NaCl crystallizer in Santa Pola, Spain. The most dominant organism in the community was shown to be the square haloarchaeon *Haloquadratum walsbyi.* Further studies took place in these same salterns in two crystallizers of intermediate salinity with 13% NaCl (w/v) and 19% NaCl (w/v) and two crystallizer ponds of 33% NaCl (w/v) and 37% NaCl (w/v) (Fernández et al., 2014; Ghai et al., 2011). Results confirmed the overwhelming dominance of *Haloquadratum walsbyi* in saturation level salinities while also revealing new microbial groups in the process including a group of low GC *Actinobacteria*, an Euryarchaeon with the lowest reported GC content in its group as well as possible new *Bacteroidetes* aside from *Salinibacter* that may be abundant in these environments. Fernández et al. (2014) also reported simplified carbon and nitrogen cycles as well as the extensive use of light by bacteriorhodopsins.

Metagenomic analyses have also been conducted in Lake Tyrell in Australia. Narasingarao et al. (2012) revealed previously unknown lineage of archaea for which the class *Nanohaloarchaea* was proposed. Unique characteristics of this novel group include small genome size (~1.2 Mbp), reduced cell size (~0.6 μm), low GC content, and atypical carbohydrate metabolism pathways. Community structure was determined based on environmental 16S rRNA sequence data revealed that "*Nanohaloarchaea"* comprised 10-25% of the

archaeal community in Lake Tyrell while typical representatives found in other hypersaline environments such as *Halorhabdus, Haloquadratum,* and *Halorubrum* were also encountered. In another study, Podell et al. (2014) employed a metagenomic approach to study seasonal succesion in Lake Tyrell in a 2-year period. *Haloquadratum* and *Nanohaloarchaea*-related sequences predominated in the summer while *Halorubrum* and *Haloarcula* predominated in the winter. High levels of potassium, magnesium and sulfide were correlated positively with *Haloquadratum* abundance while these same ion concentrations were correlated negatively with *Halorubrum, Haloarcula, Halobaculum, Halonotius,* and *Salinibacter*-related sequences. On the other hand, *Nanohaloarchaea* and *Halorhabdus* related sequences were not correlated with ionic composition. Furthermore, the presence of *Haloquadratum* was correlated negatively with the rest of the prokaryotic community. These results suggest competition for similar ecological niches in the summer.

The Atacama Desert has also been a target of metagenomic studies. Culture-independent methods carried out have demonstrated that microbial halite communities in the Atacama Desert are predominated by archaea in the *Halobacteriaceae* family as well as cyanobacteria (de los Ríos et al., 2010). Crits-Cristoph et al. (2016) provided a metagenomic analysis of endolithic halite microbial communities from the Salar Grande. Results reaffirmed the dominance of halophilic archaea in these types of communities. Furthermore, functional gene composition showed the presence of RubisCO type I genes associated to cyanobacteria as well as RubisCO type III genes associated to halophilic archaea. Finally, a genome for a novel representative of *Nanohaloarchaea* was unveiled in this study.

These studies have shown that metagenomic methods in hypersaline environments successfully provide more comprehensive answers to community composition as well as possible functions within these communities. Discovery of novel microbial groups has shifted our understanding of hypersaline environments towards new directions (Ventosa et al., 2015). New techniques will become

available as more data are released and will provide more tools towards characterizing novel taxa as well as novel biocatalysts (Ventosa et al., 2015).

2.3 Sequencing Technologies

The first DNA sequencing technology was developed by Sanger and Coulson in 1977 and transformed molecular biology and genetics by providing a method of determining complete genes and entire genomes (Kircher & Kelso, 2010; Schuster, 2008). To this day, Sanger sequencing is still relevant and has been used in systematic projects (Hajibabae et al., 2007; Shokralla et al., 2012). It is considered to be the gold standard of sequencing due to its low-error rate and long read length. However, Sanger technologies are limited to being subject to extensive library preparation and cloning bias (Thomas et al., 2012). Moreover, it is also restricted to sequencing a single individual and therefore unable to be employed in environmental samples (Shokralla et al., 2012). Due to this, metagenomic shotgun sequencing has shifted to Next Generation Sequencing (NGS) technologies.

Of the NGS technologies, perhaps the two most recognized are Roche's 454/Pyrosequencing and Illumina/Solexa. The first NGS technology available was 454, a PCR-based technique which was introduced in 2005 (Shokralla et al., 2012). This technique uses an emulsion PCR as an amplification step where random DNA fragments attached to microscopic beads are clonally amplified. These beads are subsequently inserted into the wells of a picotitre plate and individually sequenced in parallel. Nucleotides complementary to the template DNA strand are added one at a time (Kircher and Kelso, 2010; Thomas et al., 2012). If successfully incorporated, a pyrophosphate molecule is released which initiates a series of reactions that conclude in the production of light by the luciferase enzyme. Light intensity is directly proportional to the amount of nucleotides incorporated and is recorded by a charge-coupled device which then translates the actual sequence of the template strand. Pyrosequencing techniques provide an average read length between 600-800 bp which facilitates

annotation. However, as pointed out by Niu et al. (2010), emulsion PCR produces artificial replicate sequences, which in turn can provide false estimates of gene abundance. Additionally, light intensity when sequencing homopolymers can be difficult to correlate to nucleotide positions, which in turn leads to insertion/deletion errors (Thomas et al., 2012).

Illumina technologies were introduced in 2007 and are based on a sequencing-by-synthesis process (Shokralla et al., 2012). Random DNA fragments are held in the solid surface of a flow cell that is divided into eight separate lanes. These random fragments are then bridge amplified, which results in clusters with identical DNA fragments. After amplification, all four nucleotides marked with fluorescence compete and the complementary base is added to the template strand. After addition, a light signal is emitted which is read by the sequencer. Illumina machines are known for their high capacity and produce a large amount of data (Thomas et al., 2012). One of the advantages of Illumina technologies is its low cost of around~$50 USD per Gb. However read lengths using the HiSeq2000 sequencer yield ~150 bp and therefore rendering them too short for functional annotation. Therefore, assembly of these reads is recommended. Furthermore tail-ends of the reads are known for their high error rates, however trimming of these reads is suggested for error correction.

Other NGS technologies are available including the Applied Biosystems SOLiD technology that produces the lowest error rate of any NGS technology. However, read length is around 50 bp making this technology unreliable for functional annotation and highly complicated for assembly (Thomas et al., 2012). Moreover, it is also subject to substitution errors and long run times (Metzker, 2010). PacBio is another technology currently on the market based on single molecule real time detection. Although it offers improved read-length, its accuracy is around 85% (Rhoads and Au, 2015).

After NGS, two generations of sequencing have passed and we are currently into the 4th generation of sequencing. Of the most recognizable technologies in this generation is the Nanopore-based sequencing where DNA

strands are passed through a nanometer-sized pore embedded in either a biological membrane or a solid-state film. The pore is divided into cis and trans compartments containing conductive electrolytes. Under voltage, these electrolyte ions are moved through the pore and consequently generate an ionic current signal. Interruption of this signal by negatively charged DNA can lead to a statistical calculation by analyzing the amplitude and duration of said interruption (Feng et al., 2015). The Oxford Nanopore Technologies MinION is one of the first commercially available sequencers using the nanopore-based techniques. The device offers kilobases of read length and also portability which facilitates real-time on-site sequencing at a low capital cost (Laver et al., 2015). The main challenge of nanopore-based sequencing is its high error rate (Feng et al., 2015). First sequencing runs of the MinION sequencer yielded error rates of 90% (Mikheyev & Tin, 2014) although recent runs such as that of Laver et al. (2015) reported a mean error rate of 38.2%. Due to this, it is still very unreliable, however optimization of these technologies looks promising and may yield very reliable results in the future.

2.4 Bioinformatics Tools

After sequencing, processing the millions of reads for data analysis is the next step. To obtain a functional description of the community, direct annotation of those genes can be performed without the need for assembly. However, if the aim is to recover full genomes from organisms in the community as well as full length coding sequences, assembly of these reads into contigs is recommended (Thomas et al., 2012). There are two main methods of assembly: reference based assembly and *de novo* assembly.

Reference-based assembly uses efficient computational resources and can generally be performed on laptop computers. Software packages available include Newbler, AMOS, MIRA and CAR among others (Lu et al., 2014; Thomas et al., 2012). However, challenges for metagenome assembly using reference-based assembly include lack of reference genomes. Moreover, differences

between the true genome of the sample and the reference including insertions, deletions or polymorphisms can lead to an assembly with fragmentation errors. Therefore, for metagenome assembly, *de novo* assembly is usually preferred.

De novo assembly usually requires larger computational resources. A great range of assemblers has been made available based on the need for assembling single genomes. Most assemblers available employ a de Bruijn graph method (Miller et al., 2010). Using this method, assemblers build their core data structure using different variations of a K-mer graph. This graph contains short nucleotide sequences known as k-mers which are connected by edges. A transitional approach is used to connect these k-mers; for example, 4-mer TGAC can connect to GACC (Luo et al., 2013). Of these assemblers, Velvet (Zerbino & Birney, 2008), SOAPdenovo (Li et al., 2008), SPAdes (Bankevich et al., 2012) and IDBA (Peng et al., 2010) have been employed successfully in published literature.

However, when applying assembly tools to metagenomes, variation in species at the genomic level can lead to overlooks in assembly and misinterpret species abundance. Solutions to this problem have been proposed in the form of metagenomic assemblers such as MetaVelvet (Namiki et al., 2012), IDBA-UD (Peng et al., 2012), and MetaSPAdes (Nurk et al., 2013), among others. The approach used in metagenomic assemblers is to build a sub k-mer graph within the entire k-mer graph that represents related genomes. Using this approach, better assemblies have been provided when compared to traditional de Bruijn assemblers (Thomas et al., 2012). Nevertheless, metagenomic assemblers are still at an early stage and more tools will be developed in the future to improve metagenome assembly.

After obtaining an assembly, sorting these contiguous sequences into groups that may represent individual genomes in a process known as binning is suggested. Various binning algorithms have been developed including Phylopythia (McHardy et al., 2007), MaxBin (Wu et al., 2014) and MetaBAT (Kang et al., 2015). Post-assembly binning can lead to the generation of partial genomes from novel and possibly uncultured organisms. Ghai et al. (2012) and

Narasingarao et al. (2012) have employed binning methods and uncovered novel and as of yet uncultured archaeal genomes. However careful consideration should be taken when validating a genome bin as contaminating fragments can lead to into false genomic bins.

Finally, annotation of metagenomes is typically performed by first identifying features of interest (genes), also known as feature prediction, and then assigning putative function of these features (functional annotation). Feature prediction typically consists of labeling sequences as genes or genomic elements. Several software packages such as MetaGeneAnnotator (Noguchi et al., 2008) and Orphelia (Hoff et al., 2009) have been developed for feature prediction. Functional annotation is usually performed using annotation pipelines of which MG-RAST (Meyer et al., 2008) is the most commonly used. Sequences in MG-RAST are aligned to reference databases such as KEGG (Kanehisa et al., 2003), eggNOG (Muller et al., 2010), COG (Tatusov et al., 2003) and SEED subsystems. One major challenge in annotation is the fact that only 20-50% of metagenomic sequences can be annotated (Thomas et al., 2012). This leaves many sequences that cannot be mapped to a known sequence or ORFans. Causes for ORFans can be related to errors in gene prediction, being a real gene encoding for unknown biochemical functions, or lack of sequence homology with known genes but possessing structural homology with known proteins, thus representing known proteins or folds. Annotation of ORFans will most likely be improved as new characterizations of unknown proteins become available.

In conclusion, several tools exist for metagenomic analyses, however many more are yet to be developed. With the advancement and rapid cost decrease of sequencing technologies, metagenomic datasets will be more complex and comprehensive. Therefore, development of new software to correctly analyze and visualize data obtained will be needed. Moreover, optimization of bioinformatics tools will lead to answering fundamental questions in microbial ecology through the use of metagenomics. Due to the slow development of bioinformatics tools, there exists a limited knowledge of microbial community processes in hypersaline

environments. This study intends to uncover some of the answers to fundamental questions such as carbon, nitrogen, and sulfur processes in these environments using bioinformatics.

3. Objectives

1. Make a temporal metagenome study by through the analysis of three metagenomic libraries from the crystallizers supplied by the Fraternidad lagoon in the Cabo Rojo salterns using Illumina sequencing.

2. Annotate genes related to environmental processes carried out by the microbial community.

3. Recover individual genomes of putative novel species from the metagenomic dataset using binning tools.

4. Establish comparisons in terms of functional and microbial diversity with other metagenomes available from solar salterns around the world.

4. Metabolic and Microbial Diversity in the Cabo Rojo Solar Salterns

4.1 Summary

Three independent samplings of water samples from the Fraternidad crystallizer ponds were carried out on December (2014), April (2016) and July (2016). Each water sample was subsequently filtered across two membrane filters the first one of 5 μm which recovered eukaryotic cells and the second one of 0.22 μm which recovered microbial cells. DNA extraction and sequencing was carried out on the 0.22 μm membrane and yielded three metagenomic paired end libraries of high quality based on Phred score consisting of 16 GB of data each. Assembly of each paired end library produced metagenomes of about 400,000 contigs. Taxonomic composition was evaluated and resulted in a diverse community with 12 phyla encountered. The phylum *Euryarchaeota* predominated in the three metagenomes, however when assessing predominance at genus level, the community was predominated by a different genus across all three samples. Anthropogenic impact as well as precipitation events may be contributing factors to the changes observed in the microbial community composition. Furthermore, functional annotation was carried out in order to detect genes related to ecological processes such as carbon, nitrogen and sulfur cycles carried out in the crystallizer ponds. Sequences matching essential pathways in carbon, nitrogen, sulfur and phosphorus cycles were found along with hits from organisms known to perform these pathways. Results suggest that cyanobacteria are contributing to primary production. The presence of microorganisms involved in nitrogen fixation, ammonia oxidation, sulfate reduction, sulfur oxidation and finally phosphate solubilizing were also detected. Using a metagenomic approach, a large diversity previously unreported in marine salterns at 34% (w/v) NaCl has been uncovered. Moreover, analysis of gene composition highlights the importance of the microbial community in the biogeochemical cycles. Ionic composition analyses in crystallizer water along with further sampling and transcriptomic approaches may give us more information on other processes

being performed including the presence of ammonia oxidizing archaea in the Cabo Rojo salterns as well as on the existence of putative undescribed species.

4.2 Materials and Methods

4.2.1 Sample processing

Sampling (50L per sample) was carried out in December (2014), March (2016) and July (2016) in the crystallizer ponds from the Cabo Rojo salterns system served by the Fraternidad Lagoon. Each 50L sample of saltern water was differentially filtered using a Millipore® pressurized filtering system consisting of two membrane filters of different pore sizes. The first membrane possessed a pore size of 5.0 µm which was intended to retain eukaryotic cells were whereas the second membrane with a pore size of 0.22 µm was used for the collection of prokaryotic cells. DNA extraction was performed on cells retained in the 0.22 µm membrane using physical chemical methods described previously by Martín-Cuadrado et al. (2007). Concentration and purity of DNA were measured using a Nanodrop™ spectrophotometer. Furthermore, a 0.8% agarose gel electrophoresis was carried out in order to corroborate DNA quality before sequencing.

4.2.2 Sequencing and Data Processing

DNA sequencing was performed using Illumina HiSeq 2500. Library preparation and sequencing was carried out by Molecular Research DNA (MR DNA) facility in Shallowater, TX. Sequencing reads obtained were quality checked using FastQC (Andrews, 2010). Low quality reads were trimmed for assembly using BBDuk. After quality trimming, taxonomic profiles of the metagenomic reads were determined using Phylosift (Darling et al., 2016) in order to obtain a taxonomic profile of the microbial community. Afterwards, assembly of remaining reads was performed using MetaSPAdes assembler where quality of metagenome

assemblies were compared based on N50 values (median length of contigs), total contigs obtained, as well as the largest contig.

4.3 Results and Discussion

4.3.1 Sampling, DNA Extraction, Purity and Sequence Quality

Salinity and temperature for all three samples was measured *in situ.* Salinity was 34% NaCl (w/v) in all three samples. Moreover, the temperature recorded in the samples was 31.1 ˚C.

Table 1: Concentration and purity of DNA extracted from samples MFF1 (December 2014), MFF2 (April, 2016), MFF3 (July, 2016).

Sample	Concentration (ng/μL)	260/280 Ratio	260/230 Ratio
MFF1	161.8	1.61	0.63
MFF2	230.0	1.54	0.58
MFF3	136.6	1.46	0.51

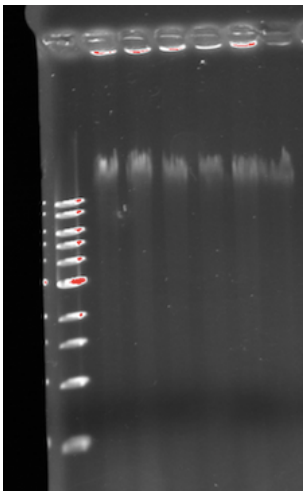

Figure 1: Agarose gel electrophoresis of replicate samples of metagenome DNA. From left to right (1 kb ladder, MFF1, MFF2, MFF3, MFF1, MFF2, MFF3).

DNA concentration and purity of each extraction was measured using a NanoDrop™ spectrophotometer (Table 1). Contamination of proteins and extracellular substances is indicated by the 260/280 absorbance ratio. A ration of 1.8 is ideal for high quality DNA. However, saltern water is a saturated salt solution containing bacteriorhodopsin and extracellular polymeric substances which can be inhibitors for downstream applications such as PCR and sequencing. In turn this can reduce the 260/280 ratio. The average of all three samples in the 260/280 ratio is around 1.54 which coincides with that from other studies in similar environments (Bey et al., 2011). The 260/230 ratio is comparatively lower in the with an average of 0.63 range. A 260/230 ratio is desirable in order to avoid interference from RNA. However, other studies have used similar ratios and have shown no inhibitory effects in downstream reactions (Solomon et al., 2016). Metagenomic DNA was further analysed by 0.8% agarose gel electrophoresis with stained with ethidium bromide. Although this method does not give us an absolute measurement for DNA quantity, it is still useful to analyze RNA contamination, DNA degradation and average size of extracted DNA (Bag et al., 2016).

Sequencing of metagenomic DNA from samples MFF1, MFF2, and MFF3 produced 3 paired-end libraries with read length ranging from 35 to 251 base pairs (bp) which were subsequently assessed for sequence quality using FastQC.

Table 2: Sequencing results for MFF1, MFF2, and MFF3

Sample	Number of Reads	GC Content	Read length
MFF1	29,432,758	56%	35-251 bp
MFF2	41,746,817	57%	35-151 bp
MFF3	40,634,465	58%	35-151 bp

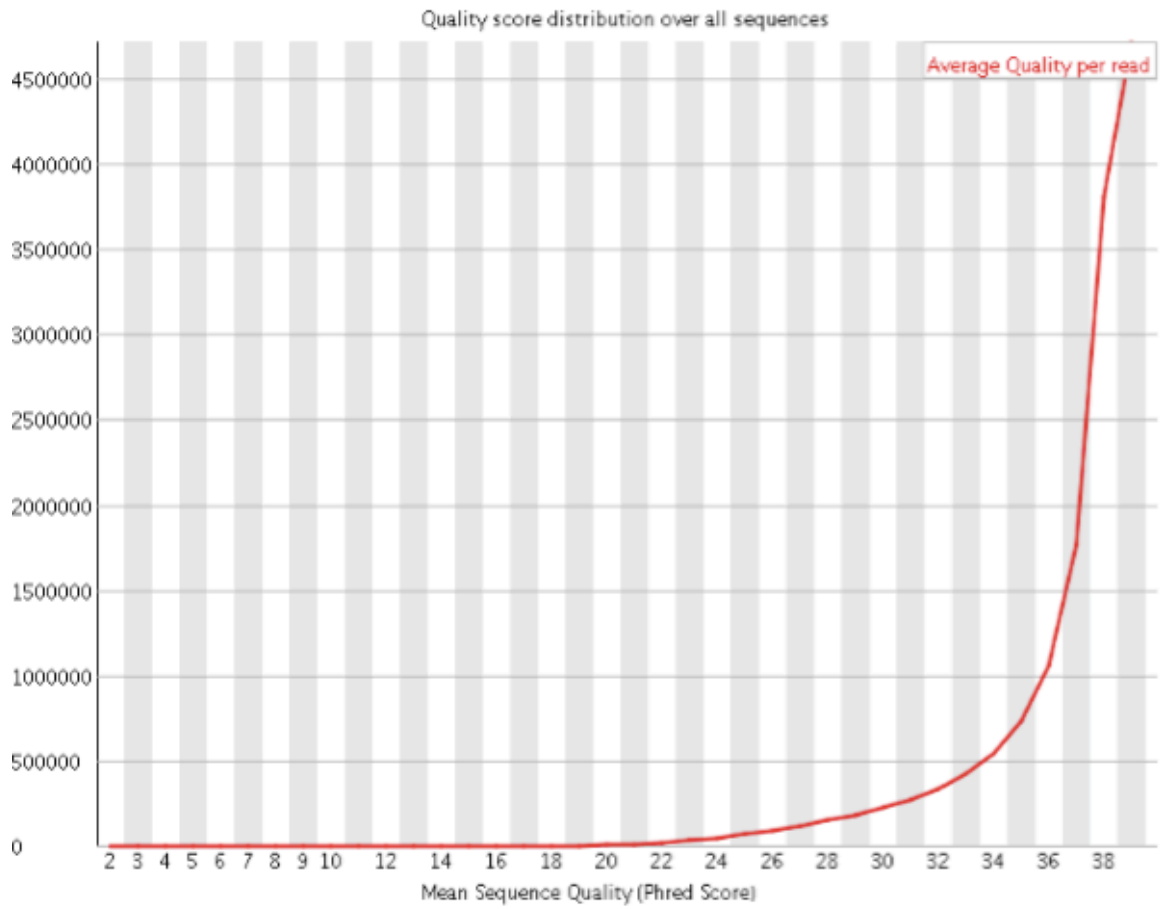

Figure 2: Phred quality score distribution across all sequences for MFF1. Most of reads exhibited a Phred quality score of 34 or higher.

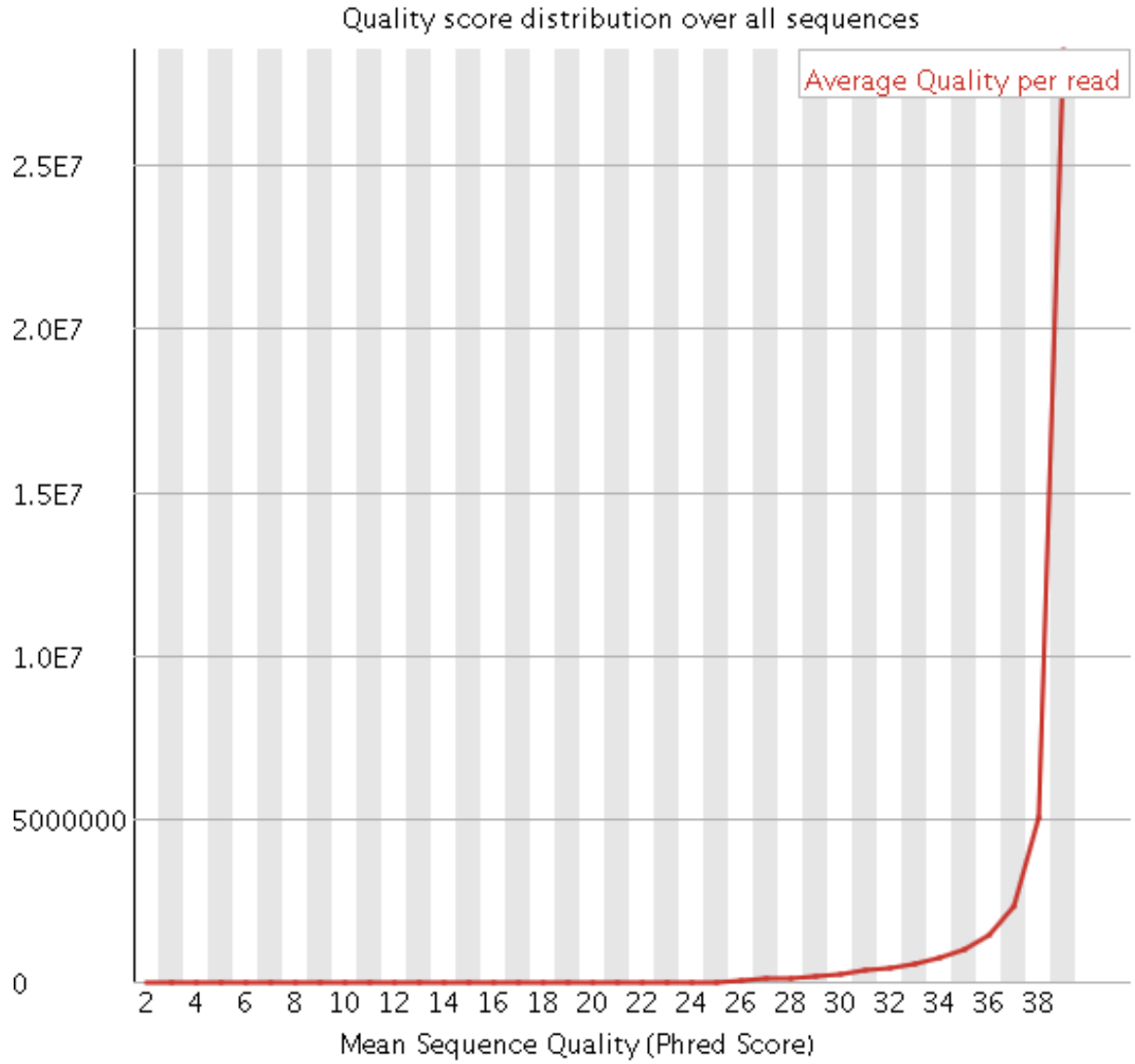

Figure 3: Phred quality score distribution across all sequences for MFF2. Most of reads exhibited a Phred quality score of 34 or higher.

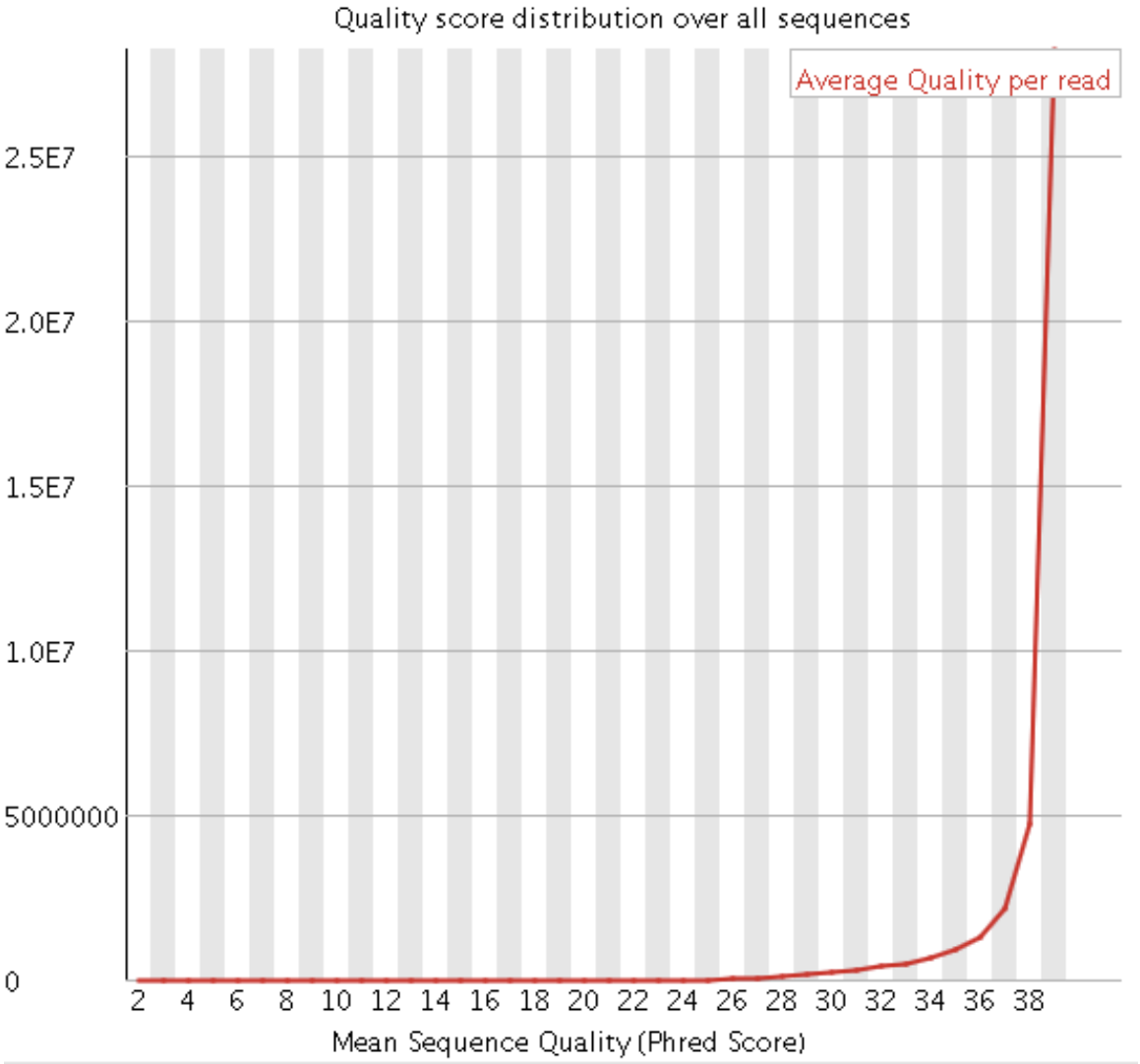

Figure 4: Phred quality score distribution across all sequences for MFF3. Most of the reads exhibited a Phred quality score of 34 or higher.

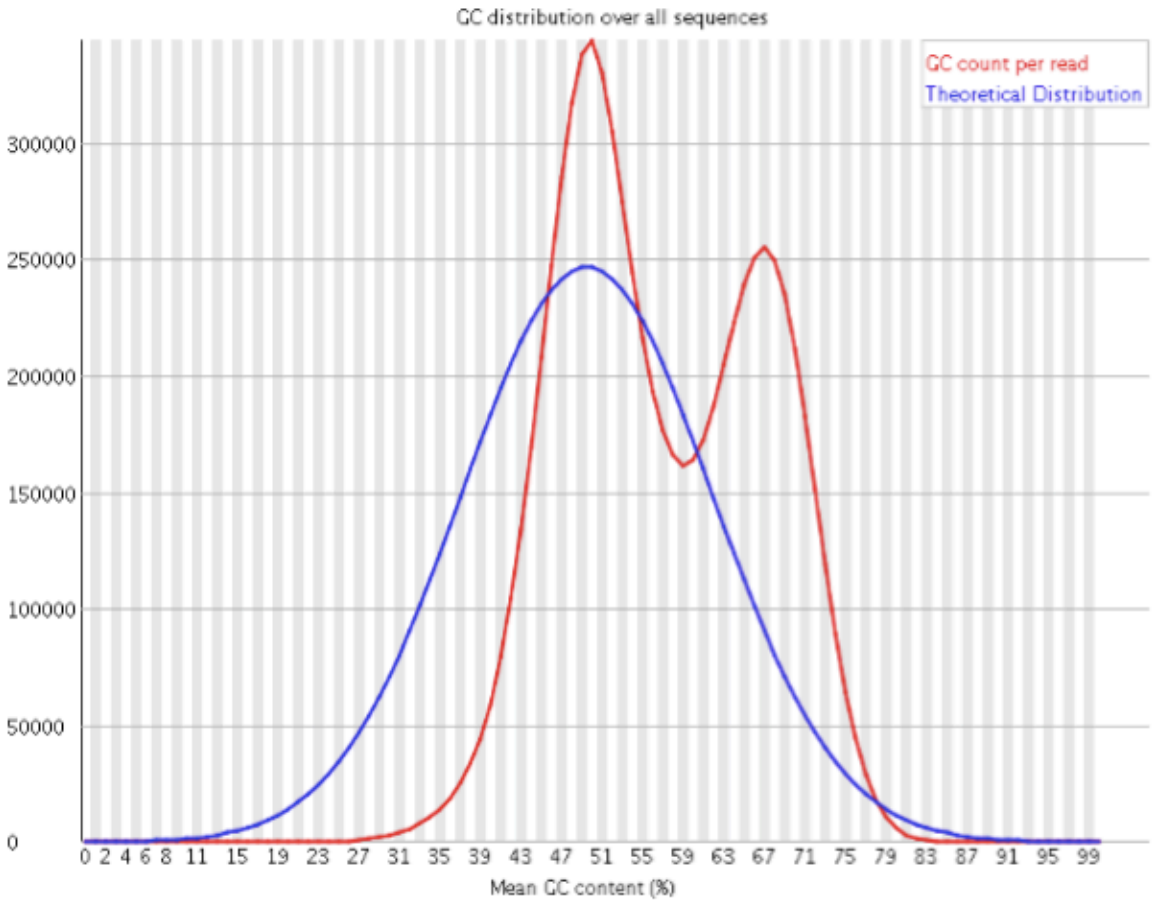

Figure 5: GC content distribution across all sequences for MFF1. A bimodal peak distribution is observed at 47% and at around 67%. The majority of the reads align with the peak at 47%.

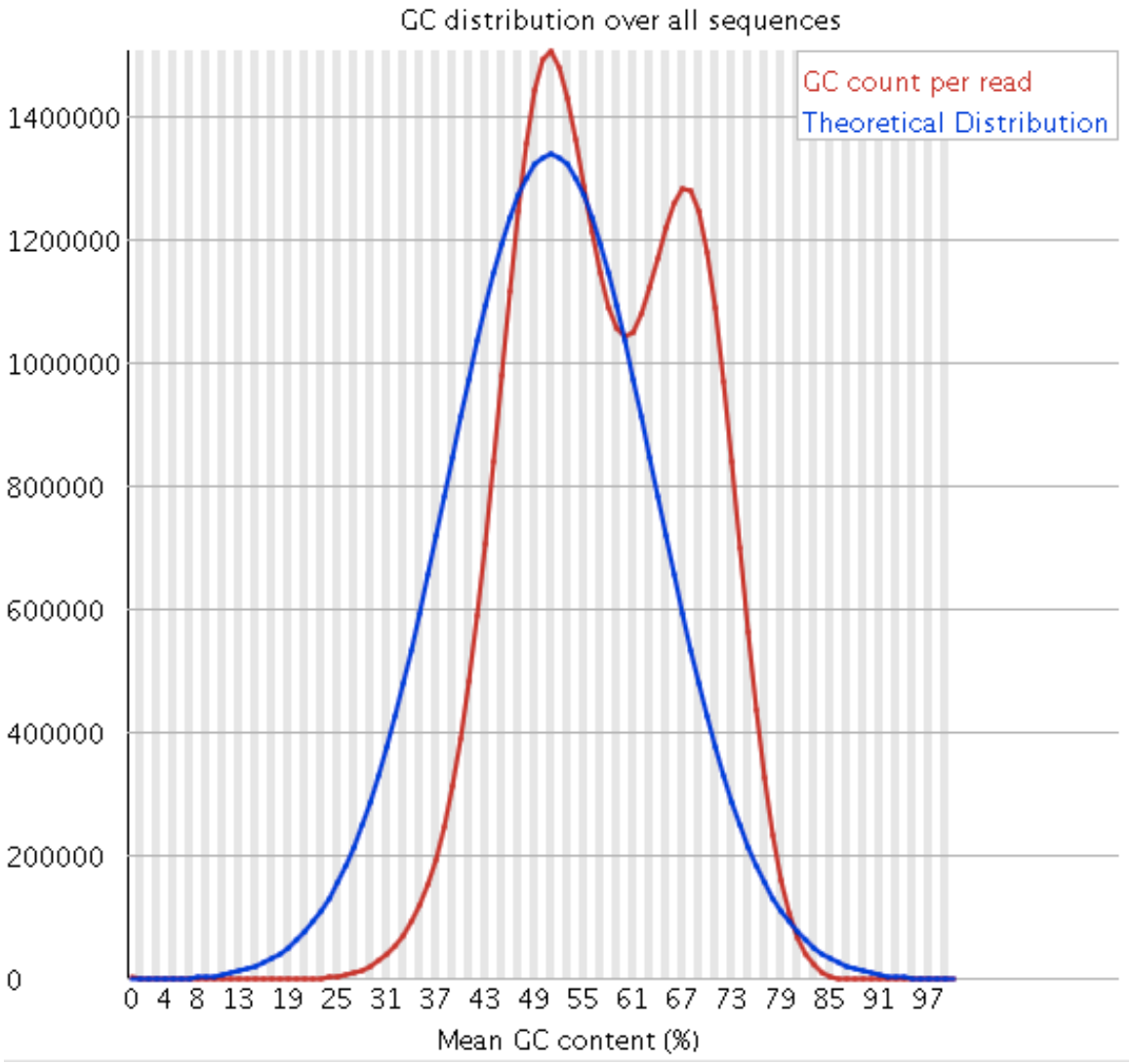

Figure 6: GC content distribution across all sequences for MFF2. A bimodal peak distribution is observed at 47% and at around 67%. The majority of the reads align with the peak at 47%.

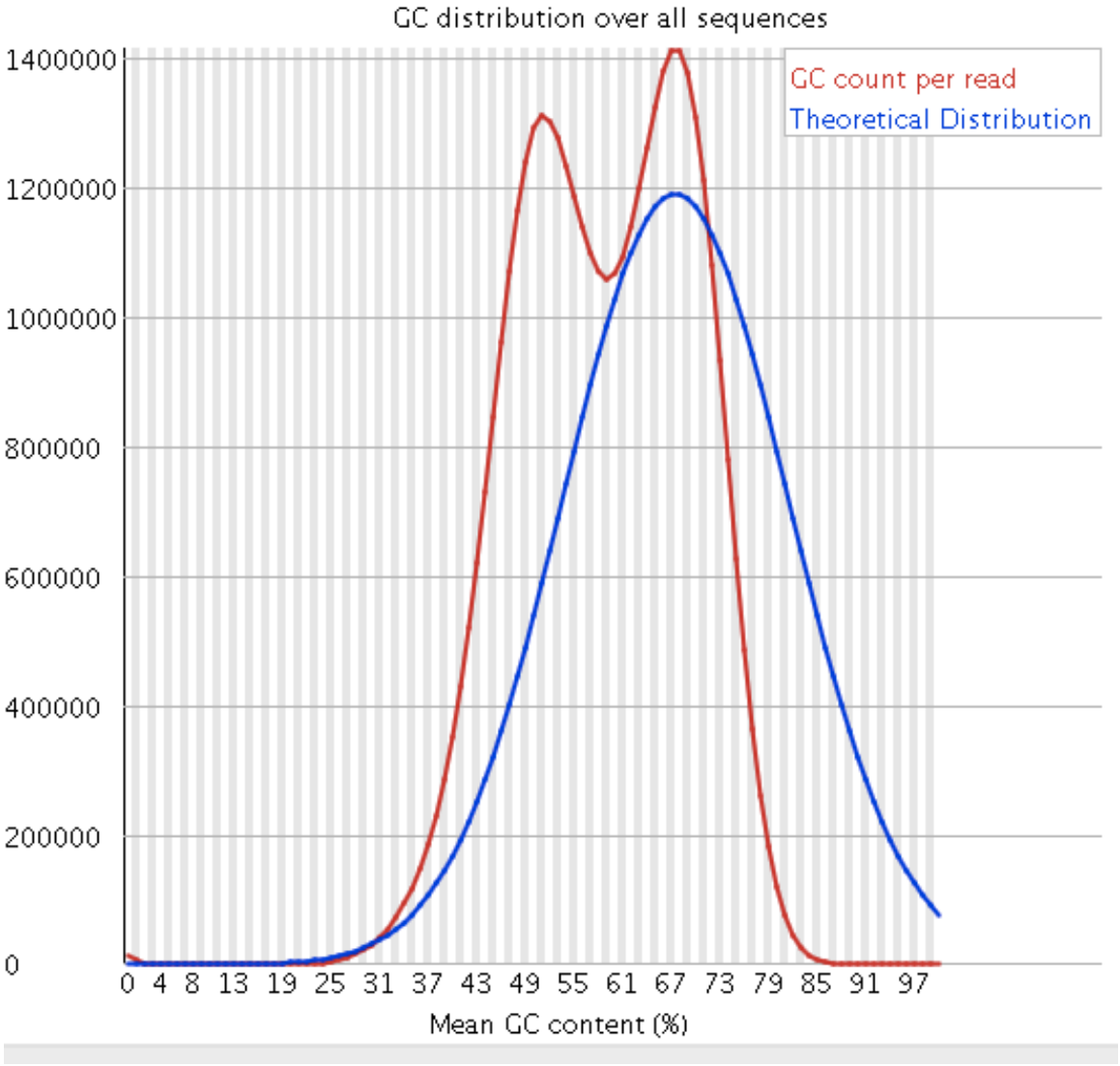

Figure 7: GC content distribution across all sequences for MFF3. A bimodal peak distribution is observed at 47% and at around 67%. The majority of the reads align with the peak at 67%.

Sequence quality results as shown in **Figures 2, 3** and **4** were satisfactory with very few low-quality reads and no unusable sequences. Furthermore, sequence quality was shown not to be hindered by RNA interference as judged by the 260/230 ratio of optical density readings. About 1 million reads were quality trimmed for assembly, however this was not shown to significantly improve quality

statistics. Furthermore, sequence quality was shown not to be hindered by RNA interference in the 260/230 ratio.

The bimodal distribution of the GC content plots observed in Figure 1 is characteristic of hypersaline environments has been previously described by Ghai et al. (2011) when studying aquatic hypersaline habitats in Alicante, Spain. The 49% GC peak is most likely associated with the square haloarchaeon *Haloquadratum* which is known to have an average GC content of 47% (Legault et al., 2006). A second peak of about 67% is most likely attributed to *Salinibacter* whose genome possesses an average GC Content of 66.52% (Mongodin et al., 2005) as well as other representative haloarchaeal genera such as *Haloferax, Halorubrum* and *Haloarcula* who possess GC content in the 65-67% range. The presence of these genera in our metagenomic samples correlates with these peaks.

4.3.2 Assembly results

Different assemblers were compared such as MetaVelvet, MetaSPAdes, SOAPdeNovo and IDBA-UD. Of these assemblers, MetaSPAdes yielded the best assembly and was subsequently used to assemble metagenomes for MFF1, MFF2 and MFF3. Assembly statistics are listed below in table 3.

Table 3: Assembly statistics for MFF1, MFF2, and MFF3.

Sample	Number of Scaffolds	N50	Longest Contig
MFF1	318,469	3,888	619,112
MFF2	420,402	4,748	469,957
MFF3	379,415	4,854	388,630

Modern DNA sequencing technologies cannot produce the entire sequence of a genome. Instead, these technologies produce millions of reads with different lengths ranging from tens to up to thousands of base pairs (Gurevich et al., 2013). These reads are sampled from different parts of a genome. Validation

of genome assemblies has been the subject of debate since the first genome assemblies were published. Most parameters used are related to assembly contiguity, where number of contiguous sequences obtained as well as longest sequence are compared. Lower number of contigs as well as high contig length is ideal. These metrics attempt to evaluate how far the assembly is from the ideal of 1 contig per chromosome (Olson et al., 2017). However, more robust measures are: the N50 values, defined as the minimum contig length in the set of contigs that comprise over half of the assembly, and the number of open reading frames (ORFs) or ORFs/Mb. Assessment software such as QUAST (Gurevich et al., 2013), GAGE (Salzberg et al., 2011), and Plantagora (Barthelson et al., 2011) have been used to compare assembly quality.

In the case of metagenomics, the sequencing reads obtained are sampled from different genomes in an environmental sample. Therefore, the data received are usually enormous, unprecise, and contain numerous fragments from a varying number of species with different abundances (Mikheenko et al., 2016). Therefore, these problems pose a challenge when it comes to assembling these reads into contigs for further analyses and assessing their quality. Most assembly assessment software are not designed for evaluating metagenomic assemblies. Furthermore, the lack of reference genomes complicates matters. However, MetaQUAST solves some problems in evaluating quality of metagenome assembly by possessing features such as using an unlimited number of reference genomes, automated species content detection, and detection of interspecies misassemblies (Mikheenko et al., 2016). **Table 2** details assembly statistics using MetaQUAST for the three metagenomes where the millions of reads were condensed to hundred thousand contigs, a significant reduction. Additionally, N50 values obtained surpass values obtained in other metagenomic studies performed in hypersaline environments (Crits-Cristoph et al., 2016). In conclusion, the number of contigs combined with N50 values and compared with other studies suggest a competent assembly of all three metagenomic paired end libraries.

4.3.3 Taxonomy

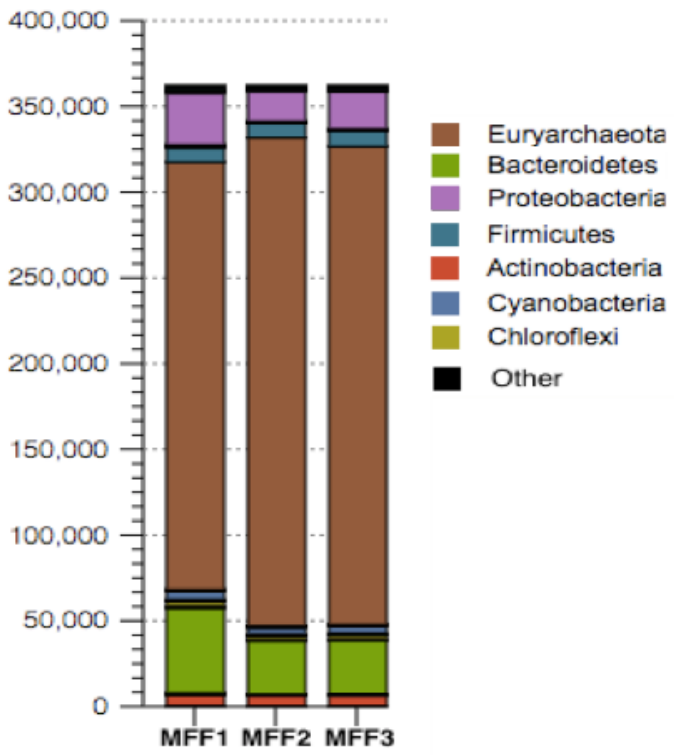

Figure 8: Taxonomic hits by phylum. Each slice indicates number of reads with predicted proteins and ribosomal RNA genes annotated to the indicated phylum. The phylum *Euryarchaeota* predominates in all three samples followed by *Bacteroidetes, Proteobacteria, Firmicutes, Actinobacteria, Cyanobacteria,* and *Chloroflexi.* Other groups present with less than 1% of sequences include *Nanoarchaeota, Chrenarchaeota* and *Chlorobi,* among others.

Table 4**:** Taxonomic composition at genus level for metagenomic reads. The percentage as well as the number of reads with predicted proteins and ribosomal RNA genes annotated to the genus are shown.

MFF1		MFF2		MFF3	
Genus	**Abundance**	**Genus**	**Abundance**	**Genus**	**Abundance**
Salinibacter	22.26 % (109,068)	*Halorubrum*	11.09 % (105,168)	*Halogeometricum*	11.01 % (88,662)
Haloquadratum	12.92 % (63,301)	*Halogeometricum*	9.46 % (89,714)	*Halorubrum*	10.54 % (84,932)
Haloarcula	8.61 % (42,188)	*Halomicrobium*	7.17 % (67,968)	*Haloarcula*	6.97 % (56,141)
Halogeometricum	7.00 % (34,300)	*Haloarcula*	7.15 % (67,785)	*Halomicrobium*	6.61 % (53,238)
Halorubrum	5.52 % (27,034)	*Haloferax*	6.39 % (60,626)	*Salinibacter*	6.17 % (49,734)
Natronomonas	5.05 % (24,736)	*Salinibacter*	6.19 % (58,711)	*Halorhabdus*	5.63 % (45,374)
Halobacterium	4.03 % (19,742)	*Halorhabdus*	5.92 % (56,151)	*Haloferax*	5.58 % (44,988)
Halomicrobium	3.70 % (18,108)	*Halobacterium*	5.41 % (51,308)	*Haloquadratum*	5.46 % (43,997)
Haloferax	3.25 % (15,916)	*Haloquadratum*	5.27 % (49,964)	*Halobacterium*	5.12 % (41,230)
Halorhabdus	3.08 % (15,109)	*Haloterrigena*	4.44 % (42,134)	*Haloterrigena*	4.00 % (32,195)
Haloterrigena	1.79 % (8,769)	*Natrialba*	3.60 % (34,119)	*Natrialba*	3.36 % (27,046)
Natrialba	1.68 % (8,211)	*Halalkalicoccus*	2.36 % (22,403)	*Halalkalicoccus*	2.13 % (17,136)
Halalkalicoccus	1.28 % (6,281)	*Natronomonas*	1.90 % (18,044)	*Natronomonas*	1.75 % (14,124)

In terms of microbial diversity present, the phylum *Euryarchaeota* predominated in all three samples with more than 70% of the reads (**Figure 3).** This abundance is expected since this group contains the halophilic representatives from the *Archaea* domain. The *Bacteroidetes* group was second in abundance with about 20% of the reads. This group posesses one halophilic representative in *Salinibacter ruber.* The third most predominant group was the *Proteobacteria* with about 3% of the reads*.* The presence of *Proteobacteria* was also expected since it contains halotolerant bacteria such as the genus *Halomonas, Halovibrio, Rhodovibrio* among others (Ventosa, 2006). Furthermore, taxonomic hits abundance, as illustrated in Figure 3, reveals a diverse representation of prokaryotic phyla. This representation is markedly different from the results found by Ghai et al. (2011) and Rhodes et al. (2012), where only *Euryarchaeota*, *Bacteroidetes* and *Proteobacteria* were encountered at a similar salinity of about 34%. This study yielded 12 distinct prokaryotic phyla, consistently across all three samples, indicating a stable microbial diversity over time. Such a number of taxa is, to our understanding, previously unreported at this salinity.

However, when assessing community composition at the genus level **(Table 3)**, community structure was shown to vary across all three samples. The first metagenome presented a community predominated by the genus *Salinibacter* followed by *Haloquadratum,* while the second and third community were predominated by *Halorubrum* and *Halogeometricum,* respectively. Other studies in hypersaline environments, with a few notable exceptions, have demonstrated that *Haloquadratum* usually predominates in salinities of 30% and higher. In contrast, herein the microbial community was never dominated by the square archaeon and, in fact, in samples MFF2 and MFF3, *Haloquadratum* was the 6^{th} and 7^{th} most predominant genus, respectively. As Podell et al. (2014) demonstrated, abundance of *Haloquadratum* was correlated positively with high levels of potassium, magnesium and sulfide and correlated negatively with increase in microbial diversity. The data obtained could suggest that a similar competition for related ecological niches could be affecting microbial community composition. Moreover, the salterns are located in an area commonly visited by tourists and are frequently emptied and refilled. Therefore, anthropogenic impacts may also affect the community structure.

Data obtained from the National Weather Service's Advance Hydrologic Prediction Services (http://water.weather.gov/precip/) reveal rainfall in the area of the Cabo Rojo salterns in November 2014, March 2016 and July 2016 amounted to 20.3 cm of rain, 0.3 cm and 7.6 cm respectively. These events were shown not to have an effect on salinity as it remained constant across all three samples. Precipitation is one of the many ways microbes are dispersed into new habitats. Therefore, during a rain event, aquatic habitats can be recipients for new microorganisms. Peter et al. (2014) performed a culture-independent study of changes in marine microbial community structure during rainfall events with and without Saharan dust influence. Puerto Rico is an area affected largely by Saharan dust, and rain events. Results of the study showed that rain water influenced by Saharan dust contributed nutrients such as Ca^{2+}, Mg^{2+} and K^{+}. The presence of these nutrients influenced the predominance of Gammaproteobacteria over other

microbial groups. In the case of hypersaline environments, concentrations of Mg^{2+} and K^{+} have been correlated positively to *Haloquadratum* growth and negatively to *Halorubrum.* Sample MFF1 which was taken in November 30, 2014, exhibited a different community composition when compared to MFF2, sampled in March 30, 2016, and to MFF3 sampled in August 1, 2016. In MFF1, *Haloquadratum* has an abundance of 12.92% whereas in MFF2 and MFF3, *Haloquadratum* possesses abundances of 5.27% and 5.46% respectively. On the other hand, in MFF1, *Halorubrum* has a frequency of 5.52%, significantly lower when compared to MFF2 and MFF3 with abundances of 11.09% and 10.54%, respectively. Furthermore, the microbial community in November, in terms of abundance, differs completely from those sampled in March and August. Likewise, Tseng et al. (2013) demonstrated the effects of precipitation on aquatic microbial communities by assessing population structure over two years using metagenomics. Results uncovered a more diverse community during precipitation events, especially during typhoon season. Thus, studies have shown that precipitation events can influence microbial population structure.

The data on precipitation could suggest that the rainfall events in November could have contributed changes in ionic composition that directly affected microbial community structure. Ionic composition data obtained by Rodríguez-García (2016) in the Cabo Rojo salterns on June 2015 (1.0 inches of precipitation in the area) showed a predominance of chloride ions (230 g/L) when compared to magnesium (28.84 g/L) and potassium (11.22 g/L). Therefore, the decrease in magnesium and potassium ions are the possible reason as to why *Haloquadratum* has low abundance in samples MFF2 and MFF3. Due to Cabo Rojo, Puerto Rico's tropical location, Saharan dust influenced precipitation could have contributed more magnesium and potassium ions which favored growth of *Haloquadratum* in samples MFF1. Similarly, increase in potassium ions could favor abundance of "salt-in" strategists, such as *Haloquadratum* and *Salinibacter,* and could perhaps be a contributing factor in their abundance for the first metagenome. Furthermore, the high concentration of chloride ions can be a

contributing factor as to why *Halogeometricum* can predominate in these salterns but possesses lower abundance when compared to previous studies.

4.3.4 Functional Annotation

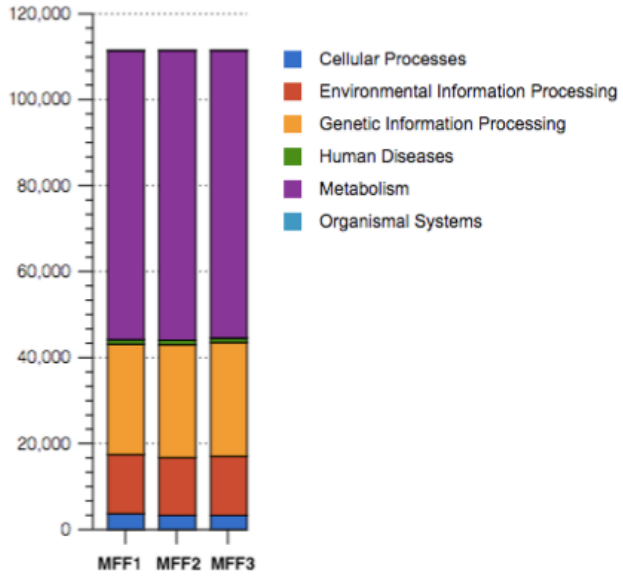

Figure 9: KEGG Onthology (KO) obtained from samples MFF1, MFF2 and MFF3. Genes related to metabolism predominate in about59% of the sequences in all three samples.

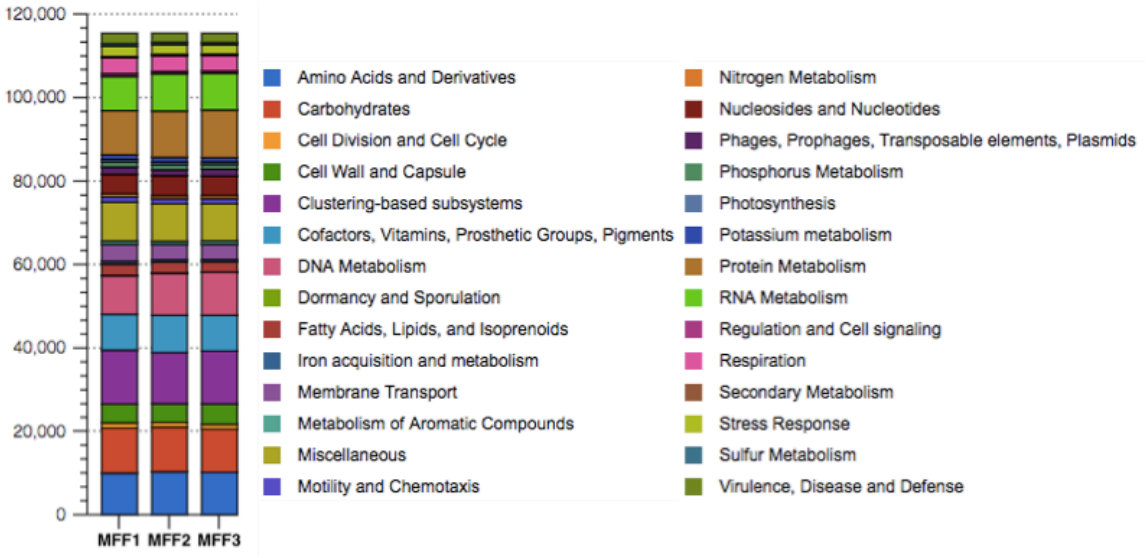

Figure 10: Subsystems distribution of annotated genes across all three metagenomes. Each slice indicates the number of reads associated with the subsystems function. Metabolism associated proteins predominate.

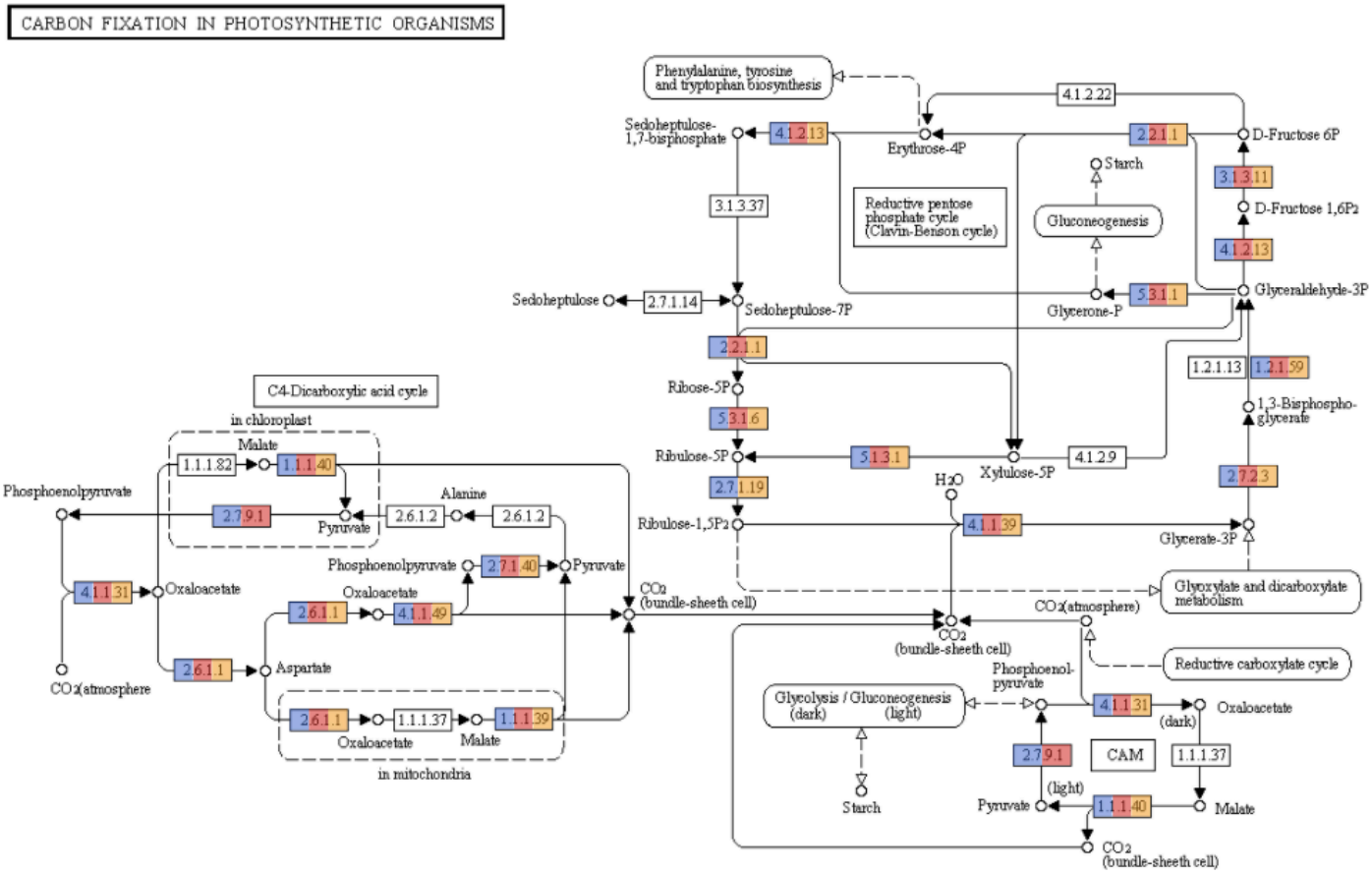

Figure 11: Carbon fixation pathways in detected metagenomes. Colored enzymes indicate the presence of the enzyme, light blue indicates presence in MFF1, red indicates presence in MFF2 and orange indicates presence in MFF3.

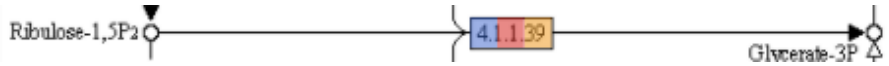

Figure 12: Reaction catalyzed by the enzyme Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO). The enzyme is present in all three metagenomes.

Figure 9 details the putative proteins obtained when aligned with the Kegg Onthology (KO) database. Around 59% of the contigs obtained in all three samples were related to metabolism. This is consistent with other results in

hypersaline environments where a great number of metabolic processes such as primary production and degradation of organic compounds are carried out (Javor, 2012). Primary production is usually carried out at salinities of 25% and above solely by the halophilic green algae *Dunaliella* (Joint et al., 2002; Oren, 2014). However, it has also been found that certain *Cyanobacteria* are also capable of thriving at high salinities. For instance, cyanobacteria phylogenetically close to the genus *Halothece* have been reported in the Atacama Desert in Chile with an average salinity of about 15% NaCl (w/v) (de los Rios et al., 2010). However, it is rare to find *Cyanobacteria* growing at these salinities and our results contrast with Ghai et al. (2011) where Cyanobacteria were absent at salinities of 19% and beyond. Genes related to carbon fixation were encountered in all three metagenomes **(Figure 11).** Particularly, the gene encoding for the Ribulose-1,5-bisphosphate carboxylase/oxygenase enzyme, more commonly known as RuBisCO was present. This enzyme is critical for carbon fixation because it catalyzes the very first step in the Calvin Cycle. Cyanobacteria carry out carbon fixation using RuBisCO enzyme in their carboxysomes; their presence associated with the detection of this enzyme suggest that Cyanobacteria might be contributing to primary production at a salinity of up to 34%, previously unreported in marine solar salterns. Transcriptomic approaches are necessary in order to determine if these cyanobacteria are metabolically active or if the hits returned were only dormant genes.

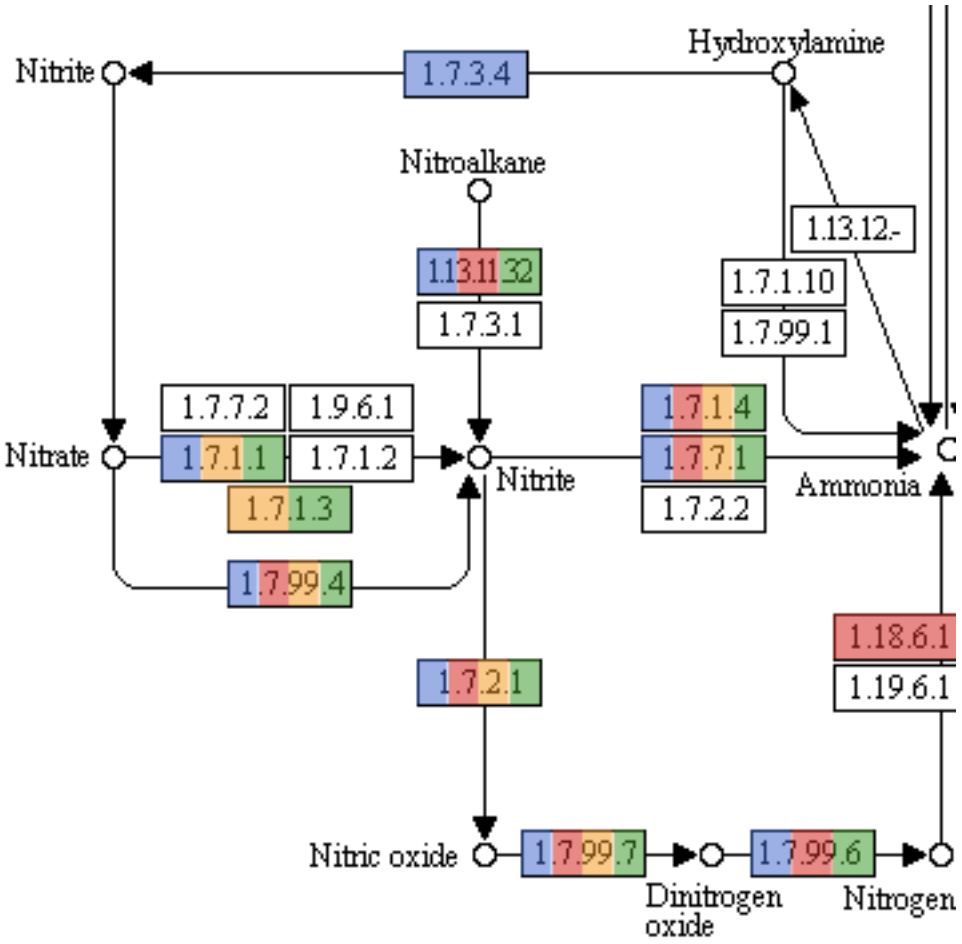

Figure 13: Nitrogen metabolism pathways present in the metagenomes. Colored enzymes (Light blue and green for MFF1, red for MFF2, orange for MFF3) indicate presence of hits related to the enzyme in the metagenome. Enzymes related to reductive pathways in the nitrogen cycle were encountered, including nitrate reductase (Enzyme Nomenclature database numbers (EC) 1.7.1.1, 1.7.1.3, 1.7.99.4), nitrite reductases (EC 1.7.2.1, 1.7.7.1, 1.7.1.4) as well as the nitrogenase needed for nitrogen fixation (EC 1.18.6.1)

Microorganisms also play a pivotal role in the nitrogen cycle. Bacteria in particular are involved in all N-cycle pathways and their nitrogen metabolism has been studied extensively (Cabello et al., 2004). However, nitrogen metabolism in archaea is not well studied and understood. Archaea are known to be participants in all the reductive pathways of the N-cycle. Archaea inhabiting extreme

environments, are considered the principal driving force of the N-cycle in these types of environments (Cabello et al., 2004). **Figure 13** illustrates the pathways concerning the nitrogen cycle and as expected, genes encoding for enzymes related to the reductive pathways of the N-cycle were present. Nitrogen fixation, the process catalyzed by a nitrogenase in which atmospheric nitrogen is converted to ammonia, is performed naturally by both bacteria and archaea. In archaea, nitrogen fixation has been reported in the methanogenic representatives of the phylum *Euryarchaeota* (Offre et al., 2013). The presence of genes associated to members from classes *Methanobacteria*, *Methanococci*, *Methanopyri* and *Methanomicrobia* in the metagenomes suggest that these organisms might be responsible for carrying out nitrogen fixation in the Cabo Rojo Salterns.

Nitrification, the conversion of ammonia to nitrate and subsequently nitrite is another pivotal process in the nitrogen cycle. Until recently, it was believed that this process was only undertaken by bacteria. However, several ammonia oxidizing archaea have been described, all representatives of the phylum *Thaumarchaeota* (Stahl & de la Torre, 2012). Sequences related to *Nitrosopumilus*, an ammonia oxidizing archaeon (Konneke et al., 2005; Walker et al., 2010), were found in the three metagenomes. However, as **Figure 13** shows, there are no sequences matching the ammonia monooxygenase (AMO) enzyme. The absence of AMO despite the presence of sequences related to ammonia oxidizing archaea highlights the limitations of metagenomics in obtaining complete genomic sequences. As sequencing technology improves combined with better bioinformatics tools to analyze the enormous amounts of data, these gaps in information will be shortened. Nevertheless, the presence of sequences from ammonia oxidizing archaea suggests the process might be performed.

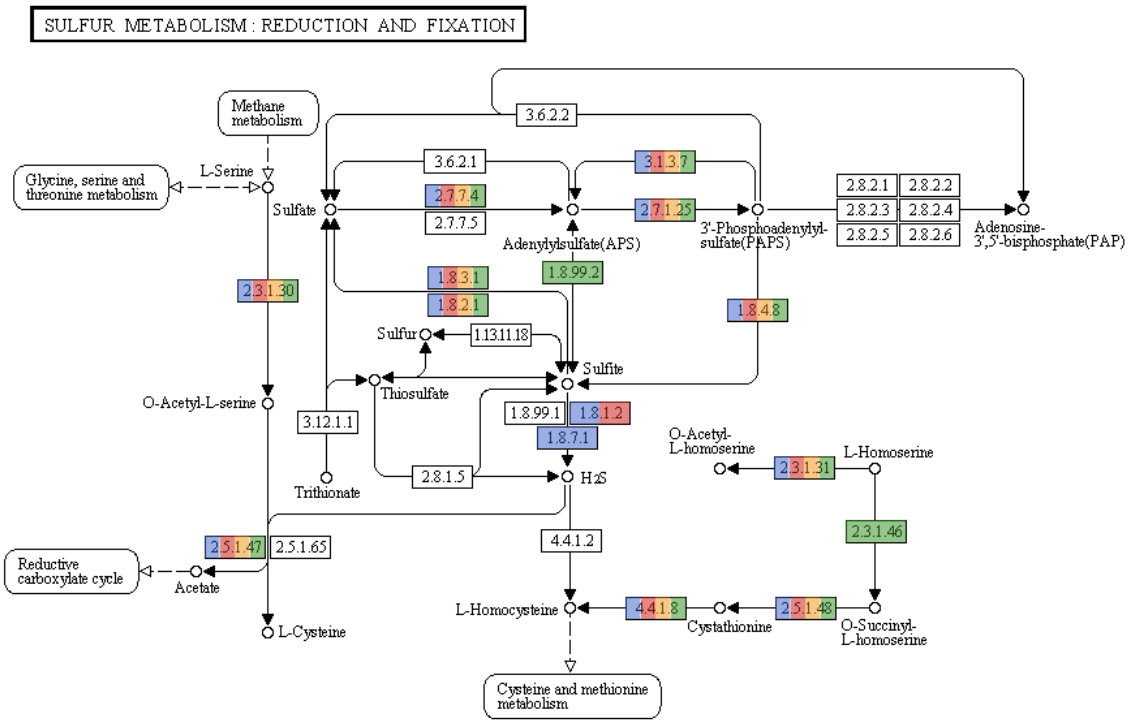

Figure 14: Sulfur metabolism pathways present in the metagenomes. Colored enzymes indicate the presence of hits related to the enzyme in the metagenome. Enzymes related to sulfate reduction (2.7.7.4, 1.8.99.2, 1.8.1.2, 1.8.7.1) were encountered.

As illustrated in **Figure 13,** hits matching sulfur metabolism were also found. The sulfur cycle is another prominent biogeochemical process undertaken in hypersaline environments. Both archaea and bacteria play a pivotal role in the cycling of sulfur. Sulfidogenesis, the production of H_2S from the reduction of elemental Sulfur (S^0), sulfate, thiosulfate or sulfite is a major step in the sulfur cycle. Bacteria possess sulfate reducing representatives in the phyla *Deltaproteobacteria* and *Firmicutes*. Sulfate reducing bacteria in *Deltaproteobacteria* include the orders *Desulfobacteriales, Desulfovibrionales, Syntrophobacteriales* (Muyzer and Stams, 2008), sequences of which were encountered in the three samples. Furthermore, the genus *Desulfotomaculum* from the *Firmicutes* phylum is also present in our study. Of the archaeal genera

known to reduce sulfate are *Archaeglobus, Thermocladium* and *Caldivirga.* However, hits related to these genera were not encountered. This result is not surprising because these organisms are not found at high salinity environments (Barton and Tomei, 1995).

The oxidation of H_2S is also another important pathway in the sulfur cycle since hydrogen sulfide is toxic to plant and animal tissue. In hypersaline environments, representatives from the *Gammaproteobacteria* such as *Halothiobacillus, Thiomicrospira* among others are classified as sulfur oxidizing bacteria (SOB) (Tourova et al., 2010). All three samples contained representatives from *Gammaproteobacteria* including protein sequences matching those of *Halothiobacillus* and *Thiomicrospira.* Sulfur oxidizing archaea (SOA) have been poorly characterized and only two genera, *Acidianus* and *Ferroglobus* are known to carry out sulfur oxidation (Offre, Spang & Schleper, 2013). Neither genus was encountered in our samples nor were they expected due to both being hyperthermophiles growing optimally at temperatures above 60°C (Segerer et al., 1986; Hafenbradl et al., 1996).

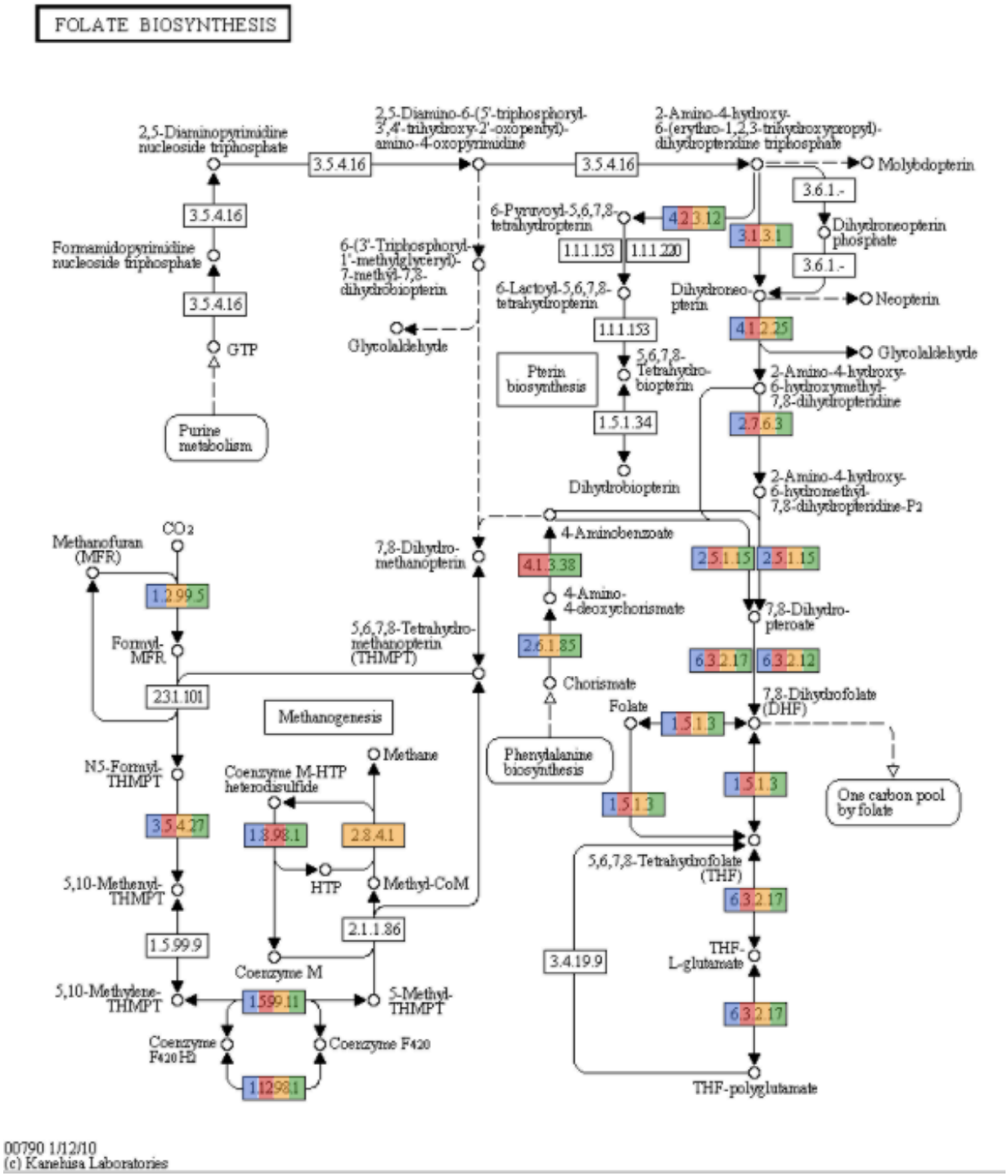

Figure 15: Folate biosynthesis pathway. Presence of alkaline phosphatase 3.1.3.1 as a catalyst in folate biosynthesis suggests the presence of phosphate solubilizing organisms.

Few studies have been undertaken regarding the presence of phosphate solubilization pathways in archaea. This process consists of utilization of inorganic phosphate and its conversion into organic compounds (Khan et al., 2014). Recent studies have encountered halophilic archaea responsible for phosphate solubilization including *Haloarcula, Halococcus, Halobacterium, Haloterrigena,* among others (Yadav et al., 2014). The presence of the alkaline phosphatase, one of the key enzymes in phosphate solubilizing pathways, along with the presence of these halophilic phosphate-solubilizing archaea, suggest that there are organisms in our data that play a key role in these processes.

This data shows that microorganisms in hypersaline environments play an important role in the biogeochemical cycles with most of the relevant pathways present in the metagenome. Furthermore, representatives known to perform processes in each of these cycles have been encountered. The absence of genes such as ammonia monooxygenase is an issue to address and highlights the limitations of metagenomics in obtaining partial sequences. With further sampling as well as the evolution of sequencing technologies a more complete assessment can be carried out as well as novel new pathways can be encountered that have not been described for the process at hypersaline environments.

4.4 Conclusion

In this study, we have collected a considerable amount of sequence data (64 GB) using a culture independent approach providing a more comprehensive perspective of microbial community structure and functional gene composition. By using a PCR unbiased culture-independent approach a large microbial diversity has been uncovered in Cabo Rojo salterns. A diverse representation of prokaryotic phyla at a salinity of 34% is previously unreported at this salinity. The microbial community structure could be influenced by weather fluctuations that contribute to changes in the ionic composition of the crystallizer ponds. Further sampling may reveal more changes to this microbial community and can possibly unveil novel microbial groups. The importance of the organisms present in the

Fraternidad crystallizer ponds is highlighted by the presence of essential genes related to the carbon, nitrogen and sulfur cycles ,with representatives from the diverse phyla encountered contributing to these cycles.

4.5 Recommendations

- Sampling in the Fraternidad lagoon and Candelaria salterns systems can give us a more complete perception of the microbial diversity in the Cabo Rojo salterns and may also provide us information on how community composition and microbial processes change across the salinity gradient.
- Analysis of ionic composition of crystallizer ponds following precipitation events.
- Conduct metatranscriptomic analyses in order to observe the active microbial community and their respective processes. Metagenomic analysis although providing a wealth of information, does not distinguish active communities from these dormant communities.

5. Putative Novel Uncultured Species

5.1 Summary

Binning of the three metagenomes was performed following the mapping of the original reads back to the assembly. Fifty five bins were obtained in MFF1, sixty four from sample MFF2, and fifty seven from sample MFF3. Quality check using CheckM provided statistics for completeness as well as contamination in the metagenomic bins obtained. Seven bins were selected based on their degree of completeness and low degree of contamination. Taxonomy was assigned using microbial genome atlas (MiGA) on the basis of amino acid identity (AAI) which establishes that organisms grouped in the same species exhibit an AAI of > 85%. BIN33 was encountered to be related to *Natronomonas moolapensis* with 59.74% of AAI. This result as well as phylogeny obtained, suggest BIN33 to be a novel genus in the family *Halobacteriaceae.* BIN36 matched with 61.90% AAI to *Natronomonas pharaonis.* The presence of ammonium uptake genes as well as RuBisCO have been found in other *Natronomonas* genomes. This organism most likely belongs to the genus *Natronomonas* and represents a novel species. BIN32 returned 60.76% AAI to *Haloferax volcanii.* The strain is proposed to be a new species within the genus *Haloferax.*

BIN24 represented the only bacterial genome obtained in our studies. Genome analyses revealed the bacterium to be halotolerant, gram negative and a new genus in the phylum *Bacteroidetes* with an AAI of 43.31%. BIN39 returned the closest species to be *Halomicrobium mukohataei.* This organism possesses necessary enzymes for anaerobic respiration using nitrate as an electron acceptor as has also been found in *H. mukohataei.* This organism is believed to be a new species in the genus *Halomicrobium* with an AAI of 62.81%.

BIN20 exhibited an AAI of 65.83% with *Haloquadratum walsbyi.* The detection of two types of bacteriorhodopsins, one halorhodopsin, gas vesicle proteins and more importantly the protein halomucin combined with

phylogenetic analyses indicate that BIN20 represents a possible novel species of *Haloquadratum.*

Finally, BIN46 returned 60.51% to *Natronomonas moolapensis.* Similar properties to BIN36 were encountered and ANI was performed in order to determine if both bins belonged to the same species. Results revealed the organisms to be a different species from the genus *Natronomonas*. Using binning methods, we have uncovered novel organisms that can be proposed as candidate species names until phenotypic characterization is developed. Genome analyses could possibly give us information for designing isolation strategies for these organisms. With further sampling it is expected to encounter even more novel organisms through binning methods.

5.2 Materials and Methods

Following assembly of the metagenomes, the original reads were mapped back to the assembly in order to obtain a coverage using the Burrows-Wheeler Aligner (Li and Durbin, 2009). Subsequently, the coverage files along with the final assembly were binned for putative genomes. Quality of genomes including completeness and contamination was assessed using CheckM tool (Parks et al., 2015). Taxonomy of quality bins obtained was assessed by means of Amino Acid Identity (AAI) using the Microbial Genome Atlas (MiGA) web interphase (Rodríguez-R and Konstantinidis, 2016). Phylogeny was determined aligning concatenated essential genes against related sequences using MEGA Software (Kumar et al., 2015). Annotation of all genomes was carried out using the Rapid Annotation Subsystems Technology (RAST) pipeline (Aziz et al., 2008).

5.3 Results

5.3.1 Binning results

Binning of all three metagenomes yielded over 50 putative genomes from all three assemblies. Assessment of quality performed utilizing CheckM indicated 7 genomes with high degree of completeness and low degree of

contamination detected among the three metagenomes. The statistics obtained from the genomes are described in **Table 5**.

Table 5: Statistics for genomic bins.

Bin Name	Completeness	Contamination	GC Content	# Contigs	N50 (bp)	Genome Size	Predicted Proteins	Source
BIN33	94.40	2.40	66.22%	11	392,903	1.5 Mb	1,576	MFF1
BIN36	92.80	5.60	65.62%	96	70,511	2.9 Mb	3,296	MFF1
BIN32	89.00	5.60	65.85%	80	39,718	1.9 Mb	2,110	MFF1
BIN24	80.20	1.80	52.79%	28	125,394	1.8 Mb	1,533	MFF1
BIN39	96.00	3.20	68.18%	75	54,228	2.4 Mb	2,502	MFF2
BIN20	88.66	3.20	50.25%	273	24,930	4.3 Mb	4,449	MFF2
BIN46	87.20	5.20	65.04%	232	25,159	2.9 Mb	3,373	MFF3

Upon assignment of taxonomic bins, it is important to avoid chimeric bins that might lead to erroneous conclusions (Thomas et al., 2012). Caution should be taken before validating genomic bins due to contaminating fragments. Of the software available, CheckM provides an accurate estimate of genome completeness and contamination (Parks et al., 2015). A high number of taxonomic bins were obtained using binning methods, however most of these exhibited either a low degree of completeness or a high degree of completeness but also a high contamination. Nevertheless, we were able to obtain 7 genomic bins of high quality from the three metagenomic libraries. All bins presented high amount of completeness and a low degree of contamination (**Table 5**).

5.3.2 Taxonomy of genomic bins

Taxonomy assigned using MiGA produced the most comprehensive analysis for the genomes.

Table 6: Amino Acid Identity (AAI) comparison for all the genomic bins obtained.

Bin	Closest Relative	AAI	Fraction of Proteins Shared
BIN33	*Natronomonas moolapensis*	59.74%	81.85%
BIN36	*Natronomonas pharaonis*	61.90%	64.33%
BIN32	*Haloferax volcanii*	60.76%	68.04%
BIN24	*Pontibacter korlensis*	43.31%	66.02%
BIN39	*Halomicrobium mukohataei*	62.81%	68.27%
BIN20	*Haloquadratum walsbyi*	65.83%	73.46%
BIN46	*Natronomonas moolapensis*	60.51%	55.45%

Assigning taxonomy to uncultured organisms poses more of a challenge due to the lack of phenotypic characterization. Therefore, the information available is based only on sequence data. The *Candidatus* status bypasses this limitation by assigning candidate names until phenotypic characters are appropriately characterized (Rodríguez-R and Konstantinidis, 2014). Several methods have been proposed in order to identify microbial species at the genome level. For cultured species, the DNA-DNA hybridization (DDH) has been the most traditional approach to differentiate closely related species with the 70% identity cutoff (Arahal, 2014). However, the Average Nucleotide Identity (ANI) has been proposed as an alternate way of distinguishing bacterial and archaeal species. The cutoff for the ANI analyses is 95% and has been employed successfully for the characterization of new microbial species (Goris et al., 2007; Konstantinidis and Tiedje, 2005; Richter and Rosselló-Mora, 2009). Special caution should be taken, however, when describing species within a population due to the members of a microbial population exhibiting gene differences of less than 5% of their total genes. Furthermore, ANI offers more robust resolution between genomes that share 80-100% ANI; organisms that show less than 80% ANI are too divergent to

be compared based on ANI measurement (Rodríguez-R and Konstantinidis, 2014). Due to this, it is recommended to use amino acid identity (AAI) to distinguish between more divergent organisms (Konstantinidis and Tiedje, 2005). Organisms exhibiting an AAI of > 85% are typically grouped within the same species whereas those grouped in the same genus exhibit an AAI from 60%-80% (Luo et al., 2014). Due to this, we used AAI to determine taxonomy for our 7 metagenomic bins **(Table 6).**

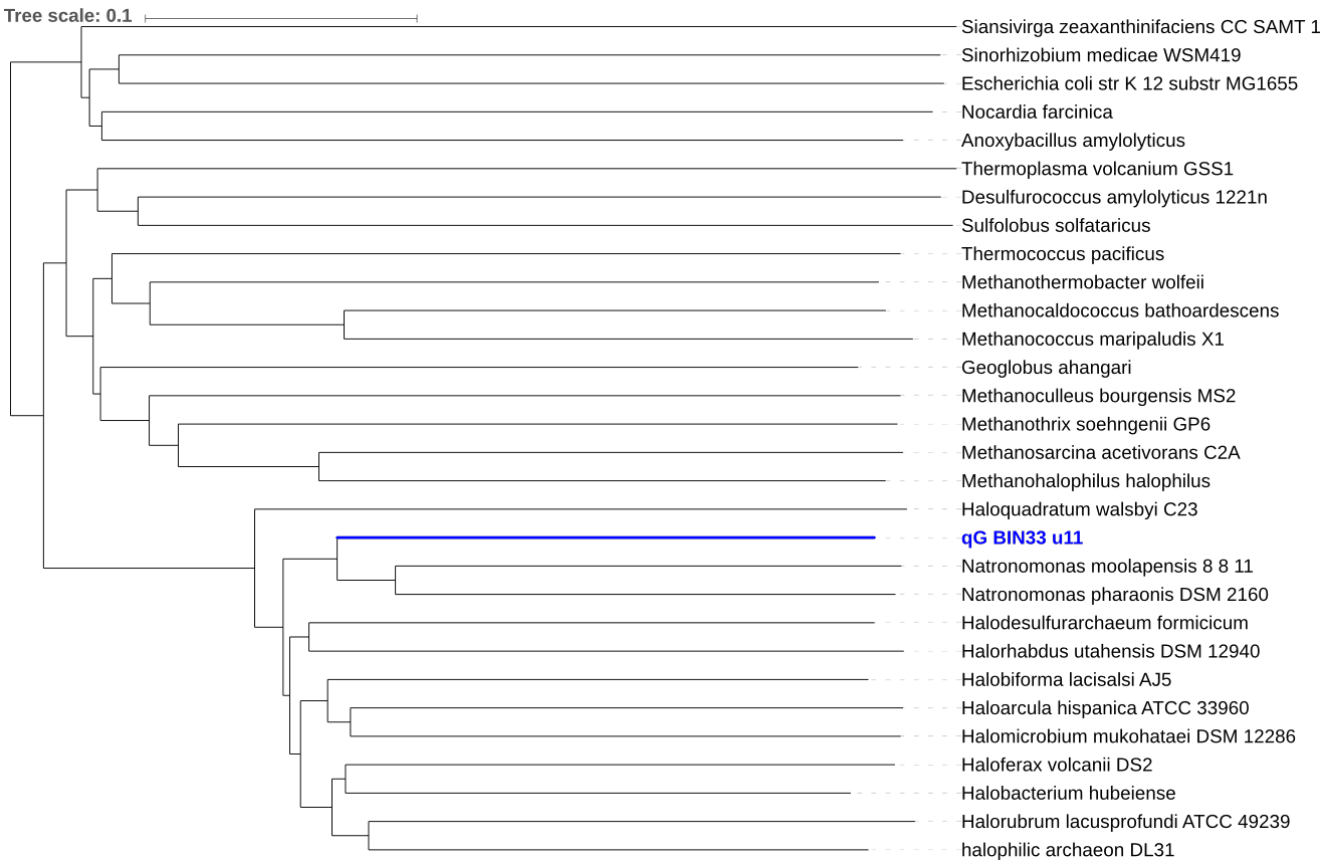

Figure 16: Phylogeny of BIN33 using Amino Acid Identity. Scale represents change of aminoacid substitution over time.

Statistical analysis using MiGA revealed that the dataset most likely belonged to a species not represented in the database with a p-value of 0.0038.

AAI results for BIN33 returned its closest relative to be *Natronomonas moolapensis* with an identity of 59.74%. Phylogeny using AAI (**Figure 15)** suggests that this genome is a member of the family *Halobacteriaceae* and could represent a new genus within the family. The high GC content as well as the presence of proteins associated with hyperosmotic stress indicate this organism is halophilic as is characteristic of organisms thriving in hypersaline environments. The strain is non-motile due to the absence of motility genes, gas vesicle cluster and chemotaxis genes. Furthermore, several enzymes from the glycerol utilization cluster included glycerol kinase and glycerol-3-phosphate dehydrogenase. The presence of these enzymes suggests that this strain could possibly grow on media containing glycerol.

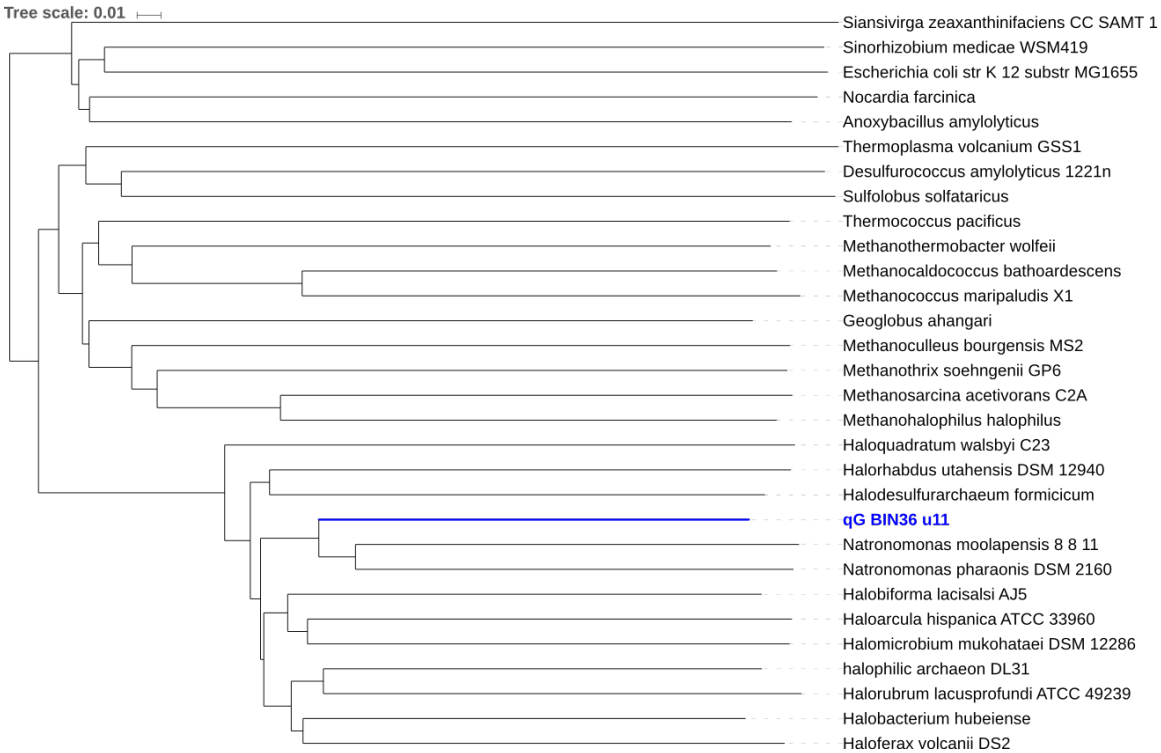

Figure 17: Phylogeny of BIN36 using Amino Acid Identity. Scale represents the amount of aminoacid subsitutions over time.

Statistical analysis returned a p-value for 0.0046 for a species not represented in the databases.

Similarly, BIN36 returned its closest relative to be *Natronomonas pharaonis* with an identity of 61.90%. The genus *Natronomonas* has two described species and was first isolated in alkaline pH (Kamekura et al., 1997). Nevertheless, *Natronomonas moolapensis* was isolated in neutral pH and grows optimally at a pH of 7 to 7.5 (Burns et al., 2010). The presence of flagellar proteins, signal transduction enzymes (CheA and CheY) suggest BIN36 to be motile. For the uptake of ammonium, BIN36 possesses several transporters for ammonium (*amtB*), nitrate/nitrite (*narK*) and urea (ABC transporter urt). Also, glutamine synthase (*glnA*) and glutamate synthase (*gltB*) genes were encountered. Therefore, this strain seems to assimilate ammonium and able to convert it to glutamate. Furthermore, it is suggested that the electron donor for the reductive conversions is ferredoxin due to the presence of ferredoxin-nitrite reductase genes. Similar nitrogen metabolism was described for the genome sequence of *Natronomonas pharaonis (*Falb et al., 2005). However, the *Natronomonas pharaonis* genome (Falb et al., up. cit.) exhibited the presence of urea conversion enzymes which were absent in this genome.

Other notable genes encountered in the genome was the enzyme RuBisCO, which in haloarchaea, occurs only in *Natronomonas pharaonis* (Falb et al., 2008; Konstantinidis et al., 2006). Sato et al. (2007), proposed the archaeal RuBisCO to be involved in AMP recycling pathway related to purine and pentose metabolism which in turn could be related to a cyclic CO_2 fixation pathway. Enzymes related to glycerol metabolism were also encountered including the glycerol kinase and glycerol-3-phosphate dehydrogenase. Due to similarities with both the genomes in the genus *Natronomonas* (Dyall-Smith et al., 2013; Falb et al., 2005), we can conclude that the organism most likely represents a new species of *Natronomonas* that can grow on neutral pH similarly to *Natronomonas moolapensis.*

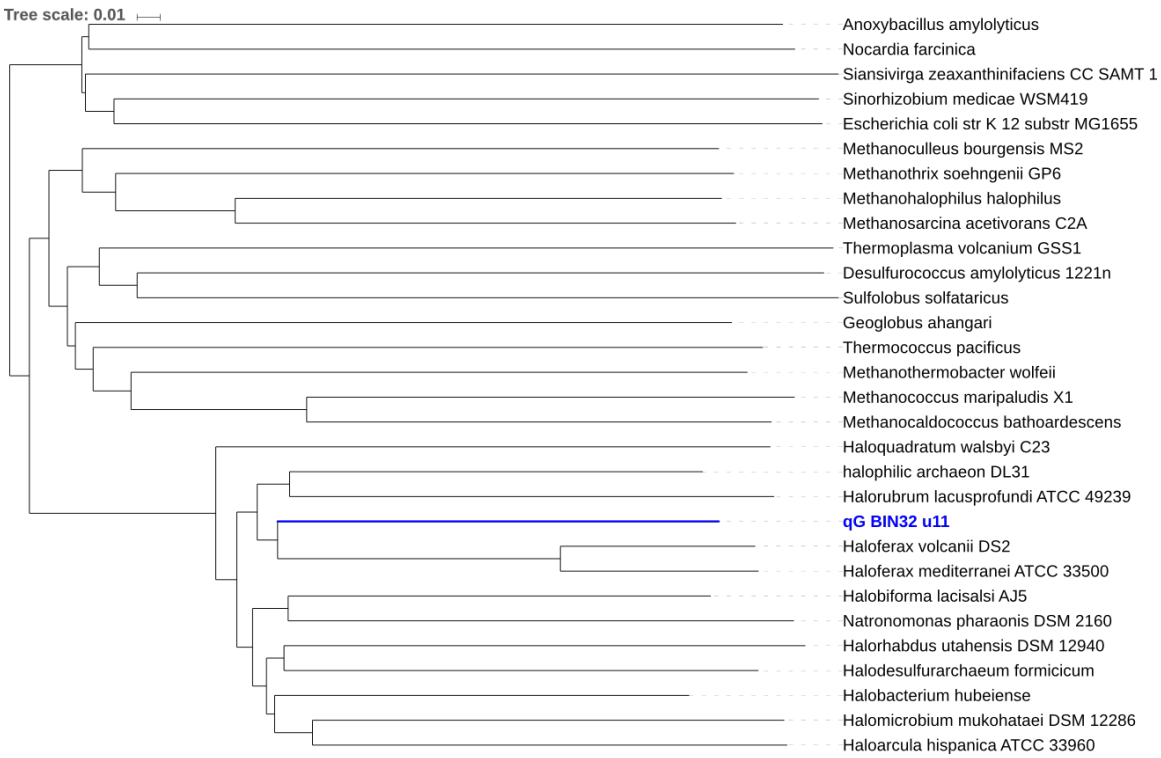

Figure 18: Phylogeny of BIN32 using Aminoacid Identity. Scale represents the amount of aminoacid substitutions over time.

Taxonomic novelty analyses returned a p-value of 0.0038 for a species that is not represented in the database.

AAI for BIN32 returned its closest relative to be *Haloferax volcanii* with an identity of 60.34%. The lack of motility genes, gas vesicles and signal transduction pathways suggest the organism is non-motile. The genome possesses all the enzymes for the Entner-Doudoroff pathway as well glycerol kinase and glycerol-3-phosphate dehydrogenase. Nitrogen metabolism genes for the reductive pathways of the nitrogen cycle were encountered including nitrite transporter and nitrite reductase. Due to the relatedness of BIN32 to *Haloferax volcanii,* ANI was conducted to compare both genomes. ANI returned an 89.94% identity. Based on phylogeny and ANI results, BIN32 is suggested to be a new species in the family *Haloferacaceae.*

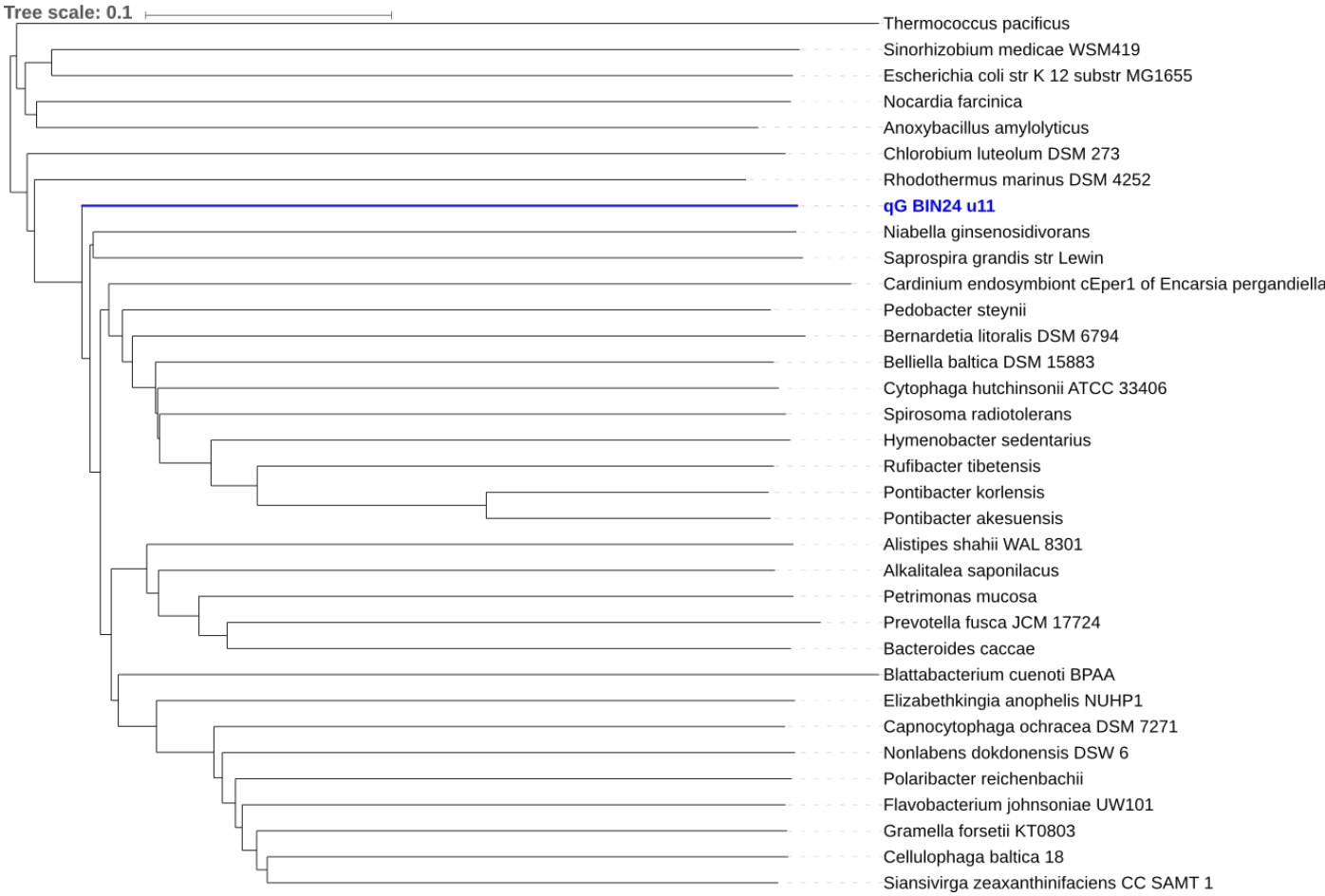

Figure 19: Phylogeny of BIN24 using Amino Acid Identity. Scale represents the amount of aminoacid substitutions over time.

Statistical analyses for BIN24 reveal p-values of 0.0026 for a genus not described in the database. P value of 0.055 could suggest that this organism could possibly belong to a family not described in the RefSeq database.

Finally, sample MFF1 also provided a bacterial bin in BIN24 which showed close relatedness to *Pontibacter korlensis.* Proteins encoding gram-negative cell wall were matched in the genome therefore classifying this organism as gram-negative. No chemotaxis proteins nor flagellar proteins were identified,

suggesting this bacterium to be non-motile. Ammonia assimilation genes, including ammonium transporter as well as nitrite reductase genes. Due to its presence in high salinity as well as its relatedness to *Pontibacter*, it is suggested to be halotolerant. Nevertheless, AAI as well as statistical analyses performed in MiGA suggest this organism to be a new genus in the phylum *Bacteroidetes*. It's low GC content is unusual compared to other organisms in hypersaline environments; however, Ghai et al. (2012) encountered a new genus of low GC Actinobacteria in the Santa Pola salterns, Spain.

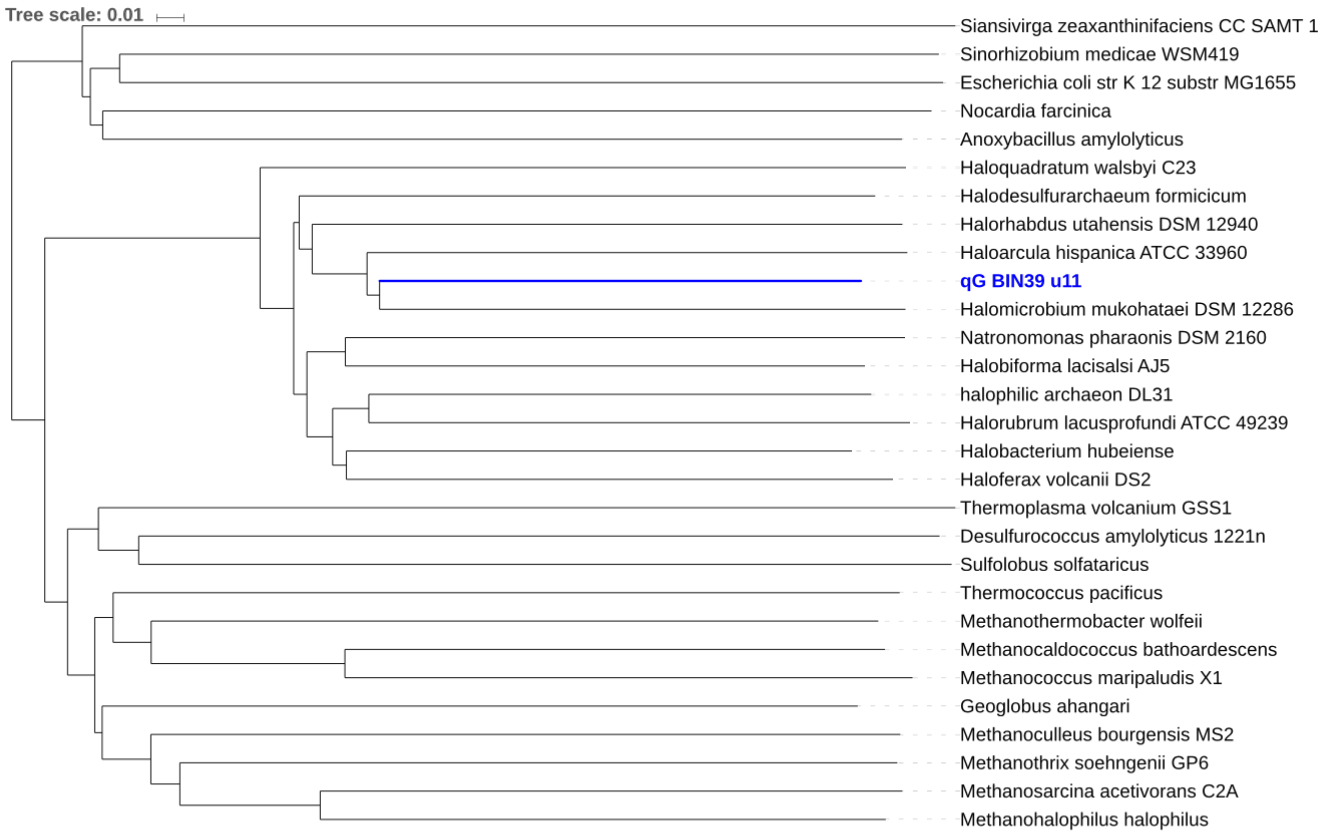

Figure 20: Phylogeny of BIN39 using Amino Acid Identity. Scale represents the amount of aminoacid substitutions over time.

Statistical analysis reveal that the dataset possibly belongs to a new species not represented in a database with a p-value of 0.0046 for species and a p-value of 0.322 for genus.

BIN39 returned similarity to *Halomicrobium mukohataei* with 62.81%. The organism is motile with genes encoding for archaeal flagellar proteins. Furthermore, this organism possesses genes necessary for ammonia assimilation as well as nitrate and nitrite reductases. *Halomicrobium mukohataei* has been described to be able to grow anaerobically under the presence of nitrate as a terminal electron acceptor and forming nitrite as an end product in anaerobic respiration (Oren et al., 2002). Similar growth has also been observed in other organisms such as *Corynebacterium glutamicum* where nitrate was used as an electron acceptor producing nitrite as an end product (Nishimura et al., 2007). Once again, the presence of glycerol kinases and glycerol-3-phosphate dihydrogenase suggests that this organism can grow on media containing glycerol. Our results suggest this organism to be a novel species of the genus *Halomicrobium.*

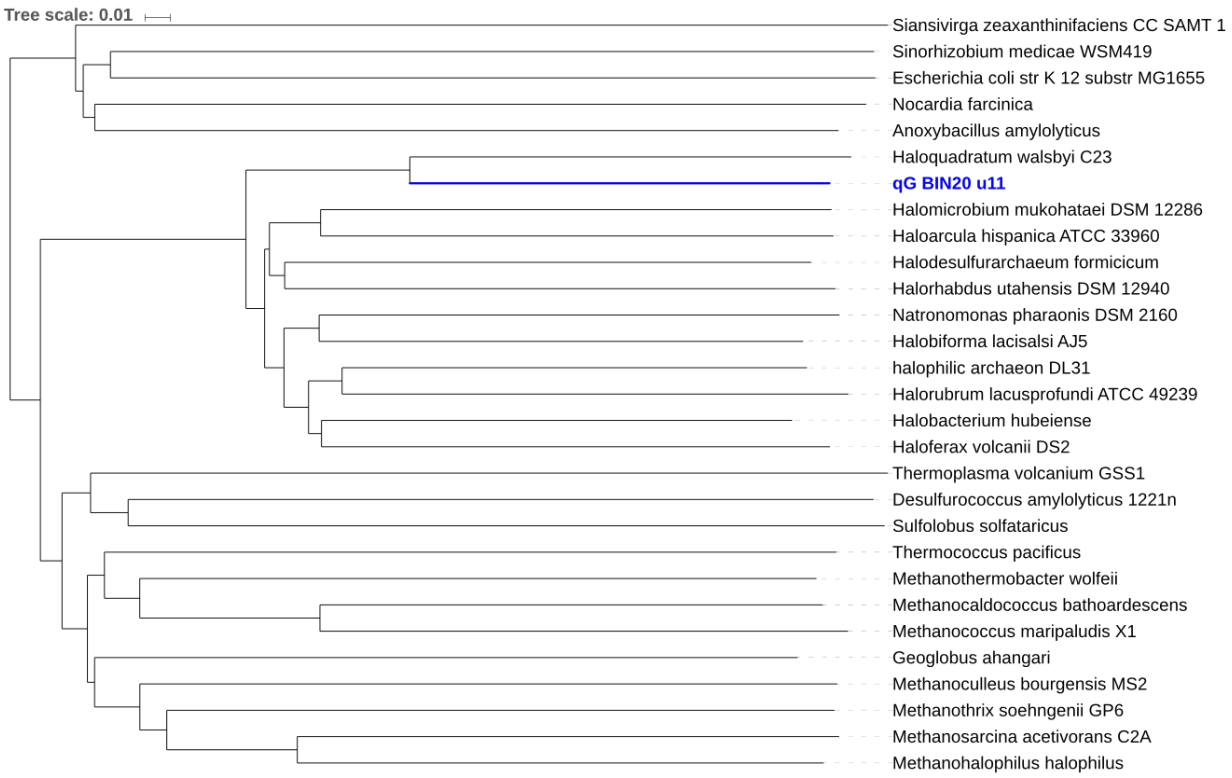

Figure 21: Phylogeny of BIN20 using Amino Acid Identity. Scale represents the amount of aminoacid identity change over time.

Statistical analysis returned p-value of 0.0054 for a species not represented in the RefSeq database and p-value of 0.437 for genus not represented in the database.

BIN20 returned an AAI of 65.83% with *Haloquadratum walsbyi.* The sequence of halomucin was blasted against the genome and was encountered. Halomucin, known as the largest archaeal protein, with 9,159 aminoacids, was described for the first time in *Haloquadratum walsbyi* (Bolhuis et al., 2006). Halomucin provides desiccation protection in saline environments to *Haloquadratum* and is probably the secret of success for this organism in these environments. The presence of the gas vesicle cluster also

coincides with the genome data of *Haloquadratum.* Bolhuis et al. (2006) also described the presence of two bacteriorhodopsins and one halorhodopsin in the *Haloquadratum* genus which were also encountered here and are the reason they are able to grow phototrophically. This genome also encodes the presence of TRAP-type C4-dicarboxylate transport system, 2 different ABC-type sulfonate transport systems and a phosphonate transport system which are only described in *Haloquadratum walsbyi.* The low GC content in this genome of 50.25% is also comparable to that of *Haloquadratum walsbyi* with 47.9%. This low GC content is uncharacteristic of halophilic archaea due to their exposure to solar radiation. Due to the close relatedness of the genome with the *Haloquadratum walsbyi* genome, ANI was conducted in order to determine further resolution. ANI returned an identity of 89.94%, indicating that BIN20 is possibly a novel species of *Haloquadratum*.

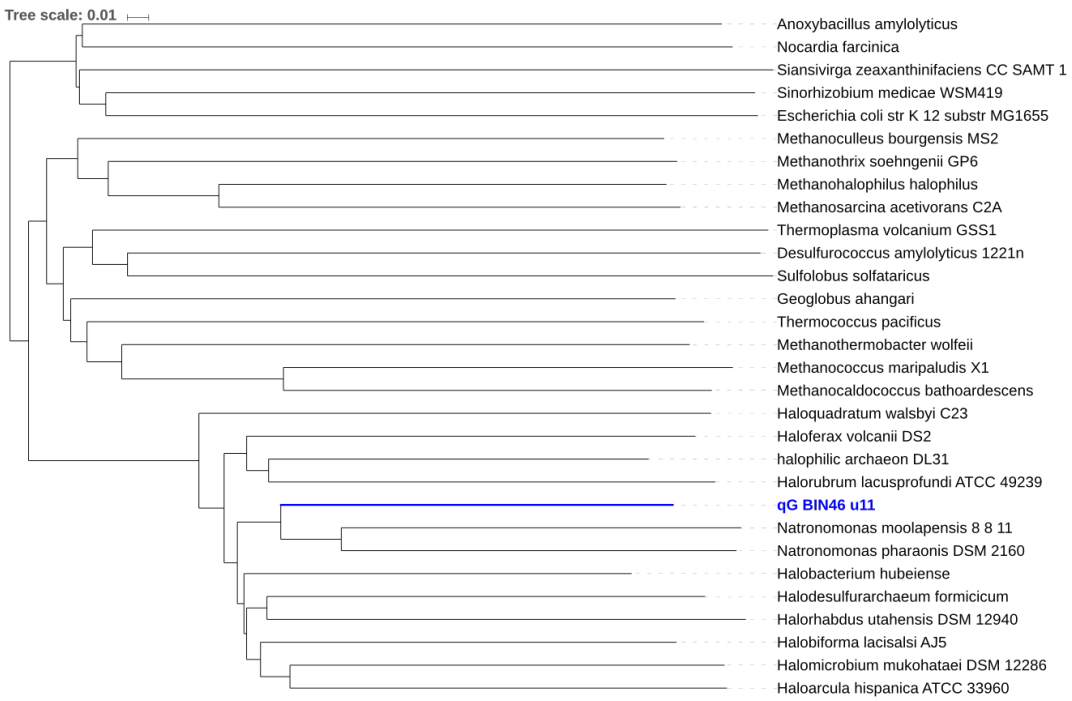

Figure 22: Phylogeny of BIN46 using Amino Acid Identity. Scale represents the amount of aminoacid identity change over time.

Statistical analysis returned p-value 0.0038 for a new species not represented in the RefSeq database and p-value 0.286 for a new genus not represented in the database.

Finally, MFF3 yielded one bin (BIN46) which returned 60.51% AAI to *Natronomonas moolapensis.* Similar to BIN46 the presence of flagellar proteins, and signal transduction enzymes (CheA and CheY) suggest BIN46 to be motile. For the uptake of ammonium, BIN36 possesses several transporters for ammonium (*amtB*) and nitrate/nitrite reductases (*narK*). Also, glutamine synthase (*glnA*) and glutamate synthase (*gltB*) genes were encountered. Therefore, this strain seems to assimilate ammonium and convert it to

glutamate. Furthermore, it is suggested that the electron donor for the reductive conversions is ferredoxin due to the presence of ferredoxin-nitrite reductase genes. As in BIN36, genes encoding for RuBisCO were encountered suggesting this organism also plays a role in the cyclic CO_2 fixation pathway described above. ANI was conducted between BIN36 and BIN46 to verify if they were identical species due to the similarity in the genomes. ANI returned a 99.9% identity suggesting these two organisms are the same species.

5.3.3 Annotation of genomes

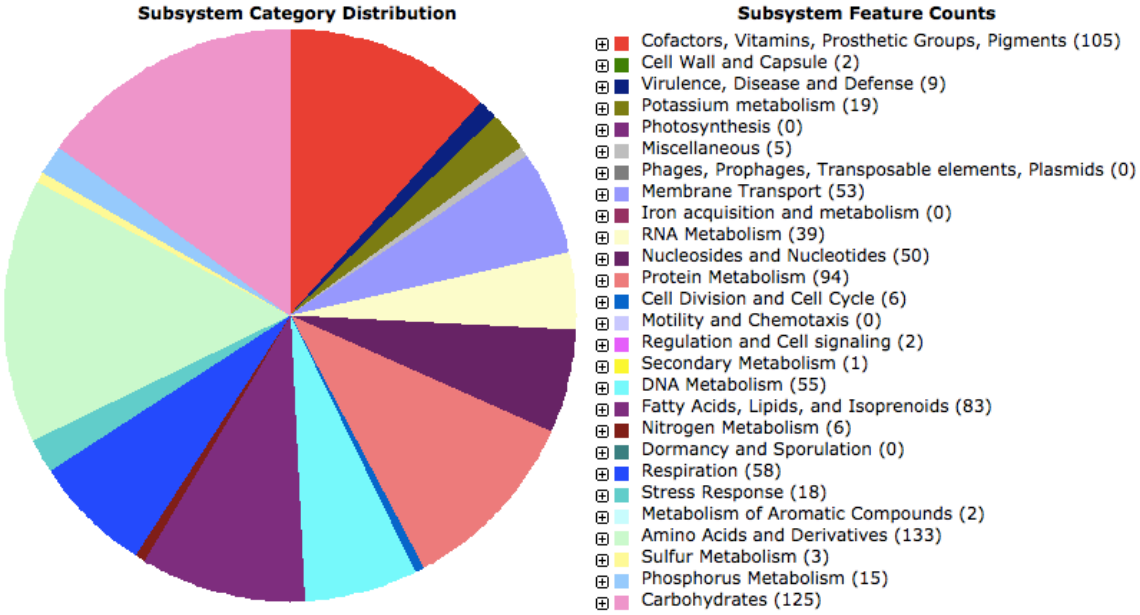

Figure 23: Subsystem category distribution of BIN33. The graph represents the number of proteins that were grouped into subsystems with each section representing a subsystem with the number of proteins associated with that subsystem (shown in parenthesis). As many as 883 from a total 1570 coding sequences were identified to fit into subsystems. This chart was generated using Rapid Annotation System Technology (RAST).

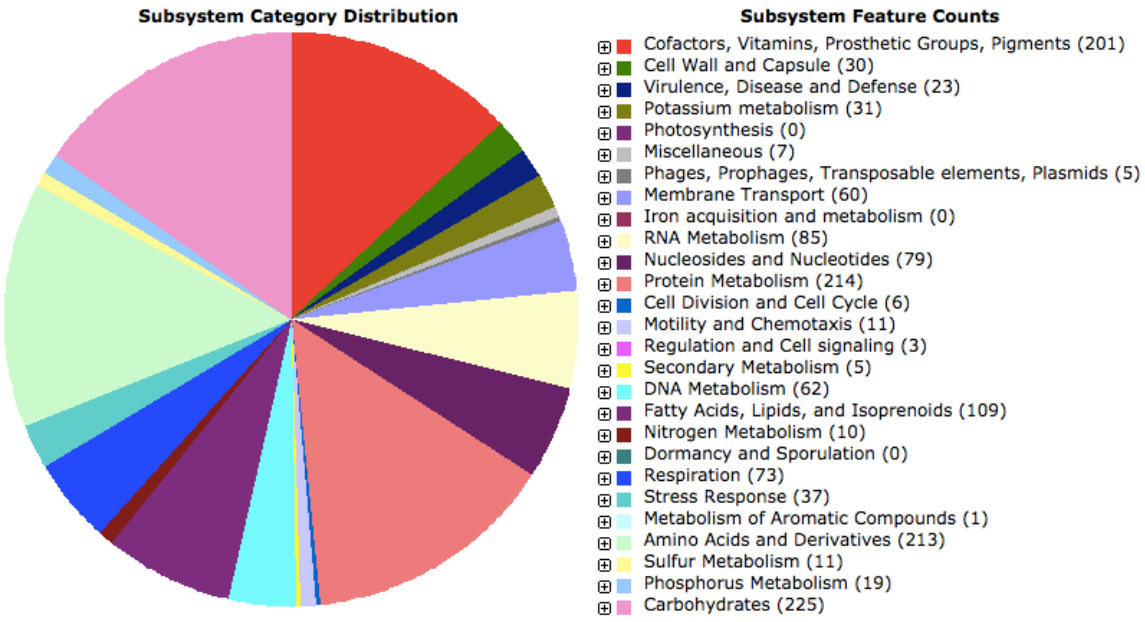

Figure 24: Subsystem category distribution of BIN36. The graph represents the number of proteins that were grouped into subsystems with each section representing a subsystem with the number of proteins associated with that subsystem (shown in parentheses). As many as 1520 of 3233 coding sequences were identified to fit into subsystems. This chart was generated using Rapid Annotation System Technology (RAST).

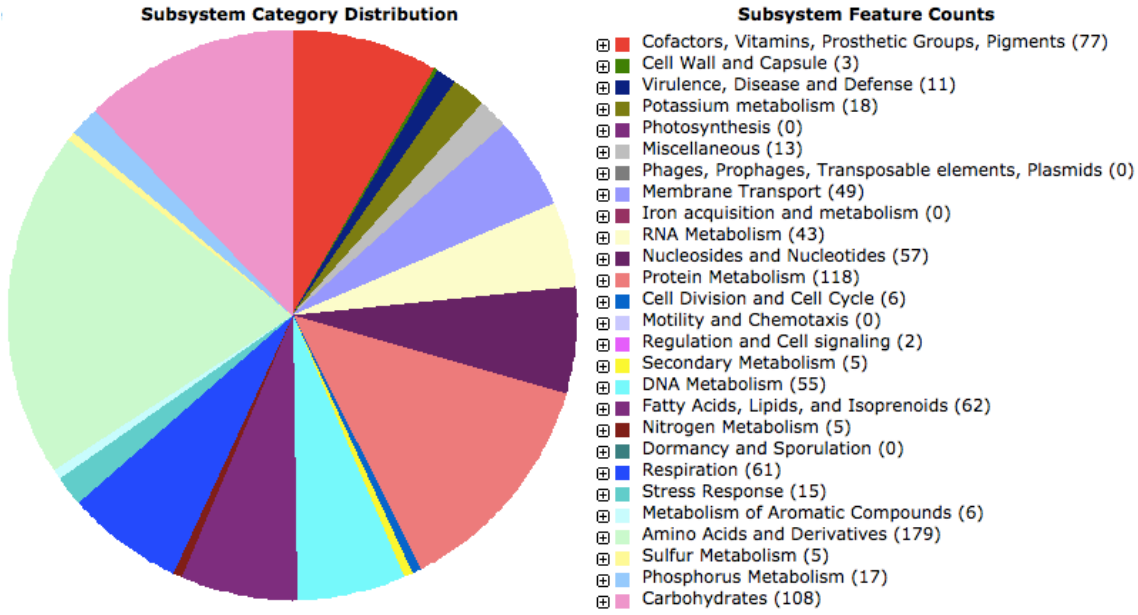

Figure 25: Subsystem category distribution of BIN32. The graph represents the number of proteins that were grouped into subsystems with each section representing a subsystem with the number of proteins associated with that subsystem (shown in parentheses). As many as 915 of 2092 coding sequences were identified to fit into subsystems. This chart was generated using Rapid Annotation System Technology (RAST).

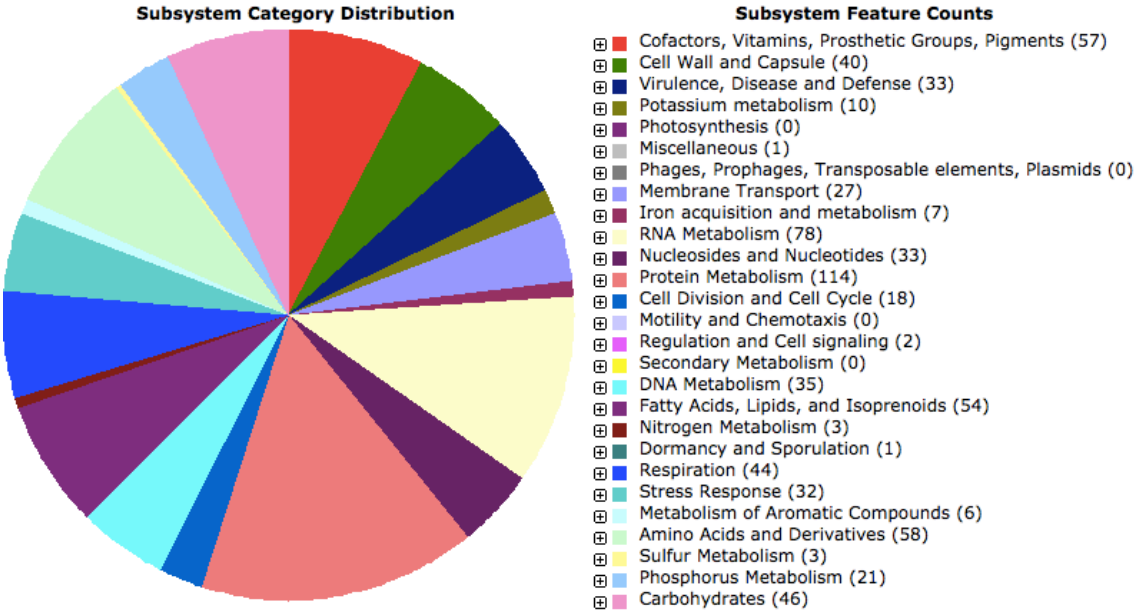

Figure 26: Subsystem category distribution of BIN24. The graph represents the number of proteins that were grouped into subsystems with each section representing a subsystem with the number of proteins associated to with that subsystem (shown in parentheses). As many as 723 of 1561 coding sequences were identified to fit into subsystems. This chart was generated using Rapid Annotation System Technology (RAST).

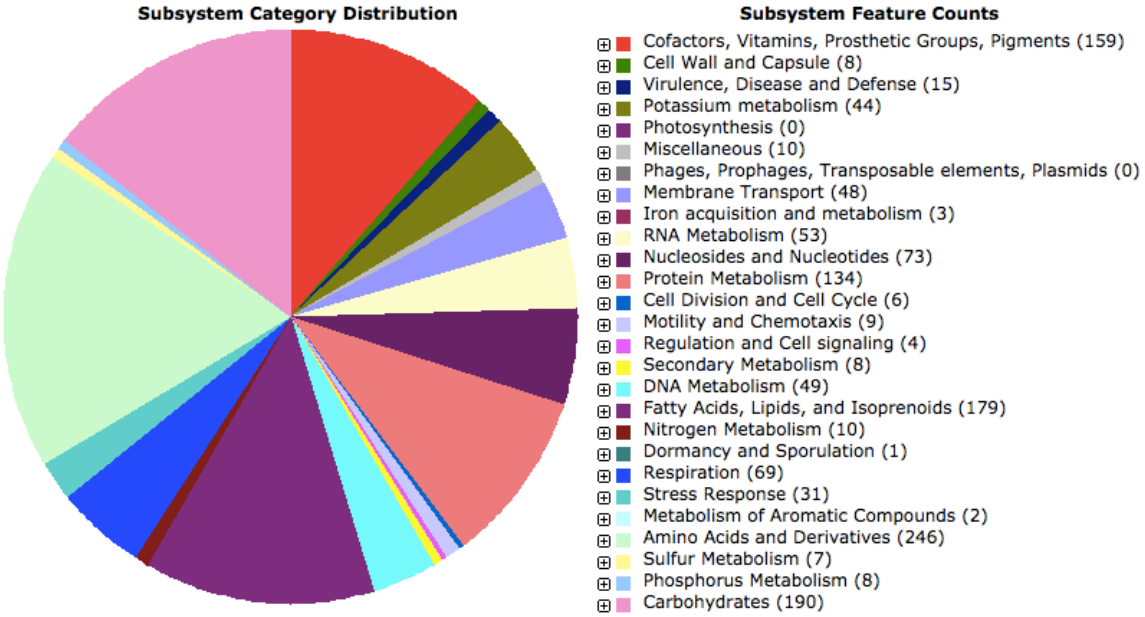

Figure 27: Subsystem category distribution of BIN39. The graph represents the number of proteins that were grouped into subsystems with each section representing a subsystem with the number of proteins associated with that subsystem (shown in parentheses). As many as 1366 of 2439 coding sequences were identified to fit into subsystems. This chart was generated using Rapid Annotation System Technology (RAST).

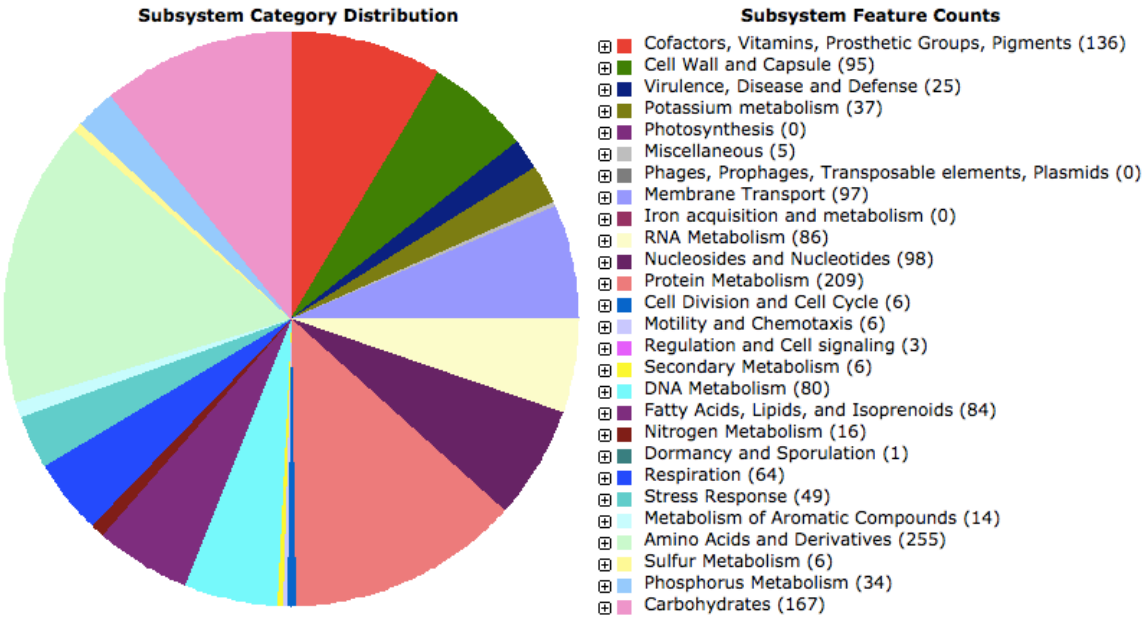

Figure 28: Subsystem category distribution of BIN20. The graph represents the number of proteins that were grouped into subsystems with each section representing a subsystem with the number of proteins associated with that subsystem (shown in parentheses). As many as 1579 of 4980 coding sequences were identified to fit into subsystems. This chart was generated using Rapid Annotation System Technology (RAST).

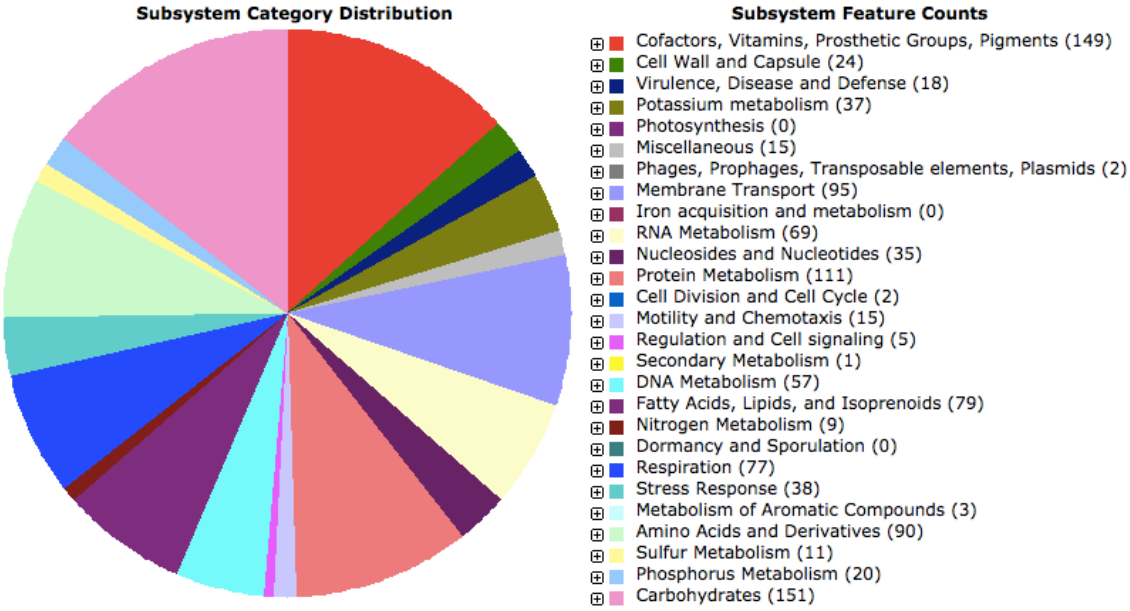

Figure 29: Subsystem category distribution of BIN46. The graph represents the number of proteins that were grouped into subsystems with each section representing a subsystem with the number of proteins associated with that subsystem (shown in parentheses). As many as 1113 of 3204 coding sequences were identified to fit into subsystems. This chart was generated using Rapid Annotation System Technology (RAST).

Binning methods have uncovered previously undescribed microbes in literature. As previously mentioned, Narasingarao et al. (2012) recovered the recently proposed *Nanohaloarchaea* through binning methods. Furthermore, Ghai et al. (2012) uncovered a novel group of low GC Actinobacteria as well as a novel lineage of *Proteobacteria* using metagenomic binning. Finally, through these methods, *Nanoarchaeum equitans* was also described. Therefore, binning methods have been proven to be reliable in describing novel species in these environments and may prove useful in providing a more non-biased assessment of the unculturable diversity in environmental samples.

5.4 Conclusions

Using binning methods, 6 novel organisms undetected with traditional culture-dependent surveys were recovered. In particular, our sequence data suggests a possible novel species of *Haloquadratum.* So far, the only described species of the genus is *Haloquadratum walsbyi.* Although lack of phenotypic characterization of these strains makes validation of the organisms complicated, candidate status can be assigned to our 6 metagenomic bins. Through the further analysis of metabolism related genes in the bins described herein isolation strategies can be designed in order to obtain these organisms in pure culture for a formal taxonomic description. As sequencing costs continue to drop, bigger amounts of metagenomic sequence data can be obtained and more novel organisms could be obtained through optimized binning methods.

5.5 Recommendations

- Use of genomic sequence data for attempts in isolation. Our analyses indicate that nontraditional culture strategies must be developed in order to isolate these organisms.
- Sampling in lagoons as well as Candelaria crystallizer ponds to see if other possible novel organisms can be obtained through these methods.

6. Literature Cited

Amoozegar, M. A., Siroosi, M., Atashgahi, S., Smidt, H., & Ventosa, A. (2017). Systematics of Haloarchaea and Biotechnological Potential of their Hydrolytic Enzymes. *Microbiology*, *163*, 623–645. https://doi.org/10.1099/mic.0.000463

Arahal, D. R. (2014). *Whole-Genome Analyses: Average Nucleotide Identity*. *Bacterial Taxonomy* (1st ed., Vol. 41, pp. 103–122). Elsevier Ltd. https://doi.org/10.1016/bs.mim.2014.07.002

Auld, R. R., Myre, M., Mykytczuk, N. C. S., Leduc, L. G., & Merritt, T. J. S. (2013). Characterization of the microbial acid mine drainage microbial community using culturing and direct sequencing techniques. *Journal of Micobiological Methods*, *93*(2), 108–115. https://doi.org/10.1016/j.mimet.2013.01.023

Bag, S., Saha, B., Mehta, O., Anbumani, D., Kumar, N., Hansen, T., … Pedersen, O. (2016). An Improved Method for High Quality Metagenomics DNA Extraction from Human and Environmental Samples. *Nature Publishing Group*, (4), 1–9. https://doi.org/10.1038/srep26775

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A, Dvorkin, M., Kulikov, A. S., … Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barthelson, R., Mcfarlin, A. J., Rounsley, S. D., & Young, S. (2011). Plantagora : Modeling Whole Genome Sequencing and Assembly of Plant Genomes. *PloS One*, *6*(12), 1–8. https://doi.org/10.1371/journal.pone.0028436

Barton, L. L. (1995). *Sulfate-Reducing Bacteria* (p. 347).

Bernhard, A. E., Torre, R. De, Walker, C. B., Waterbury, J. B., & Stahl, D. A. (2005). Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature*, *437*(September), 543–546. https://doi.org/10.1038/nature03911

Bey, B. S., Fichot, E. B., & Norman, R. S. (2011). Extraction of High Molecular Weight DNA from Microbial Mats Representative Results. *Journal of Visualized Experiments*, *53*(July), 3–7. https://doi.org/10.3791/2887

Bolhuis, H., Palm, P., Wende, A., Falb, M., Rampp, M., Rodriguez-Valera, F., … Oesterhelt, D. (2006). The genome of the square archaeon *Haloquadratum walsbyi*: life at the limits of water activity. *BMC Genomics*, *7*, 1–12. https://doi.org/10.1186/1471-2164-7-169

Cabello, P., Roldán, M. D., & Moreno-Vivián, C. (2004). Nitrate reduction and the nitrogen cycle in archaea. *Microbiology*, *150*, 3527–3546. https://doi.org/10.1099/mic.0.27303-0

Caton, T. M., Witte, L. R., Ngyuen, H. D., Buchheim, J. A., & Buchheim, M. A. (2004). Microbial Ecology Halotolerant Aerobic Plains of Oklahoma Heterotrophic. *Microbial Ecology*, *48*(4), 449–462. https://doi.org/10.1007/s00248-004-0211-7

Chevrier, V. F., Hanley, J., & Altheide, T. S. (2009). Stability of perchlorate hydrates and their liquid solutions at the Phoenix landing site, Mars. *Geophysical Research Letters*, *36*(10), L10202. https://doi.org/10.1029/2009GL037497

Crits-Christoph, A., Gelsinger, D. R., Ma, B., Wierzchos, J., Ravel, J., Davila, A., … DiRuggiero, J. (2016). Functional interactions of archaea, bacteria and

viruses in a hypersaline endolithic community. *Environmental Microbiology*, *18*(6), 2064–2077. https://doi.org/10.1111/1462-2920.13259

Cui, Y., Zhang, H., Ding, J., & Peng, Y. (2016). The effects of salinity on nitrification using halophilic nitrifiers in a Sequencing Batch Reactor treating hypersaline wastewater. *Nature Publishing Group*, (April), 1–11. https://doi.org/10.1038/srep24825

Darling, A. E., Jospin, G., Lowe, E., Matsen, F. a, Bik, H. M., & Eisen, J. a. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, *2*, e243. https://doi.org/10.7717/peerj.243

Dyall-Smith, M. L., Pfeiffer, F., Oberwinkler, T., Klee, K., Rampp, M., Palm, P., & Gross, K. (2013). Genome of the Haloarchaeon *Natronomonas moolapensis*, a Neutrophilic Member of a Previously Haloalkaliphilic Genus. *Genome Announcements*, *1*(2), 1–2. https://doi.org/10.1128/genomeA.00095-13.Copyright

Falb, M., Pfeiffer, F., Palm, P., Rodewald, K., Hickmann, V., Tittor, J., & Oesterhelt, D. (2005). Living with two extremes : Conclusions from the genome sequence of *Natronomonas pharaonis*. *Genome Research*, 1336–1343. https://doi.org/10.1101/gr.3952905.

Fendrihan, S., Bérces, A., Lammer, H., Musso, M., & Rontó, G. (2011). Investigating the Effects of Simulated Martian Ultraviolet Radiation on *Halococcus dombrowskii* and Other Extremely Halophilic Archaebacteria. *Astrobiology*, *9*(1), 104–112. https://doi.org/10.1089/ast.2007.0234.Investigating

Feng, Y., Zhang, Y., Ying, C., Wang, D., & Du, C. (2015). Nanopore-based fourth-generation DNA sequencing technology. *Genomics, Proteomics & Bioinformatics*, *13*(1), 4–16. https://doi.org/10.1016/j.gpb.2015.01.009

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2018). DNA – DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, *57*(2007), 81–91. https://doi.org/10.1099/ijs.0.64483-0

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). BIOINFORMATICS APPLICATIONS NOTE Genome analysis QUAST : quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hafenbradl, D., Keller, M., Dirmeier, R., Rachel, R., Roßnagel, P., Burggraf, S., … Stetter, K. O. (1996). *Ferroglobus placidus gen. nov., sp. nov*., a novel hyperthermophilic archaeum that oxidizes Fe 2 + at neutral pH under anoxic conditions. *Archives of Microbiology*, *2*, 308–314.

Hajibabaei, M., Singer, G. a C., Clare, E. L., & Hebert, P. D. N. (2007). Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. *BMC Biology*, *5*, 24. https://doi.org/10.1186/1741-7007-5-24

Hecht, M. H., Kounaves, S. P., Quinn, R. C., West, S. J., Young, S. M. M., Ming, D. W., … Smith, P. H. (2009). Detection of perchlorate and the soluble chemistry of martian soil at the Phoenix lander site. *Science (New York, N.Y.)*, *325*(5936), 64–67. https://doi.org/10.1126/science.1172466

Hoff, K. J., Lingner, T., Meinicke, P., & Tech, M. (2009). Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Research*, *37*(Web Server issue), W101–5. https://doi.org/10.1093/nar/gkp327

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., & Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research*, *32*(Database issue), D277–80. https://doi.org/10.1093/nar/gkh063

Kang, D. D., Froula, J., Egan, R., & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, *3*, e1165. https://doi.org/10.7717/peerj.1165

Khan, S., Zaidi, A., & Ahmad, E. (2014). *Mechanism of Phosphate Solubilization and Physiological Functions of Phosphate-Solubilizing Microorganisms* (pp. 31–64). https://doi.org/10.1007/978-3-319-08216-5

Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing--concepts and limitations. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *32*(6), 524–536. https://doi.org/10.1002/bies.200900181

Ko, Æ. L., Horn, P., Gronau, Æ. S. Von, Gonzalez, Æ. O., Pfeiffer, Æ. F., & Oesterhelt, E. B. Æ. D. (2008). Metabolism of halophilic archaea. *Extremophiles*, *12*, 177–196. https://doi.org/10.1007/s00792-008-0138-x

Konstantinidis, K. T., & Tiedje, J. M. (2005a). Genomic insights that advance the species definition for prokaryotes. *PNAS*, *102*(7), 2567–2572.

Konstantinidis, K. T., & Tiedje, J. M. (2005b). Towards a Genome-Based Taxonomy for Prokaryotes. *Journal of Bacteriology*, *187*(18), 6258–6264. https://doi.org/10.1128/JB.187.18.6258

Konstantinidis, K., Tebbe, A., Klein, C., Scheffer, B., Aivaliotis, M., Bisle, B., … Oesterhelt, D. (2007). Genome-Wide Proteomics of *Natronomonas pharaonis*. *Journal of Proteome Research*, *6*, 185–193. https://doi.org/10.1021/pr060352q

Kottemann, M., Kish, A., Iloanusi, C., Bjork, S., & DiRuggiero, J. (2005). Physiological responses of the halophilic archaeon Halobacterium sp. strain NRC1 to desiccation and gamma irradiation. *Extremophiles : Life under*

Extreme Conditions, *9*(3), 219–227. https://doi.org/10.1007/s00792-005-0437-4

Laver, T., Harrison, J., O'Neill, P. a, Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, *3*, 1–8. https://doi.org/10.1016/j.bdq.2015.02.001

Legault, B. A, López-López, A., Alba-Casado, J. C., Doolittle, W. F., Bolhuis, H., Rodriguez-Valera, F., & Papke, R. T. (2006). Environmental genomics of "Haloquadratum walsbyi" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. *BMC Genomics*, *7*, 171. https://doi.org/10.1186/1471-2164-7-171

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows – Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Litchfield, D. (1998). Survival strategies for microorganisms in hypersaline environments and their relevance to life on early Mars. *Meteor Planet Sci*, *33*, 813–819.

Luo, C., Rodríguez-R, L. M., & Konstantinidis, K. T. (2013). *A user's guide to quantitative and comparative analysis of metagenomic datasets. Methods in enzymology* (1st ed., Vol. 531, pp. 525–547). Elsevier Inc. https://doi.org/10.1016/B978-0-12-407863-5.00023-X

Luo, C., Rodríguez-R, L. M., & Konstantinidis, K. T. (2018). MyTaxa : an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids*, *42*(8), 1–12. https://doi.org/10.1093/nar/gku169

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … Wang, J. (2012). SOAPdenovo2 : an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*(18), 1–6.

Mancinelli, R. L., Landheim, R., Sanchez-Porro, C., & Dornmayr, M. (2009). *Halorubrum chaoviator sp . nov* ., a haloarchaeon isolated from sea salt in Baja California , Mexico , Western Australia and Naxos. *International Journal of Systematic and Evolutionary Microbiology*, *59*, 1908–1913. https://doi.org/10.1099/ijs.0.000463-0.Halorubrum

McHardy, A. C. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods*, *4*(1), 63–72. https://doi.org/10.1038/NMETH976

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., … Edwards, R. a. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, *9*, 386. https://doi.org/10.1186/1471-2105-9-386

Mikheenko, A., Saveliev, V., & Gurevich, A. (2016). Genome analysis MetaQUAST : evaluation of metagenome assemblies. *Bioinformatics*, *32*(7), 1088–1090. https://doi.org/10.1093/bioinformatics/btv697

Mikheyev, A. S., & Tin, M. M. Y. (2014). A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources*, *14*(6), 1097–1102. https://doi.org/10.1111/1755-0998.12324

Mongodin, E. F., Nelson, K. E., Daugherty, S., Deboy, R. T., Wister, J., Khouri, H., … Walsh, D. A. (2005). The genome of Salinibacter ruber : Convergence

and gene exchange among hyperhalophilic bacteria and archaea. *PNAS*, *102*(50), 18147–18152.

Montalvo-Rodríguez, R., Vreeland, R. H., Aharon, O., Martin, K., López-Garriga, J., & Chester, W. (1998). *Halogemetricum borinquense sp . nov* ., a novel halophilic archaeon from Puerto Rico. *International Journal of Systematic Bacteriology*, *48*, 1305–1312.

Montalvo-Rodrıguez, R., Vreeland, R. H., Lopez-garriga, J., Oren, A., Ventosa, A., Kamekura, M., & Chester, W. (2000). *Haloterrigena thermotolerans sp . nov* ., a halophilic archaeon from Puerto Rico. *International Journal of Systematic and Evolutionary Microbiology*, 1065–1071.

Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., … Bork, P. (2010). eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Research*, *38*(Database issue), D190–5. https://doi.org/10.1093/nar/gkp951

Muyzer, G., & Stams, A. J. M. (2008). The ecology and biotechnology of sulphate-reducing bacteria. *Nature Reviews*, *6*, 441–454. https://doi.org/10.1038/nrmicro1892

Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, *40*(20), e155. https://doi.org/10.1093/nar/gks678

Narasingarao, P., Podell, S., Ugalde, J. a, Brochier-Armanet, C., Emerson, J. B., Brocks, J. J., … Allen, E. E. (2012). De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial

communities. *The ISME Journal*, *6*(1), 81–93. https://doi.org/10.1038/ismej.2011.78

Nishimura, T., Vertès, A. A., & Shinoda, Y. (2007). Anaerobic growth of Corynebacterium glutamicum using nitrate as a terminal electron acceptor. *Applied Microbiology and Biotechnology*, *75*, 889–897. https://doi.org/10.1007/s00253-007-0879-y

Niu, B., Fu, L., Sun, S., & Li, W. (2010). Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics*, *11*(187), 1–11.

Offre, P., Spang, A., & Schleper, C. (2013). Archaea in Biogeochemical Cycles. *Annual Review of Microbiology*, *16*(30), 437–457. https://doi.org/10.1146/annurev-micro-092412-155614

Oguchi, H. N., Aniguchi, T. T., & Toh, T. I. (2008). MetaGeneAnnotator : Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. *DNA Research*, *15*, 387–396.

Olson, N. D., Treangen, T. J., Hill, C. M., Cepeda-espinoza, V., Ghurye, J., Koren, S., & Pop, M. (2017). Metagenomic assembly through the lens of validation : recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Briefings in Bioinformatics*, (May), 1–11. https://doi.org/10.1093/bib/bbx098

Oren, A. (2002). Diversity of halophilic microorganisms : Environments , phylogeny , physiology , and applications. *Journal of Industrial Microbiology*, *28*, 56–63.

Oren, A. (2010). Industrial and environmental applications of halophilic microorganisms. *Environmental Technology*, *31*(8-9), 825–834. https://doi.org/10.1080/09593330903370026

Oren, A. (2014). The ecology of *Dunaliella* in high-salt environments. *Journal of Biological Research*, 1–8. https://doi.org/10.1186/s40709-014-0023-y

Oren, A. (2015a). Cyanobacteria in hypersaline environments: biodiversity and physiological properties. *Biodiversity Conservation*, *24*, 781–798. https://doi.org/10.1007/s10531-015-0882-z

Oren, A. (2015b). Halophilic microbial communities and their environments. *Current Opinion in Biotechnology*, *33*(1), 119–124. https://doi.org/10.1016/j.copbio.2015.02.005

Oren, A., Elevi Bardavid, R., & Mana, L. (2014). Perchlorate and halophilic prokaryotes: implications for possible halophilic life on Mars. *Extremophiles : Life under Extreme Conditions*, *18*(1), 75–80. https://doi.org/10.1007/s00792-013-0594-9

Oren, A., Elevi, R., Watanabe, S., Ihara, K., & Corcelli, A. (2002). nov ., and emended description of *Halomicrobium mukohataei*. *International Journal of Systematic and Evolutionary Microbiology*, *52*, 1831–1835. https://doi.org/10.1099/ijs.0.02156-0.The

Ozcan, B., Ozcengiz, G., Coleri, A., & Cokmus, C. (2007). Diversity of Halophilic Archaea From Six Hypersaline Environments in Turkey. *Journal of Microbiology Biotechnology*, *17*(6), 985–992.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM : assessing the quality of microbial genomes recovered

from isolates , single cells , and metagenomes. *Genome Research*, *25*, 1043–1055. https://doi.org/10.1101/gr.186072.114.Freely

Peter, H., Hörtnagl, P., Reche, I., & Sommaruga, R. (2016). Europe PMC Funders Group Bacterial diversity and composition during rain events with and without Saharan dust influence reaching a high mountain lake in the Alps. *Environmental Microbiology*, *6*(6), 618–624.

Podell, S., Emerson, J. B., Jones, C. M., Ugalde, J. A, Welch, S., Heidelberg, K. B., … Allen, E. E. (2014). Seasonal fluctuations in ionic concentrations drive microbial succession in a hypersaline lake community. *The ISME Journal*, *8*(5), 979–990. https://doi.org/10.1038/ismej.2013.221

Puckett, M. K., McNeal, K. S., Kirkland, B. L., Corley, M. E., & Ezell, J. E. (2011). Biogeochemical Stratification and Carbonate Dissolution-Precipitation in Hypersaline Microbial Mats (Salt Pond, San Salvador, The Bahamas). *Aquatic Geochemistry*, *17*(4-5), 397–418. https://doi.org/10.1007/s10498-011-9141-4

Rennó, N. O., Bos, B. J., Catling, D., Clark, B. C., Drube, L., Fisher, D., … Young, S. M. M. (2009). Possible physical and thermodynamical evidence for liquid water at the Phoenix landing site. *Journal of Geophysical Research*, *114*, E00E03. https://doi.org/10.1029/2009JE003362

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Richter, M., & Rosselló, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *PNAS*, *106*(45), 19126–1931.

Rodríguez-García, C. M. (2016). Metagenomic Analysis of Prokaryotic Communities from Hypersaline Environments at Cabo Rojo, Puerto Rico through Pyrosequencing of 16S rRNA Genes. (Master's Thesis). University of Puerto Rico, Mayaguez Campus.

Rodriguez-R, L. M., Konstantinidis, K. T (2014). Identify Bacterial Species. *Microbe*, *9*(3), 111–118.

Rodriguez-R L.M. & Konstantinidis K.T. (2016). The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Preprints 4:e1900v1.

Rodriguez-Valera, F., Rodriguez-, B., Thingstad, T. F., Rohwer, F., & Mira, A. (2009). Explaining microbial population genomics through phage predation. Nature Reviews, 7, 828–836. https://doi.org/10.1038/nrmicro2235

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., … Yorke, J. A. (2012). GAGE : A critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, 557–567. https://doi.org/10.1101/gr.131383.111.22

Sánchez-Nieves, R., Facciotti, M., Saavedra-Collado, S., Dávila-Santiago, L., Rodríguez-Carrero, R., & Montalvo-Rodríguez, R. (2016). Draft genome of Haloarcula rubripromontorii strain SL3, a novel halophilic archaeon isolated from the solar salterns of Cabo Rojo, Puerto Rico. *Genomics Data*, *7*(9001), 287–289. https://doi.org/10.1016/j.gdata.2016.02.005

Sánchez-Porro, C., Martı, S., Mellado, E., & Ventosa, A. (2003). Diversity of moderately halophilic bacteria producing extracellular hydrolytic enzymes. *Journal of Applied Microbiology*, *94*, 295–300.

Schmidt, H., & Woebken, D. (2017). Diversity and activity of nitrogen fixing archaea and bacteria associated with micro-environments of wetland rice. *Geophysical Research Abstracts*, *19*.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, *5*(1), 16–18. https://doi.org/10.1038/NMETH1156

Segerer, A., Neuner, A., Kristjansson, I. J. K., & Stetter, K. (2018). *Acidianus infernus gen nov., sp. nov., and Acidianus brierleyi comb. nov.*: Falcutatively Aerobic, Extremely Acidophilic Thermophilic Sulfur Metabolizing Archaebacteria. *International Journal of Systematic Bacteriology*, *36*(4), 559–564.

Shokralla, S., Spall, J. L., Gibson, J. F., & Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, *21*(8), 1794–1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x

Solomon, S., Kachiprath, B., & Sajeevan, G. J. T. P. (2016). High-quality metagenomic DNA from marine sediment samples for genomic studies through a preprocessing approach. *3 Biotech*, *6*(2), 1–5. https://doi.org/10.1007/s13205-016-0482-y

Soto-Ramírez, N., Sánchez-Porro, C., Rosas, S., González, W., Quiñones, M., Ventosa, A., & Montalvo-Rodríguez, R. (2007). *Halomonas avicenniae* sp. nov., isolated from the salty leaves of the black mangrove *Avicennia germinans* in Puerto Rico. *International Journal of Systematic and Evolutionary Microbiology*, *57*(Pt 5), 900–905. https://doi.org/10.1099/ijs.0.64818-0

Soto-Ramírez, N., Sánchez-Porro, C., Rosas-Padilla, S., Almodóvar, K., Jiménez, G., Machado-Rodríguez, M., … Montalvo-Rodríguez, R. (2008).

Halobacillus mangrovi sp. nov., a moderately halophilic bacterium isolated from the black mangrove *Avicennia germinans*. *International Journal of Systematic and Evolutionary Microbiology*, *58*(Pt 1), 125–130. https://doi.org/10.1099/ijs.0.65008-0

Stahl, D. A., & Torre, R. De. (2012). Physiology and Diversity of Ammonia-Oxidizing Archaea. *Annual Review of Microbiology*, *66*, 83–101. https://doi.org/10.1146/annurev-micro-092611-150128

Streit, W. R., & Schmitz, R. A. (2004). Metagenomics--the key to the uncultured microbes. *Current Opinion in Microbiology*, *7*(5), 492–498. https://doi.org/10.1016/j.mib.2004.08.002

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V, … Natale, D. A. (2003). The COG database : an updated version includes eukaryotes. *BMC Bioinformatics*, *14*, 1–14.

Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, *2*(1), 3. https://doi.org/10.1186/2042-5783-2-3

Tourova, T. P., Kovaleva, O. L., Sorokin, D. Y., & Muyzer, G. (2010). Ribulose-1,5 biphosphate carboxylase/oxygenase genes as a functional marker for chemolithoautotrophic halophilic sulfur-oxidizing bacteria in hypersaline habitats. *Microbiology*, *156*, 2016–2025. https://doi.org/10.1099/mic.0.034603-0

Tseng, C., Chiang, P., Shiah, F., Chen, Y., Liou, J., Hsu, T., … Halgamuge, S. (2013). Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances. *The ISME Journal*, *7*(12), 2374–2386. https://doi.org/10.1038/ismej.2013.118

Ventosa, A. (2006a). Unusual micro-organisms from unusual habitats : hypersaline environments. In N. A. Logan, H. M. Lappin-Scott, & P. C. F. Oyston (Eds.), *Prokaryotic Diversity - Mechanisms and Significanse* (pp. 223–253). Cambridge: Cambridge University Press.

Ventosa, A. (2006b). *Unusual microorganisms from unusual habitats: hypersaline environments.* In: Logan NA, Lappin-Scott HM, Oyston PCF (eds) Prokaryotic diversity-mechanism and significance. Cambridge University Press, Cambridge, pp 223–253

Ventosa, A., de la Haba, R. R., Sánchez-Porro, C., & Papke, R. T. (2015). Microbial diversity of hypersaline environments: a metagenomic approach. *Current Opinion in Microbiology*, *25*, 80–87. https://doi.org/10.1016/j.mib.2015.05.002

Ventosa, A., Fernández, A. B., León, M. J., Sánchez-Porro, C., & Rodriguez-Valera, F. (2014). The Santa Pola saltern as a model for studying the microbiota of hypersaline environments. *Extremophiles : Life under Extreme Conditions*, *18*(5), 811–824. https://doi.org/10.1007/s00792-014-0681-6

Ventosa, A., Márquez, M. C., Garabito, M. J., & Arahal, D. R. (1998). Moderately halophilic gram-positive bacterial diversity in hypersaline environments. *Extremophiles*, *2*, 297–304.

Ventosa, A., Mellado, E., Sánchez-Porro, C., & Marquez, M. C. (2008). Halophilic and Halotolerant Micro-Organisms from Soils. In P. Dion & C. S. Nautiyal (Eds.), *Microbiology of Extreme Soils* (pp. 87–115). Berlin: Springer-Verlag.

Walker, C. B., Torre, J. R. De, Klotz, M. G., Urakawa, H., Pinel, N., Arp, D. J., … Chain, P. S. G. (2010). *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine

crenarchaea. *PNAS*, *107*(19), 8818–8823. https://doi.org/10.1073/pnas.0913533107

Wu, Y., Tang, Y., Tringe, S. G., Simmons, B. A., & Singer, S. W. (2014). MaxBin : an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, *2*(26), 1–18.

Yadav, A. N., Sharma, D., Gulati, S., Singh, S., & Dey, R. (2015). Haloarchaea Endowed with Phosphorus Solubilization Attribute Implicated in Phosphorus Cycle. *Nature Publishing Group*, (October 2014), 1–10. https://doi.org/10.1038/srep12293

Yannarell, A. C., Steppe, T. F., & Paerl, H. W. (2007). Disturbance and recovery of microbial community structure and function following Hurricane Frances. *Environmental Microbiology*, *9*(3), 576–583. https://doi.org/10.1111/j.1462-2920.2006.01173.x

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. https://doi.org/10.1101/gr.074492.107

Zhaxybayeva, O., Stepanauskas, R., Ram, N., & Thane, M. R. (2013). Cell sorting analysis of geographically separated hypersaline environments. *Extremophiles*, *13*. https://doi.org/10.1007/s00792-013-0514-