UNIVERSITY OF CALIFORNIA, SAN DIEGO

Metagenomic Characterization of a Hypersaline Microbial Ecosystem

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Marine Biology

by

Juan A. Ugalde

Committee in charge:

Professor Eric E. Allen, Chair
Professor Farooq Azam
Professor Douglas H. Bartlett
Professor. Philip Bourne
Professor David T. Pride
Professor Forest Rohwer

2014

The dissertation of Juan A. Ugalde is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2014

DEDICATION

A mi familia, sin su apoyo nada de esto hubiera sido posible.

EPIGRAPH

A beginning is the time for taking the most delicate care that the balances are correct.

—Frank Herbert, *Dune*.

This is not the place to go into the specifies of which microbial genomes would be most useful. I would suggest, however, that a phylogenetic tree hang on the wall of every laboratory in which microbial genomes are being sequenced for inspiration.

—Carl Woese, *A manifesto for microbial genomics*.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This work could not have been posible without the support of numerous persons over the last few years. I feel very fortunate for the opportunity that I had to met such amazing people over my years in San Diego.

First, I have to thank Eric Allen, for giving me the opportunity to join his research group. He provided me with enough guidance, but at the same time, with enough freedom to pursue multiple projects and ideas. His enthusiasm provided inspiration for this work, and for all my future projects. I also want to thank all my committee members, Farooq Azam, Douglas Bartlett, Philip Bourn, David Pride, and Forest Rohwer, for their comments, feedback and support through my scientific endeavors.

Multiple funding sources helped me along the years. In particular I want to acknowledge the financial support provided by a Fulbright-Conicyt fellowship during my first four years of graduate school.

I also have to thank all of the current members of the Allen Lab. In particular Sheila Podell, for all the help both with bioinformatics and scientific questions, and for being an amazing officemate. Also I have to thank Christine Shulse, for being such a great guide and for her friendship during my first years of grad school, and Jessica Blanton, for all the support and friendship during my last year of graduate school.

I am very grateful of the support of my friends outside the SIO community. In particular, I have to thank Jorge and Paula, for their constant support and friendship, from the first day that I arrived in San Diego.

Of course, nothing of this could have been possible without the support of my family. Thanks for always supporting me and believe in me. And of course, most of all, I want to thank Andrea, for her constant support, company, and patience. Without you, none of this could have been possible. Thank you.

Chapter 2 is a full reprint of: De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline communnities. P. Narasingarao, S. Podell, J.A. Ugalde, C. Brochier-Armanet, J.B. Emerson, J.J. Brocks, K.B. Heidelberg, J.F. Banfield and E.E. Allen. *ISME Journal*, **6**,81-93. 2012 (doi: 10.1038/ismej.2011.78), with permission from all coauthors.

Chapter 3 is a full reprint of: Xenorhodopsins, an enigmatic new class of microbial rhodopsins horizontally transferred between archaea and bacteria. J.A. Ugalde, S. Podell, P. Narasingarao and E.E. Allen. *Biology Direct*, **6**,52. 2011 (doi: 10.1186/1745-6150-6-52), with permission from all coauthors.

Chapter 4 is a full reprint of: Assembly-driven community genomics of a hypersaline microbial ecosystem. S. Podell, J.A. Ugalde, P. Narasingarao, J.F. Banfield, K.B. Heidelberg and E.E. Allen. *PLoS One*, **8**4:e61692. 2013 (doi: 10.1371/journal.pone.0061692), with permission from all coauthors.

Appendix A is a full reprint of: Draft Genome Sequence of "*Candidatus* Halobonum tyrrellensis" Strain G22, Isolated from the Hypersaline Waters of Lake Tyrrell, Australia. J.A. Ugalde, P. Narasingarao, S. Kuo, S. Podell and E.E. Allen. *Genome Announcements*, **1**(6):e01001-13. 2013. With permission from all coauthors.

VITA

2001-2003	Research assistant, Laboratory of Bioinformatics and Gene Expression, Universidad de Chile, Santiago, Chile.
2003-2004	Research assistant, Whitney Laboratory for Marine Biosciences, University of Florida.
2006	*Licenciatura* in Molecular Biotechnology Engineering. Universidad de Chile, Santiago, Chile.
2004-2008	Project Engineer, Laboratory of Bioinformatics and Mathematics of the Genome, Universidad de Chile, Santiago, Chile.
2008-2012	Fulbright fellow.
2014	Doctor of Philosophy, Scripps Institution of Oceanography, University of California, San Diego

PUBLICATIONS

Martin LJ, Adams R, Bateman A, Bik HM, Haws J, Hird SM, Hughes D, Kembel SW, Kinney K, Kolokotronis SO, Levy G, Mclain C, Meadow JF, Medina RF, Mhuireach G, Moreau CS, Munshi-South J, Nichols, LM, Palmer C, Popova L, Schal C, Siegel J, Taubel M, Trautwein M, Ugalde JA, Dunn RR. Evolution in the Indoor Biome. *Proc R Soc B. Accepted.*

Valenzuela C, Ugalde JA, Mora GC, Alvarez S, Contreras I, Santiviago CA. Draft Genome Sequence of *Salmonella enterica* Serovar *Typhi*. *Genome Announc* 2(1): e00104-14. 2014.

Malfatti F, Turk V, Tinta T, Mozetič P, Manganelli M, Samo TJ, Ugalde JA, Kovač N, Stefanelli M, Antonioli M, Fonda-Umani S, Del Negro P, Cataletto B, Hozić A, Ivošević DeNardis N, Mišić Radić T, Radić T, Fuks D, Azam F. Microbial mechanisms coupling carbon and phosphorus cycles in phosphorous-limited northern Adriatic Sea. *Sci Total Environ* 470: 1173-1183. 2014.

Ugalde JA, Narasingarao P, Kuo P, Podell S, Allen EE. Draft genome sequence of "*Candididatus* Halobonum tyrrellensis" Strain G22, Isolated from the Hypersaline Waters of Lake Tyrrell, Australia. *Genome Announc* 1(6): e01001-13. 2013.

Podell S, Emerson JB, Jones, CM, Ugalde JA, Welch, S, Heidelberg KB, Banfield JF, Allen EE. Seasonal fluctuations in ionic concentrations drive microbial succession in a hypersaline lake community. *ISME J*, advance online publication. doi:10.1038/ismej.2013.221.

Kharbush JJ, Ugalde JA, Hogle SL, Allen EE, Aluwihare LI. Composite bacterial hopanoids and their microbial producers across oxygen gradients in the water column of the California Current. *Appl Env Microbiol* 79(23): 7491-7501. 2013.

Ugalde JA, Gallard MJ, Belmar C, Muñoz P, Ruiz-Tagle N, Ferrada-Fuentes S, Espinoza C, Allen EE, Gallardo VA. Microbial Life in a Fjord: Metagenomic Analysis of a Microbial Mat in Chilean Patagonia. *PLoS One* 8(8):e71952. 2013.

Podell S, Ugalde JA, Narasingarao P, Banfield JF, Heidelberg EE. Assembly-driven community genomics of a hypersaline microbial ecosystem. *PLoS One* 8(4):e61692. 2013.

Narasingaro P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, Heidelberg KB, Banfield JF, Allen EE. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *ISME J* 6:81-93. 2012.

Ugalde JA, Podell S, Narasingarao P, Allen EE. Xenorhodopsins, an enigmatic new class of microbial rhodopsins horizontally transferred between Archaea and Bacteria. *Biol Direct* 6:52. 2011.

Levicán G, Ugalde JA, Ehrenfeld N, Maass A, Parada P. Comparative genomic analysis of carbon and nitrogen assimilation mechanisms in three indigenous bioleaching bacteria: predictions and validations. *BMC Genomics* 9(1):581. 2008.

Chang BSW, Ugalde JA, Matz MV. Applications of ancestral protein reconstruction in understanding protein function: GFP-like proteins. *Meth Enyzmol* 395:652-670. 2005.

Matz MV, Labas YA, Ugalde J. Evolution of function and color in GFP-like proteins. *Method of Biochemical Analysis, Green Fluorescent Protein.* Chalfie M, Kain SR, Eds. John Wiley & Sons. 2005.

Ugalde JA, Chang BSW, Matz MV. Evolution of Coral Pigments Recreated. *Science* 305(5689): 1433. 2004.

Shagin DA, Barsova EV, Yanushevich YG, Fradkov AF, Lukyanov KA, Labas YA, Semenova TN, Ugalde JA, Meyers A, Nunez JM, Widder EA, Lukyanov SA, Matz MV. GFP-like proteins as ubiquitous metazoan superfamily: evolution of functional features and structural complexity. *Mol Biol Evol* 21(5): 841-850. 2004.

ABSTRACT OF THE DISSERTATION

Metagenomic Characterization of a Hypersaline Microbial Ecosystem

by

Juan A. Ugalde

Doctor of Philosophy in Marine Biology

University of California, San Diego, 2014

Professor Eric E. Allen, Chair

The use of metagenomic approximations to study natural microbial communities has allowed us to understand the phylogenetic and functional composition of these communities. Gene fragments, derived from the sequence reads, can be phylogenetically classified by binning methods, and compared against reference databases for functional assignments. However, these classifications are limited by the availability of reference genomes in the databases. One way to overcome such limitations, is through the use of assembly-based metagenomics, where the *de novo* sequence assembly, which does not rely on external reference sequences, can allow us to identify novel organisms that are present in the community.

The work presented here, shows the results of the assembly-based metagenomic characterization of a hypersaline microbial community, from Lake Tyrrell,

Australia. The main objective was the reconstruction of the most abundant members of this microbial community using the sequence information, and generate a set of habitat-specific genomes that can be used for future studies.

The assembly and characterization of this metagenomic dataset allowed the discovery of a novel archaeal Class, the *Nanohaloarchaea* (Chapter 2), an ubiquitous lineage late found to be present in other hypersaline environments.

The work on the *Nanohaloarchaea*, led to the discovery a novel type of rhodopsin protein, Xenorhodopsins (Chapter 3), that is present on the genomes of members of this group. Phylogenetic analysis indicates that this new rhodopsin has been horizontally transferred between Bacteria and Archaea.

The complete assembly analysis of the metagenome dataset, allowed the description of the most abundante members present in this microbial community (Chapter 4), allowing estimations of relative abundance, phylogenetic and functional diversity.

With the availability of habitat-specific genomes, it is possible to study not only the phylogenetic and functional diversity, but also the fine-scale genetic diversity of the members of the community. Deep sampling of four samples from the Lake Tyrrell microbial community, using high-throughput sequencing, and the availability of habitat-specific genomes allowed the characterization of this genetic diversity (Chapter 5). This information was used to compare the genetic diversity between populations, and identify signatures of environmental adaptation at the sequence level.

Chapter 1

Introduction to the Thesis

Bacteria and Archaea represent an abundant component of Earth's biomass, with estimates between 9.2-31.7 10^{29} cells [53] to 41.8-64.3 10^{29} cells [133] globally. Species diversity estimates suggest that there are close to 10^7 different species of Bacteria and Archaea[21], although some authors suggest this number is an overestimation [108]. Nevertheless, Bacteria and Archaea represent the most diverse group of organisms on Earth, and yet we have scarcely begun to characterize this diversity [141, 102].

Our inability to cultivate most of the microorganisms present in the environment limits our understanding of the phylogenetic and functional diversity of microbial communities. Known as the "Great Plate-Count Anomaly" [115], our current culture collections may not be representative of what can be observed in natural samples by culture-independent methods [5]. Most organisms are currently uncultured, due to our lack of information on their physiology and nutritional requirements [116]. With new and unique cultivation methods, and information derived from genomic surveys [122], we may be able to culture these organisms in the near future.

Over the last decade, the development of advanced technologies and experimental methods has allowed the study of natural microbial communities via culture-independent techniques. Specifically, DNA-based methods, driven by the development of next-generation sequencing [68], have allowed the investigation of natural microbial communities without relying on cultivation. DNA-based surveys

span two broad categories: marker-based analysis (such as the 16S rRNA gene) and whole genome sequencing.

Marker-based analysis provides a characterization,or census, of the phylogenetic diversity present in the community, allowing for estimation of community diversity[15, 99, 14] and identification of community members based on sequence similarity against a reference database [72]. This approach provides only a partial picture of the metabolic repertoire present in natural microbial communities, as it requires extrapolation of genomic data from isolated microorganisms to the entire community. Even strains with very similar 16S rRNA sequences may have different metabolic and genetic properties, which are not captured in the diversity found using a marker gene such as the 16S rRNA gene [49].

Metagenomics, also referred to as community genomics, relies on the direct sequencing of genetic material isolated from a microbial community *en masse* [140], and provides the means to evaluate the phylogenetic and functional diversity of microbial populations directly from the environment. Depending on the complexity of the community under study, not all of the members of the microbial community will be observed in the sequence data, as the most abundant members will dominate the metagenomic analyses [10]. Using metagenomic methods, several authors report studies of microbial communities with low to moderate species diversity (e.g. [123, 96, 25, 121, 13]). Highly diverse communities, such as those found in soils and sediments, require high amounts of sequence information to provide a full picture of the microorganisms that are present [52]. The decrease in sequencing cost, driven by the development of novel sequencing technologies, has enhanced our ability to deeply sample complex communities using metagenomics approaches. Consequently, the challenge will be developing or acquiring the computational resources and algorithms necessary to analyze vast amounts of genetic sequence information [92, 52].

Microbial communities in extreme environments, such as those associated with hypersaline habitats, represent tractable systems that can be studied using metagenomic approaches. Because of their relatively low species diversity [6, 96, 40], it is possible to study the phylogenetic and metabolic diversity that is present in

the community. In this first chapter, I provide an overview of metagenomics, with particular emphasis in assembly-based approximations, followed by an introduction into the microbiology of hypersaline environments and finally an overview of our study site, the hypersaline waters of Lake Tyrrell located in Victoria, Australia.

1.1 Metagenomics

The study of natural microbial communities by analyzing DNA obtained directly from environmental samples was first proposed by Handelsman *et al.* in 1998 [42] to access the genomic information of uncultured environmental microorganisms. The first metagenomic studies were performed by isolating DNA from environmental samples, plasmid-based cloning of this DNA and subsequence sequencing of these clones using the Sanger method [121, 126]. Limited throughput, cloning biases, and the expesive and laborious nature of this process limited the scope of early metagenomic studies, often focused on microbial communities with low to medium species diversity [121, 96], or to studies aimed at describing the general functional and phylogenetic composition of a community [126, 106]. The development of *next generation* sequencing technologies [69], removed some of these limitations, allowing for direct sequencing of DNA samples without cloning, higher throughputs, and a lower cost per base [69]. One of the remaining limitations in current technologies is the length of the sequence reads, which are not yet close to the length of reads generated by Sanger sequencing. Improvements to current methods and the development of further advanced technologies, such as single molecule sequencing [31] and nanopores [11], holds promise to further alleviate these limitations.

Metagenomic studies are driven, in most cases, by discovery, data mining and comparative research, rather than by a specific hypothesis [140]. Accordingly, both the sequence information and the contextual environmental data, or metadata, associated with a sample are of great importance, allowing for both biological and physicochemical context to be applied to the microbial community under study. These data include environmental parameters associated with a sample and

procedures used to process a sample (e.g. filtration procedure, sample processing and library construction). All of these variables become highly important in the bioinformatic analysis of a sample data set [79].

Based on the complexity of the microbial community under study, and the choice of sampling and sequencing methods, two main approaches are used to study microbial communities using metagenomics: gene-centric and assembly-based approaches (Figure 1.1). Gene-centric studies focus on complex communities or projects with a shallow level of sequencing, where assembly of larger contiguous genome fragments from community members is not feasible given the number of reads or size of the data set obtained. In these cases, the focus is on the phylogenetic and functional diversity as a profile of the community [123, 13], and the comparison of different environments based on these profiles [25, 138, 100]. Gene-centric studies focus on the phylogenetic and functional classification of the metagenomic reads, either by analyzing every read present in the dataset [9, 89, 91], or by using marker-genes as a proxy to create either taxonomic and/or functional profiles of the community [22, 110]. One main drawbacks of this method is that community profiles are limited by the reference databases being used, which can lead to missing unique groups that are present in the microbial community under study.

Assembly-based metagenomics, involves the bioinformatic assembly of the community sequence information, with the goal of recovering larger contiguous genome fragments that can improve the phylogenetic and functional classification of the organisms in the community. This approach allows the recovery of longer fragments, generate better gene models and can identify previously unknown microbial groups [79, 96, 54, 24]. The information obtained from the assembled population genomes can be used to potentially guide cultivation efforts for these organisms [122], in population genetic studies [114, 130, 4], and provide a better picture of the interactions among the members of the microbial community [121, 36]. In addition, the recovery of near-complete genomes from metagenomics datasets allows for the discovery of novel functions that may be missed by viewing the unassembled sequence reads. For example, Chapter 2 of this thesis shows

how the assembly of sequence reads from the Lake Tyrrell metagenome led to the discovery of a previously unknown Phylum of Archaea and to the discovery of a novel family of microbial rhodopsins, called the xenorhodopsins, and evidence that suggests its acquisition through horizontal gene transfer (Chapter 3).

Assembly-based methods are limited by the complexity of the microbial community under study, and the amount of sequencing needed to recover the genomes of the most abundant members of the community. Furthermore, some of the more rare members of the community are inaccessible by sequencing, because the most abundant organisms dominate the metagenomic dataset. For example, estimations of the number of sequences needed to address highly diverse samples, such as from soils, show that billions of sequences are needed to be able to sample some of these most abundant organisms [45]. Even in simpler systems, the main challenge is to classify the assembled genomic fragments into unique phylogenetic bins, each representing a different population. By combining various pieces of information, such as sample characteristics, nucleotide composition, amino acid counts, and library abundance, it is possible to classify these genomic fragments into unique populations, a process known as phylogenetic binning [96, 1]. Larger datasets represent a computational challenge because of the high memory requirements of short-read assembly software programs. In this case, the use of methods for digital normalization may reduce the computational problems [92, 52].

The work presented in this thesis describes a study of the microbial community inhabiting the hypersaline waters of Lake Tyrrell, Australia using an assembly-based approach. The combination of a relatively low-species diversity, driven by the extreme conditions found in this habitat [6], and a deep sequencing approach, allowed for the genomic reconstruction of some of the most abundant members of the community and the discovery of novel microorganisms.

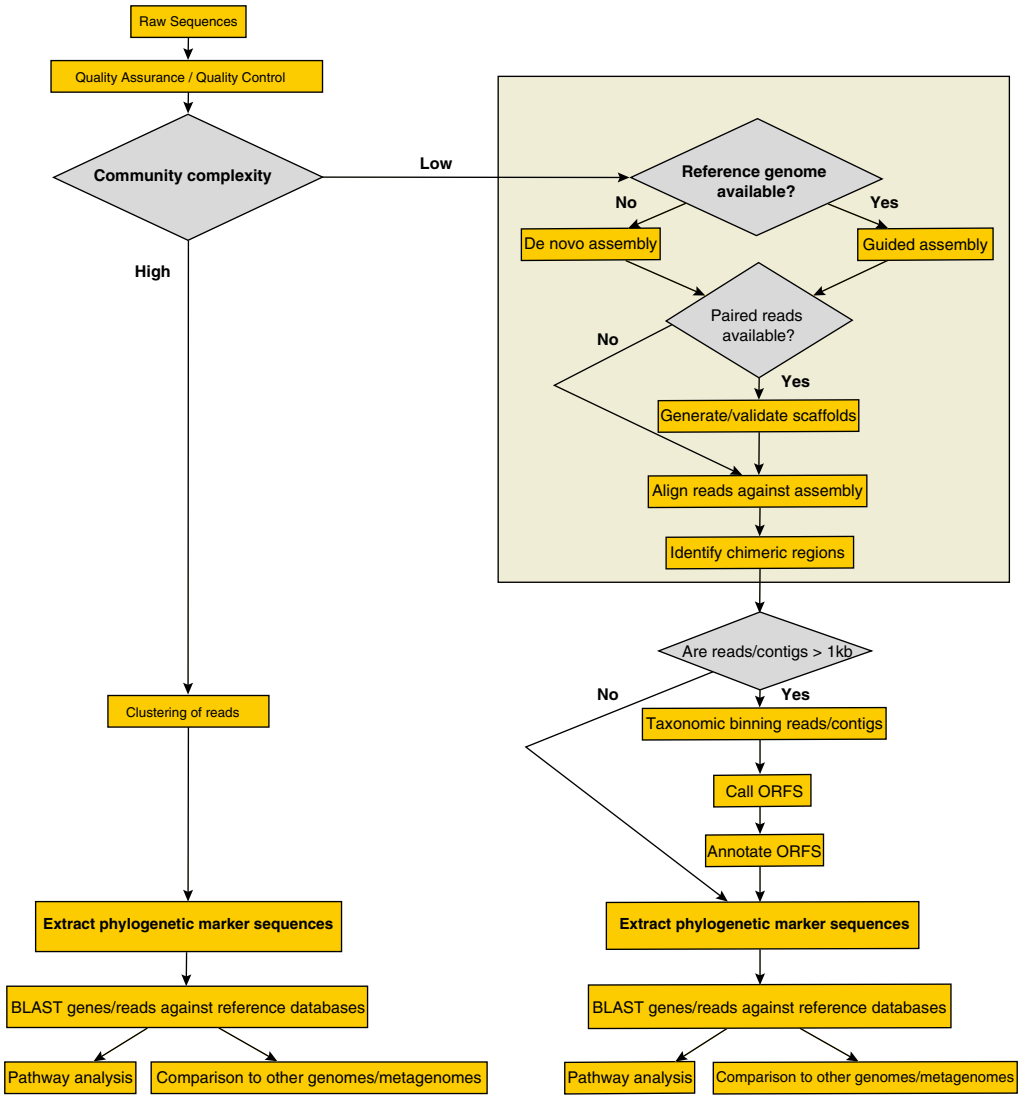


Figure 1.1: Diagram comparing two possible approaches for the analysis of metagenomic data from natural microbial communities. Figure from Bragg and Tyson, 2014 [10].

1.2 Microbial communities in hypersaline environments

Hypersaline habitats are found worldwide, in a variety of natural and man-made environments. Examples include salt lakes, saline soils, salt ats, solar salterns, brine pools, salted foods, and fermented foods [82]. Aquatic hypersaline systems are the most studied, and are either of marine origin (thalassohaline), or formed by the dissolution of mineral salt deposits (athalassohaline).

Within these saline systems, a variety of microbial species are adapted to these environments. Moderate halophiles can be found between 30-150 g/L NaCl, while extreme halophiles exist in the range of 150-300 g/L NaCl [6]. In addition to the high NaCl concentrations, other salts are also important to consider when measuring the ionic composition of these systems, including Mg^{2+} and Ca^{2+}, which can inuence the microbial community in these habitats [74, 95] (Figure 1.4).

Genera from both the Bacteria and the Archaea exist in moderate and extreme saline systems. Within the Archaea, the phylogenetic diversity appears to be limited to the *Euryarchaeota*, specifically in the classes *Halobacteria* and *Methanomicrobia* (Table 1.1) [127]. The discovery of the *Nanohaloarchaea* (described in Chapter 2, and [79]), a novel Class, within the Euryarchaeota,, in globally dispersed hypersaline systems expanded our understanding of the phylogenetic diversity of halophilic Archaea [79]. Recently, a phylogenomic analysis of novel archaeal groups, isolated via single-cell genomics, suggested that the *Nanohaloarchaea* are a new phylum, sister to the *Euryarchaeota* [102], although more work (and more genomes and isolates) is needed to fully resolve the phylogenetic relationships between these groups [136].

Even within the *Halobacteria*, novel taxonomic groups remain unidentified. Our group recently described the genome of a newly isolated halophilic Archaea, *Candidatus* Halobonum tyrrellensis strain G22 [124], which phylogenetic analysis suggests is a new genus within the *Halobacteria* (Appendix A). Analysis of the 16S rRNA gene, and a phylogenomic approach using the markers genes implemented in the software, PhyloPhlan, has further supported this phylogenetic placement

(Figures 1.2 & 1.3).

The phylogenetic breadth of bacterial species found in saline systems is wider than that of Archaea (Table 1.2). In moderate saline environments, Bacteria are more abundant than Archaea [83, 40, 39, 16], but as salinity increases, Archaeal groups become more abundant [40, 16, 78]. One of the most abundant bacterial species found in extreme hypersaline systems is *Salinibacter ruber* [7]. This bacterium shares many phenotypic characteristics with halophilic Archaea [87], and multiple gene clusters appear to have been acquired via horizontal gene transfer from the Archaea [76].

Viruses are also very abundant in hypersaline systems, with reports of counts of at least 10^7 viral-like particles per mL [29]. Viruses could be playing a dual role in these systems; as predators, contributing to the cycling of nutrients through cell lysis; and as a form of information storage, by providing access to an auxiliary gene pool that can be utilized by other microorganisms [62, 137, 48]. Studies of viral dynamics in hypersaline systems have showed that they play a fundamental role in shaping the population structure of microbial communities [103, 104].

Microorganisms that live in hypersaline systems require particular physiological strategies to deal with the high salt concentrations and the potential osmotic pressure gradient that can generate across the cell membrane. Two different osmotic adaptation strategies can be found in halophiles: a salt-in strategy, which involves the accumulation of inorganic ions in the cytoplasm; and a salt-out strategy, that involves pumping ions out the cell and the accumulation of compatible solutes in the cytoplasm [86]. The salt-in strategy is found in the archaeal populations, specifically within the Order *Halobacteriales*, in Bacteria of the Order *Halanerobiales* and in *S. ruber* [86]. These organisms accumulate inorganic ions, such as K^+ and Cl^-, which requires special enzymatic adaptations reected through protein amino acid compositional changes that favor acidic amino acids, such as glutamate and aspartate. Based on their amino acid composition profiles, it has been suggested that the uncultured members of the Class *Nanohaloarchaea* also utilize this strategy. The salt-out strategy is found in most of the halophilic Bac-

teria and halophilic methanogenic Archaea [86]. These organisms have outward-directed sodium transporters that pump the Na^+ ions out of the cytoplasm, but more importantly, they accumulate a large number of organic solutes to mantain the osmotic potential in the cytoplasm [86].

Halophilic microorganisms show a diverse suite of metabolic capabilities. Several dissimilatory metabolic pathways have been described in halophiles (Figure 1.5) across a wide range of salinity concentrations. The majority of bacterial and archaeal groups isolated from hypersaline sites are aerobic chemoorganotrophs. In addition, as oxygen has a low solubility in salt saturated brines [113], several anaerobic energy pathways are available, including electron acceptor substrates such as nitrate, fumarate, and dimethyl sulfoxide [83, 86]. Primary productivity in saline systems varies according to the salinity. In moderate saline environments (between 100 to 250 g/L), primary productivity occurs in microbial mats dominated by members of the *Cyanobacteria*; in high salt environments, the primary producers are planktonic algae of the genus *Dunaliella* [84].

Another important characteristic of halophilic organisms, particularly among the halophilic Archaea, is the presence of microbial rhodopsins, photoactive proteins found in all three domains of life. Rhodopsins serve as light-driven proton pumps (bacteriorhodopsin), chloride pumps (halorhodopsin), or phototactic and photophobic receptors (sensory rhodopsins) [12]. Halorhodopsins and sensory rhodopsins appear to be limited in their distribution to the *Halobacteria*, with just a few examples found in other organisms [112]. Chapter 3 describes the discovery a novel type of microbial rhodopsin, called the xenorhodopsin, which was found in the genome of the *Nanohaloarchaea*, and that appear to broadly dispered via horizontal gene transfer between Archaea and Bacteria (although it is not possible yet to establish the directionality of acquisition).Indeed, the closest homolog to the Nanohaloarchaea xenorhodopsin is a putative sensory rhodopsin found in the cyanobacterium *Anabaena* sp. PCC 712 [128, 125].

Even in highly characterized microbial communities [83, 85]], recent studies using culture independent techniques, such as metagenomics and single-cell genomics, recovered novel microbial groups [79, 40, 65]. The limited microbial

diversity, driven by extreme salinity systems, makes them ideal model systems to study microbial diversity using metagenomic methods. We can fully characterize a microbial community using deep-sequencing approaches, with the goal to reconstruct the genomes of each member of the community. Through this approach, we identify not only the functional and phylogenetic diversity of the community, but also the association of such functions to members of the community. Lastly, it allows for the study of population genetics within the community, with the goal of understanding how these organisms adapt and respond to variations in the environment.

It is important to highlight that other saline ecosystems have been studied using metagenomic approaches. Prior hypersaline metagenomic studies have focused on the population genomics of single species, such as *Haloquadratum walsbyi* [60] and *Salinibacter ruber* [90], or on the dynamic changes of the microbial and viral diversity over salinity gradientes and over time [138, 103]. Only recent studies have explored the microbial and viral diversity of these systems by using assembly-based metagenomics and single-cell genomics approaches [79, 96, 40, 39, 33, 34].

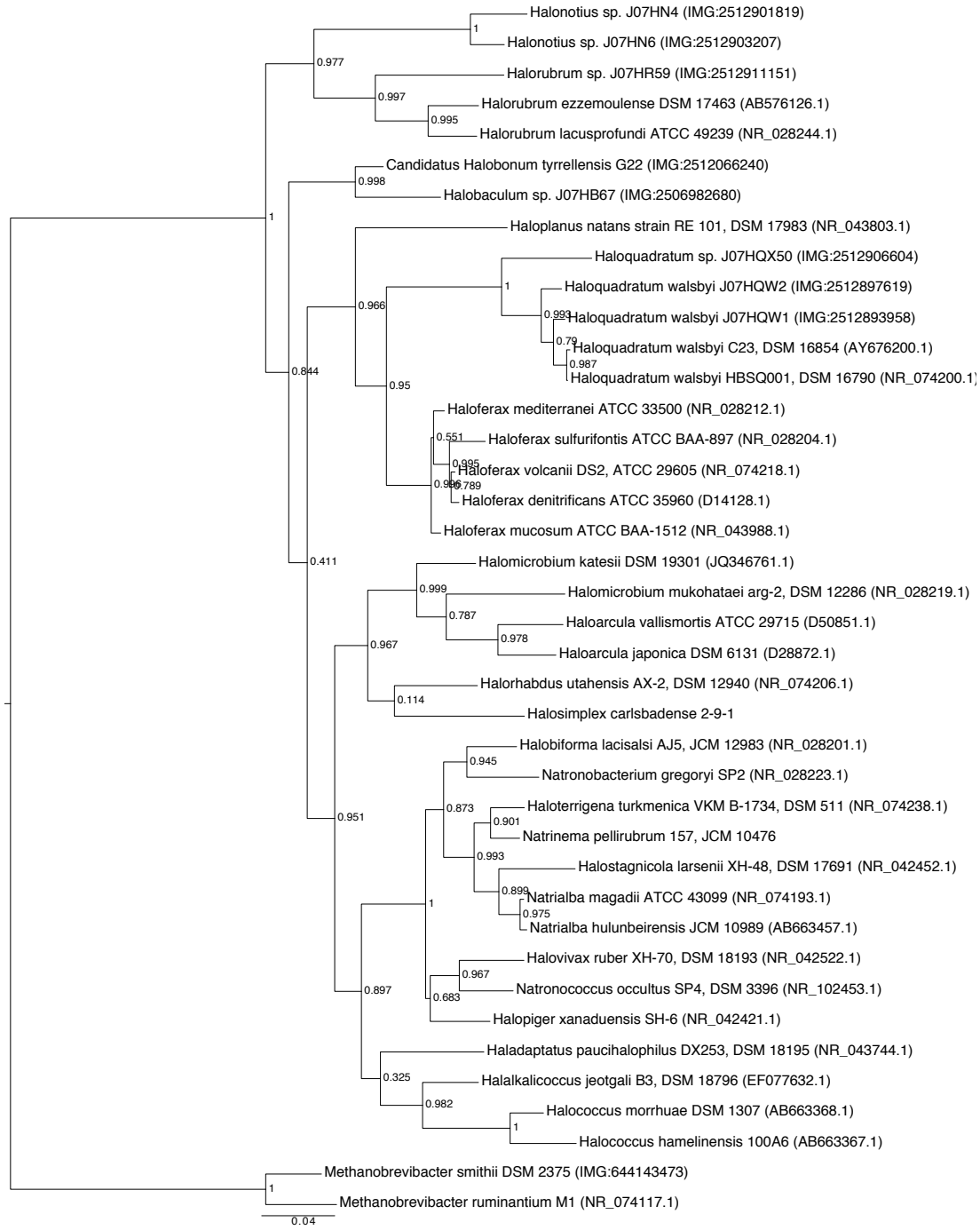


Figure 1.2: Phylogenetic tree of *Candidatus* Halobonum tyrrellensis G22 and related microorganisms, based on 16S rRNA sequences.

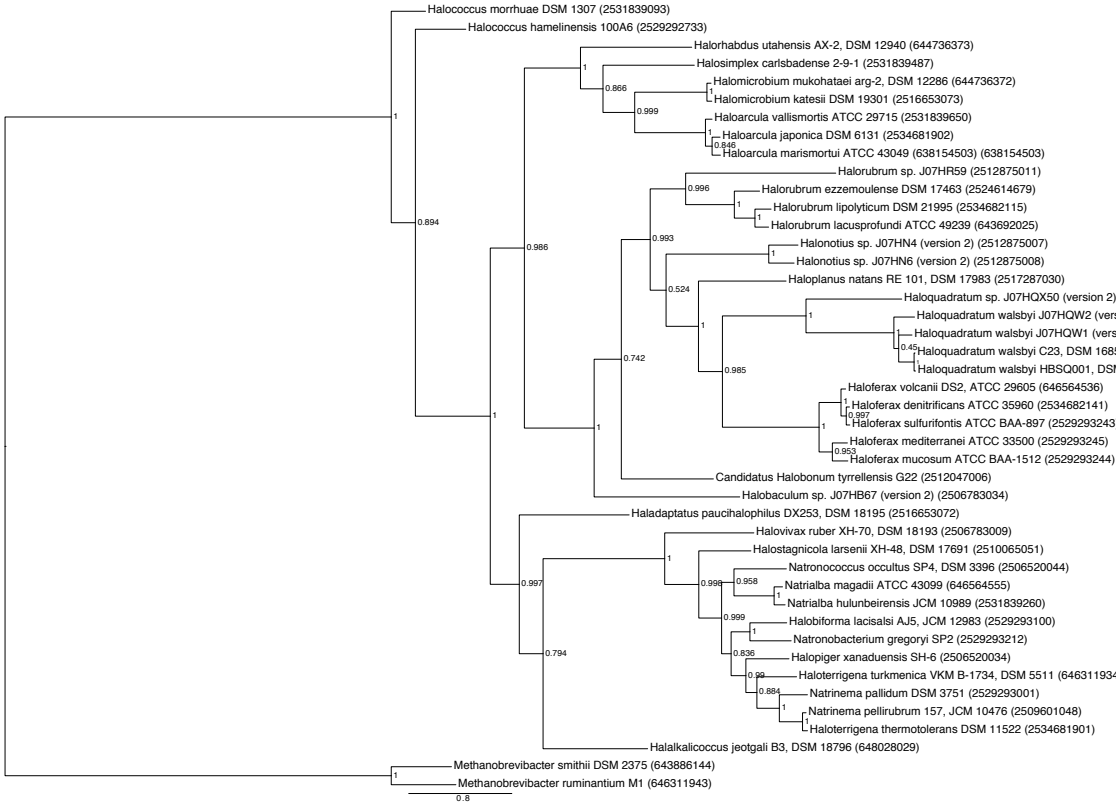


Figure 1.3: Phylogenetic tree of *Candidatus* Halobonum tyrrellensis G22 and related microorganisms, based on phylogenetic marker sequences implemented in the Phylophlan software [109]

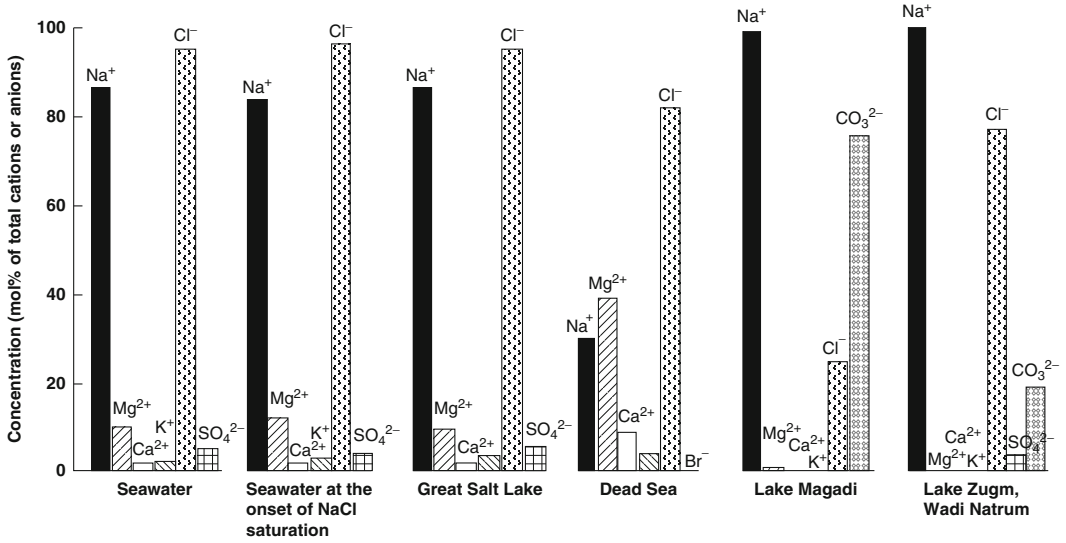


Figure 1.4: Ionic composition of several aquatic systems. Figure from Oren, 2013. [86]

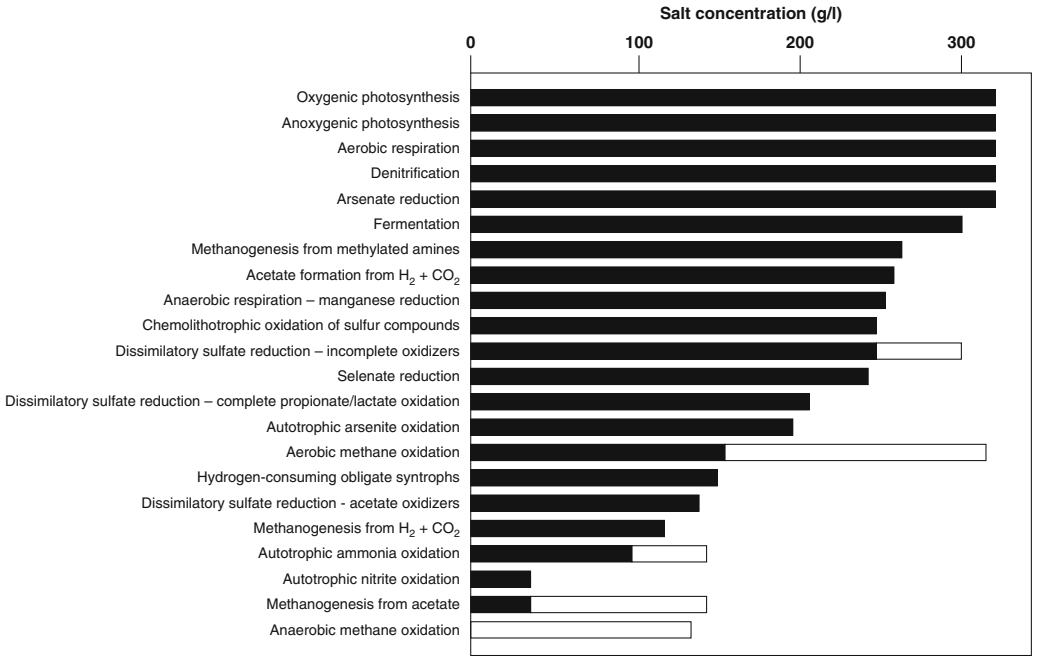


Figure 1.5: Salt concentration limits for some microbial metabolic processes. Black bars indicate information based on laboratory studies, while white bars indicate activities measured in natural microbial communities. Figure from Oren, 2013 [86].

Table 1.1: Halophilic Archaea, modified from [127] to include the recently discovered *Nanohaloarchaea* [79].

Phylum	**Class**	**Genera**
Euryarchaeota	*Halobacteria*	*Halobacteria, Haladaptatus, Halalkalicoccus, Halarchaeum, Haloarcula, Halobaculum, Halobiforma, Halococcus, Haloferax, Halogeometricum, Halogranum, Halomicrobium, Halonotius, Halopelagius, Halopiger, Haloplanus, Haloquadratum, Halorhabdus, Halorubrum, Halorussus, Halosarcina, Halosimplex, Halostagnicola, Haloterrigena, Halovivax, Natrialba, Natrinema, Natronoarchaeum, Natronobacterium, Natronococcus, Natronolimnobius, Natronomonas, Natronorubrum, Salarchaeum*
	Methanomicrobia	*Methanohalobium, Methanocalculus, Methanohalohilus, Methanosalsum*
Nanohaloarchaea		*Nanosalina, Nanosalinarum*

Table 1.2: Halophilic Bacteria, modified from [127].

Phylum	**Class**	**Genera**
Actinobacteria	*Actinobacteria*	*Actinopolyspora, Amycolatopsis, Georgenia, Corynebacterium, Haloactinobacterium, Haloactinopolyspora, Haloechinothrix, Haloglycomyces, Nesterenjonia, Nocardiopsis, Haloactinospora, Streptomonospora, Isoptericola, Prauserella, Saccharomonospora, Saccharopolyspora*
Bacteroidetes	*Bacteroidia*	*Anaerophaga*
	Flavobacteria	*Gramella, Psychroflexus*
	Sphingobacteria	*Salinibacter, Salisaeta*
Cyanobacteria		*Rubidibacter, Prochlorococcus, Halospirulina*

Continued on next page

Table 1.2 – *Continued from previous page*

Phylum	**Class**	**Genera**
Firmicutes	*Bacilli*	*Alkalibacillus, Aquisalibacillus, Bacillus, Filobacillus, Gracilibacillus, Halalkalibacillus, Halolactibacillus, Halobacillus, Jeotgalibacillus, Lentibacillus, Oceanobacillus,Ornithinibacillus, Paraliobacillus, Piscibacillus, Pontibacillus, Salimicrobium, Salinibacillus, Salirhabdus, Salsuginibacillus, Sediminibacillus, Salinicoccus, Tenuibacillus, Thalassobacillus, Virgibacillus*
	Clostridia	*Acetohalobium, Halanaerobacter, Halanaerobium, Halobacteroides, Halocella, Halonatronum, Halothermothrix, Natranaerobius, Natrionella, Natronovirga, Orenia, Selenihalanaerobacter, Sporohalobacter*

Continued on next page

Table 1.2 – *Continued from previous page*

Phylum	**Class**	**Genera**
Proteobacteria	*Alphaproteobacteria*	*Antarctobacter, Citreimonas, Dichotomicrobium, Fodinicurvata, Hwanghaeicola, Hyphomonas, Jannaschia, Maribaculum, Maribius, Marispirillum, Methylarcula, Oceanibulbus, Oceanicola, Palleronia, Paracoccus, Ponticoccus, Rhodobium, Rhodotalassium, Rhodovibrio, Rhodovulum, Roseicitreum, Roseinatronobacter, Roseisalinus, Roseospira, Roseovarius, Salinihabitans, Salipiger, Sediminimonas, Shimia, Sulfitobacter, Tropicibacter, Woodsholea, Yangia*

Continued on next page

Table 1.2 – *Continued from previous page*

Phylum	**Class**	**Genera**
	Gammaproteobacteria	*Aidingimonas, Alcanivorax, Alkalilmnicola, Alteromonas, Aestuariibacter, Aquisalimonas, Arhodomonas, Carnimonas, Chromohalobacter, Cobetia, Ectothiorhodspira, Ectothiorhodosinus, Glaciecola, Gilvimarinus, Haliea, Halochromatium, Halomonas, Halorhodospira, Halospina, Halothiobacillus, Idiomarina, Kushneria, Marichromatium, Marinobacter, Marinobacterium, Melitea, Methylohalomonas, Microbulbifer, Modicisalibacter, Nitrincola, Oleispira, Pseudidiomarina, Pseudoaltermonas, Psychromonas, Pseudomonas, Saccarospirillum, Salicola, Salinicola, Salinisphaera, Salinivibrio, Thioalkalibacter, Thioalkalivibrio, Thiohalobacter, Thiohalorhabdus, Thiohalocapsa, Thiohalomonas, Thiohalophilus, Thiohalospira, Thiomicrospira*

Continued on next page

Table 1.2 – *Continued from previous page*

Phylum	**Class**	**Genera**
	Deltaproteobacteria	*Desulfocella, Desulfohalobium, Desulfonatronospira, Desulfosalsimonas, Desulfovermiculus, Desulfovibrio, Desulfurivibrio*
	Epsilonproteobacteria	*Arcobacter, Sulfurimonas, Sulfurovum*
Spirochaetes	*Spirochaetes*	*Spirochaeta*
Tenericutes	*Mollicutes*	*Haloplasma*
Thermotogae	*Thermotogae*	*Petrotoga*

1.3 Lake Tyrrell, Australia, as a model ecosystem

Lake Tyrrell, Victoria, Australia, is located in the center of the Murray Basin Plains (Figure 1.6), in a region with a semi-arid climate, average rainfall of 300 mm/year, and an evaporation rate of 2000 mm/year [66]. The lake is considered an acid-hypersaline system, where low-pH, oxygenated, saline, metal-rich groundwater from springs is evapo-concentrated and mixed with near-neutral pH waters, rich in sulfides [64]. The lake shows seasonal salinity variations. During winter, the salt content is approximately >250 g/L; in summer, due to water evaporation, the residual brines reach concentrations >330 g/L [66].

Lake Tyrrell has been described and studied in detail in terms of its hydrological and geochemical features [64, 66, 50]], making it a great candidate for microbiological characterization. Recent projects have used a diverse array of microbiological techniques to study the microbial diversity of Lake Tyrrell, including Eukaryotes [55], Archaea and Bacteria [96, 79, 124], and Viruses [33, 34]. Future work will combine the metagenomic, proteomic, and available geochemical

information to provide a more integrative description of the interactions between microbes, viruses, and the environment.

In this thesis, we explore the microbial diversity of the Lake Tyrrell ecosystem, based on the data generated through a metagenomic study. In particular, our study highlights how, by assembling metagenomic data, we obtain a more complete picture of the microbial diversity present in the community, and also how we can exploit this information to obtain a broad picture of the phylogenetic and functional diversity, and to explore in detail the genetic diversity of the members of the microbial community.

Chapter 2 describes how the assembly of metagenomic datasets allowed the recovery and identification of novel microbial groups that are abundant in the hypersaline waters of Lake Tyrrell, and other hypersaline ecosystems. Using additional metadata, such as size fractionation, sequence nucleotide composition, and phylogenetic binning, two near-complete genomes from a novel Class of Archaea, the *Nanohaloarchaea*, were recovered from the metagenomic samples. This chapter represents work that has already been published [79]. My role, in this study was in the classification of sequences into phylogenetic groups using statistical approaches, such as non-metric multidimensional scaling, and exploring the functional diversity of the two novel genomes.

Chapter 3 corresponds to the bioinformatic analysis and description of a novel type of microbial rhodopsin, xenorhodopsin, identified from the genomes of the Nanohaloarchaea.The results convey how this rhodopsin appears to be a new class of microbial rhodopsin based on phylogenetic analysis and on the presence of unique amino acid signatures. This work has already been published [125], where I was the lead author of the study.

Chapter 4 describes the assembly of several genomes from the Lake Tyrrell metagenome, based on the combination of assembly-based approaches and metadata. These results provide a framework for future analyses of this ecosystem, providing a set of habitat-specific genomes, including phylogenetic, functional, and genetic diversity. This work is already published [96]. My role in this study was developing the methods for classification of the assembled scaffolds into different

phylogenetic groups and developing novel visualization and analysis approaches to compare the functional repertoire of community members.

Chapter 5, leverages the assembled habitat-specific genomes, and describes a bioinformatic approach for the analysis of genetic diversity using metagenomic approaches. Using this framework, a deep-sequencing approach was used to characterize the population composition and genetic heterogeneity of two different temporal samples (Summer and Winter) from the microbial members of the Lake Tyrrell community.

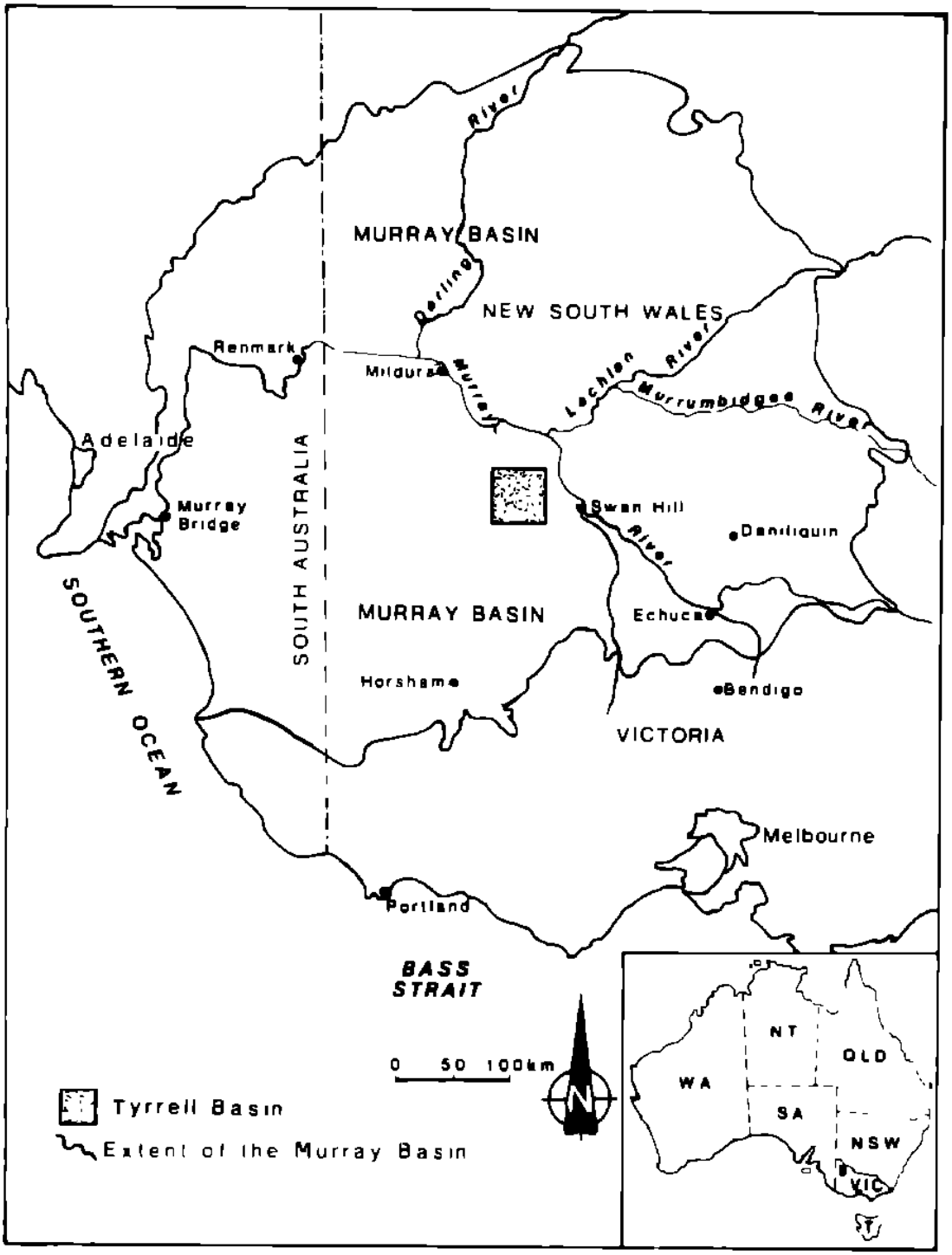


Figure 1.6: Location of Lake Tyrrell in the southeastern region of Australia. Figure from Macumber, 1992. [66].

Chapter 2

De novo Metagenomic Assembly Reveals Abundant Novel Major Lineage of Archaea in Hypersaline Communities

The ISME Journal (2011), 1–13

www.nature.com/ismej

ORIGINAL ARTICLE

De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities

Priya Narasingarao[1,8], Sheila Podell[1,8], Juan A Ugalde[1], Céline Brochier-Armanet[2], Joanne B Emerson[3], Jochen J Brocks[4], Karla B Heidelberg[5], Jillian F Banfield[3,6] and Eric E Allen[1,7]

[1]*Marine Biology Research Division, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA;* [2]*Université de Provence, Aix-Marseille Université, CNRS, UPR 9043, Laboratoire de Chimie Bactérienne, Institut de Microbiologie de la Méditerranée (IFR88), Marseille, France;* [3]*Department of Earth and Planetary Sciences, University of California, Berkeley, Berkeley, CA, USA;* [4]*Research School of Earth Sciences, The Australian National University, Canberra, ACT, Australia;* [5]*Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA;* [6]*Department of Environmental Science, Policy, and Management, University of California, Berkeley, Berkeley, CA, USA and* [7]*Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA*

This study describes reconstruction of two highly unusual archaeal genomes by *de novo* metagenomic assembly of multiple, deeply sequenced libraries from surface waters of Lake Tyrrell (LT), a hypersaline lake in NW Victoria, Australia. Lineage-specific probes were designed using the assembled genomes to visualize these novel archaea, which were highly abundant in the 0.1–0.8 μm size fraction of lake water samples. Gene content and inferred metabolic capabilities were highly dissimilar to all previously identified hypersaline microbial species. Distinctive characteristics included unique amino acid composition, absence of Gvp gas vesicle proteins, atypical archaeal metabolic pathways and unusually small cell size (approximately 0.6 μm diameter). Multi-locus phylogenetic analyses demonstrated that these organisms belong to a new major euryarchaeal lineage, distantly related to halophilic archaea of class Halobacteria. Consistent with these findings, we propose creation of a new archaeal class, provisionally named 'Nanohaloarchaea'. In addition to their high abundance in LT surface waters, we report the prevalence of Nanohaloarchaea in other hypersaline environments worldwide. The simultaneous discovery and genome sequencing of a novel yet ubiquitous lineage of uncultivated microorganisms demonstrates that even historically well-characterized environments can reveal unexpected diversity when analyzed by metagenomics, and advances our understanding of the ecology of hypersaline environments and the evolutionary history of the archaea.

The ISME Journal advance online publication, 30 June 2011; doi:10.1038/ismej.2011.78
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: assembly; halophile; hypersaline; metagenome; Nanohaloarchaea

Introduction

Cultivation-independent molecular ecology techniques currently used to survey environmental microbiota include analysis of phylogenetic marker genes, targeted functional gene inventories and direct sequencing of DNA recovered from environmental samples (reviewed in Hugenholtz and Tyson, 2008; Wooley *et al.*, 2010). Direct metagenomic sequencing is an appealing route for investigating microbial community composition because it provides simultaneous insight into phylogenetic composition and metabolic capabilities of uncultivated populations (Allen and Banfield, 2005; Wilmes *et al.*, 2009). Gene fragments from individual sequencing reads and small assembled contigs can be annotated and assigned to approximate phylogenetic bins based on comparison with databases of known reference genomes (Mavromatis *et al.*, 2007). However, cultivation biases limit the phylogenetic and physiological breadth of available reference genomes (Wu *et al.*, 2009). Single cell genomics can potentially broaden genomic databases, but often provides highly fragmented data because of amplification biases (Lasken, 2007; Woyke *et al.*, 2009). As a result

Correspondence: EE Allen, Marine Biology Research Division, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093-0202, USA.
E-mail: eallen@ucsd.edu
[8]These authors contributed equally to this work.
Received 13 January 2011; revised 10 May 2011; accepted 14 May 2011

of skewed genomic representations in reference data sets, metagenome analysis methods that rely on previously described sequence examples (for example, fragment recruitment approaches) share an inherent potential bias against novel findings. This anti-novelty bias can be overcome by *de novo* sequence assembly, which does not rely on external reference sequences, and can facilitate resolution of phylogeny-to-function linkages for individual community members. Yet *de novo* sequence assembly techniques are rarely applied to metagenomic sequences because of sampling deficiencies and/or computational challenges (Allen and Banfield, 2005; Baker *et al.*, 2010).

Habitats characterized by low diversity microbial communities have proven useful for validating molecular (eco-)systems biology approaches to examine the genetic and functional organization of native microbial consortia (Tyson *et al.*, 2004; Allen and Banfield, 2005; Ram *et al.*, 2005; Lo *et al.*, 2007; Raes and Bork, 2008; Wilmes *et al.*, 2009). High salt-impacted habitats are distributed globally in the form of hypersaline lakes, salt ponds and solar (marine) salterns, where evaporative processes result in salt concentrations close to and exceeding saturation. These environments contain microbial communities of intermediate complexity (Oren, 2008), providing excellent model systems for developing scalable analytical techniques applicable to environments with greater species richness and evenness.

The biochemical and physiological challenges faced by extremely halophilic organisms have resulted in unique adaptations to maintain osmotic balance, overcome reduced water activity because of the hygroscopic effects of saturating salt concentrations, and deter DNA damage induced by intense solar irradiation (Bolhuis *et al.*, 2006; Hallsworth *et al.*, 2007). The most extreme halophiles maintain osmotic balance using a 'high salt-in' strategy, which allows intracellular salt concentrations to reach levels approximately isosmotic with the external environment (Oren, 2008). Microorganisms using the salt-in strategy not only endure extreme ionic strength, they require it for growth. Although salt-in adaptation can be energetically more favorable than transporting salt out and the accumulation of compatible solutes (Oren, 1999), it requires significant modifications to the intracellular machinery, including specialized protein amino acid compositions to maintain solubility, structural flexibility, and water availability necessary for enzyme function (Fukuchi *et al.*, 2003; Bolhuis *et al.*, 2008; Paul *et al.*, 2008; Rhodes *et al.*, 2010).

The study of microbial populations in extreme hypersaline environments is well established; the first cultivated halophilic microorganism appeared in Bergey's manual over a century ago (Oren, 2002a). Despite the extreme conditions in salt-saturated habitats, microbial cell densities often exceed 10^7 cells ml^{-1} (Oren, 2002b). Although salt-adapted organisms derive from all three domains of life, most extreme hypersaline habitats are dominated by halophilic archaea belonging to the monophyletic class Halobacteria (phylum Euryarchaeota), including members of the genera *Haloquadratum*, *Halobacterium*, *Halorubrum* and *Haloarcula* (Oren, 2008). Pure isolates of halophilic archaea currently include >96 species distributed among 27 genera, with genome sequence information available for more than a dozen species (Oren *et al.*, 2009). Numerous cultivation-independent biodiversity surveys have been performed in hypersaline environments using PCR amplification of archaeal and bacterial 16S ribosomal RNA (rRNA) genes, as well as direct metagenomic sequencing of community DNA (Grant *et al.*, 1999; Benlloch *et al.*, 2001; Ochsenreiter *et al.*, 2002; Burns *et al.*, 2004; Demergasso *et al.*, 2004; Jiang *et al.*, 2006; Maturrano *et al.*, 2006; Mutlu *et al.*, 2008; Pagaling *et al.*, 2009; Sabet *et al.*, 2009; Oh *et al.*, 2010; Rodriguez-Brito *et al.*, 2010). These studies confirm high abundance of a few dominant species with widespread geographical distribution, but the intermittent recovery of atypical, unconfirmed sequence fragments hints at additional, unrecognized diversity among halophilic archaea (Grant *et al.*, 1999; Lopez-Garcia *et al.*, 2001; Pagaling *et al.*, 2009; Oh *et al.*, 2010; Sime-Ngando *et al.*, 2010).

The lure of uncovering biological novelty is a major incentive driving metagenomic investigations in many habitats worldwide. This study demonstrates that even historically well-characterized habitats like extreme hypersaline lakes and solar salterns can reveal unexpected genes, metabolic features and entire lineages overlooked previously. The 'assembly-driven' community metagenomic approach applied in the current study has led to the discovery and reconstruction of near-complete genomes for two new archaeal genera representing the first members of a previously undescribed taxonomic class of halophilic archaea. We demonstrate that members of this new archaeal class are present in high abundance and broadly distributed in other hypersaline habitats worldwide.

Materials and methods

Sample collection

Surface water samples (0.3 m depth) were collected from Lake Tyrrell (LT), Victoria, Australia and a high salinity crystallizer pond at South Bay Salt Works, Chula Vista (CV) California. Detailed locations, sampling dates, and physical characteristics of the collection sites are provided in Supplementary Figure S1.

Water samples of 20 l each were passed through a 20 μm Nytex prefilter, followed by sequential filtration through a series of polyethersulfone, 142 mm diameter membrane filters (Pall Corporation, Port Washington, NY, USA) of decreasing porosities (3 μm > 0.8 μm > 0.1 μm) using a peristaltic pump. After each stage of filtration, filters were frozen for

future DNA extraction, 16S rRNA gene analysis and metagenomic sequencing. Aliquots of filtered water were fixed with formaldehyde (7% final concentration) overnight at 4 °C. Fixed water samples were collected on 0.2 μm polycarbonate GTTP filters (Millipore, Billerica, MA, USA) for fluorescence *in situ* hybridization (FISH) and direct count microscopy.

Library construction and assembly

Genomic DNA was extracted from individual, bar-coded 0.8 and 0.1 μm filters. Filter-specific DNA libraries were constructed with insert sizes of 8–10 kbp and/or 40 kb (fosmids) at the J Craig Venter Institute, as described previously (Goldberg *et al.*, 2006). Details of genomic DNA sequence libraries are provided in Supplementary Table S1.

16S rRNA gene clone libraries were constructed by amplification of LT metagenomic DNA using universal archaeal primer sequences Arc21F and Arc529R (Table 1), as previously described (Bik *et al.*, 2010). A group-specific primer for Nanohaloarchaea (LT_1215R) was designed using the NCBI primer design tool, and used together with universal archaeal primer Arc21F to amplify both LT and CV community DNA. Amplification products were cloned using the TOPO TA cloning kit (Invitrogen, Carlsbad, CA, USA) and sequenced bi-directionally with M13F and M13R primers.

Sanger and pyrosequencing read libraries were assembled both individually and in various combinations, using Celera Assembler software version 5.4 (Myers *et al.*, 2000), in a series of iterative assemblies guided by phylogenetic binning. Detailed genome assembly procedures are provided in Supplementary Information.

Genome annotation

J07AB43 and J07AB56 draft genomes were annotated using the Integrated Microbial Genome Expert Review service of the Joint Genome Institute (Markowitz *et al.*, 2009b). Genome completeness was estimated for the J07AB56 and J07AB43 scaffold groups by comparing genes involved in transcription, translation and replication to those identified as highly conserved in previously sequenced archaeal genomes (Ciccarelli *et al.*, 2006; Wu and Eisen, 2008; Puigbo *et al.*, 2009). Orthologs shared between the J07AB43 and J07AB56 proteomes were detected using the reciprocal smallest distance algorithm (threshold *e*-value = 1e-05; sequence divergence = 0.4) (Wall and Deluca, 2007).

Amino acid composition analysis

Amino acid frequencies in predicted proteins from J07AB56, J07AB43 and 1455 archaeal and bacterial genomes were compared using the Primer 6 software program (Clarke and Gorley, 2006) to perform Non-Metric Multidimensional Scaling (NM-MDS) analysis (Ramette, 2007). For each genome, the frequency of each amino acid for all predicted proteins was calculated using a custom perl script. These values were standardized by *Z*-score, then used to calculate a Euclidean distance similarity matrix. NM-MDS analysis was performed using default program parameters (25 random starts, Krustal fit scheme of 1 and a minimum stress value of 0.01). In addition to NM-MDS analysis, a cluster analysis was performed to define groups within the NM-MDS plot using a multidimensional distance parameter of 4%.

Table 1 Primers and probes for detecting 16S rRNA sequences

Use	*Target*	*Name*	*Sequence (5′ to 3′)*	*Reference*
PCR	NHA	LT_1215R	*ggccgcgtgtatcccagagc*	This study
	A	Arc21F	*ttc***C***ggttgatccygcc***C***ga*	DeLong (1992)
	A	Arc529R	*accgcggckgctggc*	DasSarma and Fleischman (1995)
PCR[a]	A	ArcF1	*attc***C***ggttgatcctgc*	Ihara *et al.* (1997)
	A	Arc27Fa	*tcyggttgatcctgsc***G***g*	Raes and Bork (2008)
	U	Univ515F	*tgccagc*A*gccgcggtaa*	Lane (1991)
	A	Arc751F	**C***c*GA*cggtg*A*g*R*grygaa*	Baker *et al.* (2003)
	A	Arc958R	*y*C*c*G*gcgtt*GA*mtc*C*aatt*	DeLong (1992)
	U	Univ1390R	*ac*G*ggc***G***gtgtgtrcaa*	Brunk and Eis (1998)
	A	UA1406R	*ac*G*ggc***G***gtgwgtrcaa*	Baker *et al.* (2003)
	A	Arc1492R	A*C*GG*h*TACC*ttg*T*ta*C*gactt*	Grant *et al.* (1999)
	U	Univ1492R	GGT*T*ACC*ttg*T*ta*C*gactt*	Lane (1991)
FISH	A	Arc915	*gtgctcccccgccaattcct*	Amann *et al.* (1995)
	NHA	Narc_1214	*ccgcgtgtatcccagagc*	This study
	NHA	LT_ 1198h1	*attcgggccatactgacct*	This study
	NHA	LT_ 976-h2	*ggctctggtagrgtrc*	This study
	NHA	LT_ 1237h3	*tytstttgthccggccattg*	This study
	B	Eub338	*gctgcctcccgtaggagt*	Amann *et al.* (1990)
	B	Eub338plus	*gcwgccacccgtaggtgt*	Daims *et al.* (1999)

Target specificity abbreviations: A, archaea; B, bacteria; NHA, nanohaloarchaea; U, universal. [a]PCR primer mismatches are capitalized. Bold indicates primer mismatches to J07AB43 only, underline to J07AB56 only, and boxed to both J07AB43 and J07AB56.
Lower case italic letters indicate exact matches to the organisms described in the text. Upper case non-italic letters indicate mismatches of three different types (bold, underline, or boxed).
[a]These primers were not used in this study; sequences are shown for comparison only.

Phylogenetic analysis
16S rRNA sequences and ribosomal proteins from euryarchaeal genomes in the JGI-IMG database (Markowitz *et al.*, 2009a) and GenBank were compared with metagenomic gene sequences obtained by (i) extraction from assembled scaffolds and (ii) amplification and sequencing of 16S rRNA genes from LT and CV clone libraries. Maximum likelihood trees were constructed using TreeFinder v.10.08 (Jobb *et al.*, 2004) and PhyML v.3.0 (Guindon and Gascuel, 2003). The robustness of each maximum likelihood tree was estimated using non-parametric bootstrap analysis. Details of alignment curation and tree construction are provided in Supplementary Information.

Predicted proteins in assembled genomes were evaluated for phylogenetic relatedness to known sequences in NCBI GenBank nr using the DarkHorse program, version 1.3, with a threshold filter setting of 0.05 (Podell and Gaasterland, 2007; Podell *et al.*, 2008). Minimum quality criteria for match inclusion in the DarkHorse analysis were that BLASTP alignments to GenBank nr sequences cover at least 70% of total query length and have *e*-value scores of 1e-5 or better.

Fluorescence in situ *hybridization*
Fluorophore-conjugated custom 16S rRNA probes (Table 1) were designed using ARB (Ludwig *et al.*, 2004), screened for specificity *in silico* using ProbeCheck (Loy *et al.*, 2008) and synthesized by Integrated DNA Technologies Inc. (San Diego, CA, USA). FISH was performed on CV and LT water samples collected on 0.2 μm polycarbonate GTTP filters (Millipore) at every stage of filtration (post 20 μm, post 3 μm and post 0.8 μm). The Nanohaloarchaea-specific probe Narc_1214 conjugated with Cy3 along with unlabeled helper probes LT_1198h1, LT_976h2 and LT_127h3 (Fuchs *et al.*, 2000) were used for FISH analysis. Universal probes Arc915 (archaeal) and EubMix (a bacterial probe consisting of an equimolar mixture of Eub338 and Eub338plus) were also used for the purpose of cell counts. Hybridization conditions were optimized at 46 °C for 2 h, as previously described (Pernthaler *et al.*, 2001). Filters were mounted with Vectashield medium (Vector Laboratories, Burlingame, CA, USA), and imaged at 1000 × with a Nikon Eclipse TE-2000U inverted microscope (Nikon Instruments Inc., Irvine, CA, USA). Cell counts were performed on multiple fields per slide, normalizing 16S rRNA-specific probe counts to total number of cells stained with the DNA-binding dye 4′,6-diamidino-2-phenylindole.

Accession numbers
16S rRNA gene sequences have been deposited to DDBJ/EMBL/GenBank under accession numbers HQ197754 to HQ197794. Assembled genomes with annotations have been deposited as Whole Genome Shotgun projects under accession numbers AEIY01000000 (J07AB43) and AEIX01000000 (J07AB56).

Results

Metagenomic assembly
Seven independent DNA sequencing libraries were constructed from size-fractionated surface water samples collected at LT, Australia (Supplementary Figure S1 and Supplementary Table S1). Initial assembly of the combined 632 903 Sanger sequencing reads produced 15 008 scaffolds (maximum length = 2 764 168 bp; scaffold N50 = 29 346 bp). These scaffolds included at least six different relatively abundant microbial populations, each with a distinct nucleotide percent G + C composition. A length-weighted histogram of percent G + C versus total assembled scaffold nucleotides showed peaks corresponding to these populations (Figure 1). The largest peak in this histogram, at 48% G + C, included scaffolds containing 16S rRNA sequences from multiple strains of *Haloquadratum walsbyi*, consistent with previous observations noting the dominance of this species in similar hypersaline environments (Cuadros-Orellana *et al.*, 2007; Oh *et al.*, 2010). Three additional peaks at 60% G + C or higher included scaffolds containing 16S rRNA genes with 89–99% identity to clone sequences annotated as uncharacterized halophilic archaea (class Halobacteria). Microbial populations associated with these peaks are currently under investigation, but fall outside the scope of the present report.

Two groups of scaffolds, with peaks at 43% and 56% G + C, shared an intriguing pattern of unusual characteristics. In addition to distinctive G + C content, > 90% of the reads that co-assembled in

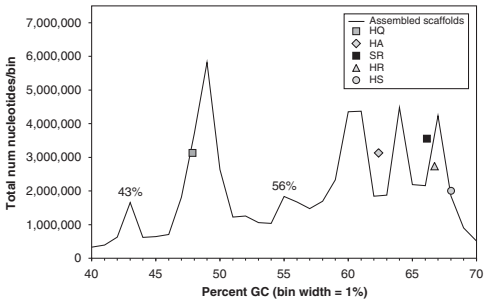


Figure 1 Length-weighted histogram of percent G + C for all scaffolds assembled from the LT community, binned in 1% GC increments. Symbols represent reference control points, indicating where five previously sequenced halophile genomes would have fallen, if they had been present in this data set. Data points are plotted based on total number of nucleotides in each scaffold (*y* axis) versus average percent GC for the entire scaffold (*x* axis). HA, *Haloarcula marismortui*; HQ, *Haloquadratum walsbyi*; HR, *Halorubrum lacusprofundi*; HS, *Halobacterium salinarum* R1; SR, *Salinibacter ruber*. Peaks labeled at 43% and 56% GC are the focus of this study.

Table 2 General features of the J07AB43 and J07AB56 draft genomes

	J07AB43	*J07AB56*
Genome size, bp	1 227 157	1 215 802
G+C percentage	44%	56%
Scaffold number	7	3
rRNA operons	1	1
tRNAs	59	38
Predicted CDSs	1678	1411
CDSs w/func. Pred.	773	719

Abbreviations: CDS, coding sequence; rRNA, ribosomal RNA; tRNA, transfer RNA.

these scaffolds were obtained from microorganisms that had passed through a 0.8 μm filter, but were retained on a 0.1 μm filter, suggesting small cell size. The 16S rRNA gene sequences contained in these scaffolds were <78% identical to any previously known cultured isolate, although they did resemble 16S rRNA gene fragments periodically recovered in culture-independent surveys of microbial diversity in hypersaline waters (Grant *et al.*, 1999; Oh *et al.*, 2010; Sime-Ngando *et al.*, 2010).

To optimize assembly efficiency for these unusual populations, the full set of metagenomic reads were subjected to a series of iterative assemblies guided by phylogenetic binning. The 43% G+C peak was thereby consolidated into seven major scaffolds (J07AB43) and the 56% G+C peak into three major scaffolds (J07AB56) (Supplementary Table S2). The J07AB43 and J07AB56 scaffold groups were subsequently analyzed as draft genomes, each representing the consensus sequence of an individual microbial population. Overall properties of these draft genomes are summarized in Table 2. These properties differ substantially from previously sequenced extreme halophiles in both nucleotide composition, expressed as percent G+C, and total genome size (Markowitz *et al.*, 2009a). With the exception of *H. walsbyi*, at 48% G+C, all other previously described halophilic archaea, as well as the halophilic bacterium *Salinibacter ruber*, have nucleotide compositions of 60% or greater G+C, compared with 43% and 56% for these new organisms. Estimated total genome size and predicted number of coding sequences for J07AB43 and J07AB56 (Table 2) were also considerably smaller than other known extreme halophiles, which currently range from 2.7 to 5.4 Mbp.

Genome completeness

To estimate the extent of genome completeness of J07AB43 and J07AB56, functional annotations for all predicted proteins were searched against a set of 53 housekeeping genes, previously identified as universally present in all archaeal genomes sequenced as of 2009 (Puigbo *et al.*, 2009). These highly conserved genes are physically dispersed throughout the genome (non-clustered) and include ribosomal proteins, amino acid tRNA synthetases, translation initiation and elongation factors, molecular chaperones and proteins essential for DNA replication and repair. All 53 of the universal archaeal housekeeping proteins were identified in J07AB56 while 44/53 (83%) were found in the J07AB43 draft genome (Supplementary Table S3). The presence of these core proteins, a single rRNA operon and transfer RNAs enabling translation of all 20 amino acids, suggests that both draft genomes are nearly complete.

Community abundance

Community abundance of J07AB43 and J07AB56 was initially assessed by sequencing 16S rRNA gene clone libraries, constructed by amplifying LT community DNA with universal archaeal primers Arc21F and Arc529R (Table 1). Amplified sequences with >91% identity to the J07AB43 and J07AB56 draft genomes were found in 124/315 (39%) of archaeal clones obtained from organisms retained on 0.1 μm pore filters, but only 24/254 (9%) of clones retained on 0.8 μm pore filters. These results are consistent with the observed enrichment of J07AB43 and J07AB56 reads specifically derived from 0.1 μm filter fractions in the assembled data set.

As a second, independent test of community abundance, new lineage-specific 16S rRNA probes were designed to visualize J07AB43 and J07AB56 cells in environmental samples by FISH (Table 1). These probes were used in combination with the DNA-binding dye 4′,6-diamidino-2-phenylindole and universal bacterial and archaeal probes to obtain direct cell counts in LT and CV water samples (Figure 2). Cells approximately 0.6 μm in diameter were labeled with lineage-specific probe NArc_1214 in samples from both locations. These results are consistent with size estimates of <0.8 μm but >0.1 μm based on filter-specific composition for both amplified 16S rRNA clones and metagenomic reads. Direct counts of fluorescently labeled cells indicated that the combined abundance of strains matching the new, lineage-specific probes was approximately 14% of all 4′,6-diamidino-2-phenylindole-labeled cells in water samples from LT, and 8–11% in samples from CV (Supplementary Table S4).

Community abundance of the organisms responsible for the J07AB43 and J07AB56 draft genomes was further examined using statistical properties of the assembled metagenomic sequence data. The number of reads that co-assembled to create each composite population scaffold group was divided by the total number of reads available and normalized for estimated genome size. Assuming the two new genomes are approximately 1.2 Mbp each, and other microbial species sampled from LT have an average genome size of 3 Mbp, J07AB43 was estimated to represent at least 6.7% of the LT sampled community (17 066 reads) and J07AB56 at least 3.4% (8652 reads), totaling approximately 10% for the two

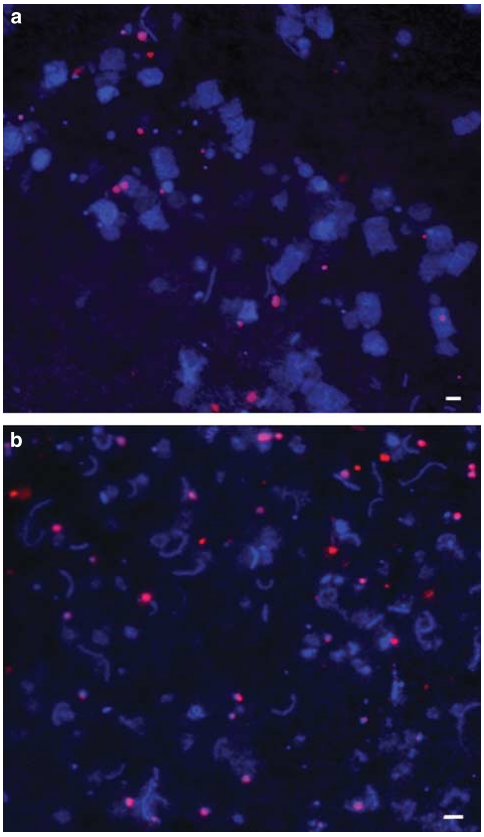


Figure 2 FISH micrographs. (**a**) LT (0.1 to 3 μm filter fraction), (**b**) CV South Bay Salt Works (0.1 to 0.8 μm filter fraction). All cells are stained with 4′,6-diamidino-2-phenylindole (blue). Nanohaloarchaea cells shown are stained with lineage-specific Cy3 probe Narc_1214 (red). Scale bar = 2 μm.

populations combined (3.0/1.2*25 718/632 903). Calculations based on metagenomic assembly most likely underestimate true population abundance, because they may exclude closely related polymorphic strains containing DNA sequence variations that were not incorporated into the consensus population assembly.

Taxonomic position of J07AB43 and J07AB56

J07AB43 and J07AB56 16S rRNA shared sequence identities of 68% to 75% with previously sequenced, cultured representatives of class Halobacteria (Supplementary Table S5). An unrooted maximum likelihood phylogenetic tree of euryarchaeotal 16S rRNA gene sequences placed J07AB43 and J07AB56 as a deep sister group of class Halobacteria (Figure 3), with significant bootstrap support.

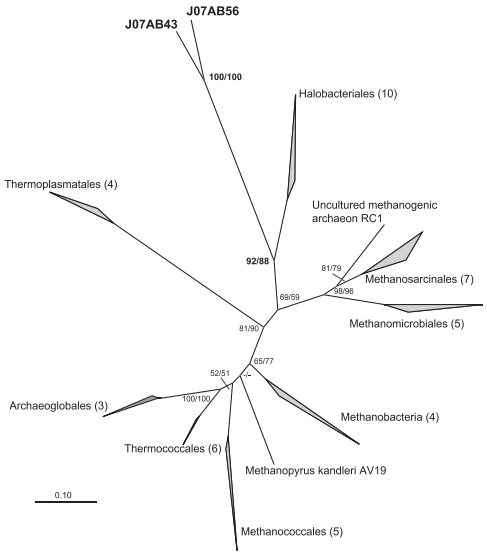


Figure 3 Unrooted maximum-likelihood 16S rRNA gene phylogenetic tree of the Euryarchaeota. Tree is based on 48 sequences, 1275 positions. Numbers of sequences in each collapsed node are indicated in parentheses. Numbers at nodes represent bootstrap values inferred by TreeFinder/PhyML. Bootstrap values <50% are indicated by a '–' sign. Scale bar represents 0.1 substitutions per site. A full, uncollapsed version of this tree is presented in Supplementary Figure S2a.

Concatenated ribosomal protein data sets have been shown to be particularly useful for resolving deep evolutionary relationships (Brochier *et al.*, 2002; Matte-Tailliez *et al.*, 2002; Rokas *et al.*, 2003; Rannala and Yang, 2008). Phylogenetic analysis of 57 ribosomal proteins from the J07AB43 and J07AB56 draft genomes showed, like the 16S rRNA tree, robust placement of these genomes as a deeply branching sister group of class Halobacteria, with bootstrap values of 98% (PhyML) and 74% (TreeFinder). This relationship was corroborated using Dayhoff04 recoding of ribosomal protein alignments (Hrdy *et al.*, 2004; Susko and Roger, 2007), to rule out possible artifacts of biased amino acid composition or fast-evolving lineages (Supplementary Figure S2b). The long branch lengths separating J07AB43 and J07AB56 from members of class Halobacteria indicate that these two sister-lineages are only distantly related, consistent with the average divergence of 35% observed between Halobacteria and J07AB43 and J07AB56 16S rRNA gene sequences (Supplementary Table S5). By contrast, 16S rRNA variability within the Halobacteria is <16%.

Nearly 60% of predicted proteins in J07AB43 and J07AB56 had no GenBank database matches close enough to enable confident phylogenetic assignment. Of those that could be assigned, fewer than 20% matched proteins from members of class

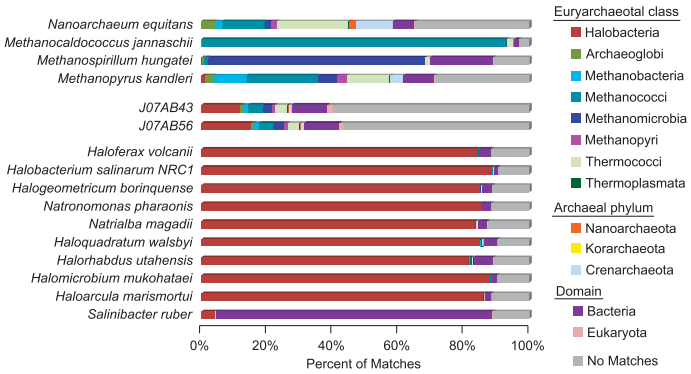


Figure 4 Phylogenetic distribution of non-self-protein BLAST matches for euryarchaeotal genomes. Searches against the GenBank nr database were classified by euryarchaeotal class, archaeal phylum, domain or no match using the DarkHorse algorithm, as described in Materials and methods section.

Halobacteria (Figure 4). In contrast, >80% of predicted proteins in the genomes of previously sequenced Halobacteria had closest non-self matches to other members of their own class, leaving fewer than 20% unmatched. Similar patterns of protein sequence conservation were observed in organisms with many sequenced database relatives, including *Methanocaldococcus janaschii*, *Methanospirillum hungatei* and *Salinibacter ruber*, but not in sparsely sampled species that are only distantly related to other known lineages, such as *Nanoarchaeum equitans* and *Methanopyrus kandleri* (Branciamore *et al.*, 2008).

Genome characteristics of J07AB43 and J07AB56

Although the J07AB43 and J07AB56 genomes are more closely related to each other than to any previously sequenced organisms, gene content analysis identified only 480 (30%) shared protein ortholog pairs between them. Of these, 143 (approximately 10% of each genome) were not found in other halophilic archaea. The majority of these shared lineage-specific sequences were too dissimilar to previously characterized proteins to assign a functional annotation. The remainder was dominated by housekeeping proteins involved in translation and ribosomal structure. Each genome included only one rhodopsin-like gene, compared with multiple paralogs present in the genomes of other extreme halophilic archaea (Ihara *et al.*, 1999), and the extremely halophilic bacterium *Salinibacter ruber* (Mongodin *et al.*, 2005). Notably absent from both genomes were homologs to the highly conserved Gvp family of gas vesicle proteins found in most halophilic archaea, Cyanobacteria and purple photosynthetic bacteria (Walsby, 1994).

Both J07AB43 and J07AB56 have highly unusual amino acid compositions compared with previously sequenced archaeal and bacterial genomes. These unusual compositions appear to support a 'salt-in' strategy of maintaining osmotic balance, as evidenced by the over-representation of amino acids with negatively charged side chains (aspartic and glutamic acid) and the under-representation of residues with bulky hydrophobic side chains (tryptophan, phenylalanine and isoleucine), to enhance protein structural flexibility and solubility under intracellular conditions of high ionic strength and low water availability. Although a similar salt-in strategy is employed by other extreme halophiles, J07AB43 and J07AB56 use their own, distinct combination of amino acids to achieve this end, preferring glutamic to aspartic acid, serine to threonine, and reduced frequencies of alanine, proline and histidine (Supplementary Table S6). The large number of proteins annotated with 'hypothetical' functions in the J07AB43 and J07AB56 genomes may be at least partially because of their unusual amino acid compositions, which can hinder recognition of database homologs in sequence-based similarity searches.

The peculiar amino acid compositions of J07AB43 and J07AB56 compared with other halophilic archaea are highlighted in a NM-MDS plot of intergenomic distances based on frequencies for all 20 standard amino acids (Figure 5). The data used to construct this matrix included all protein sequences from euryarchaeal genomes used to build the phylogenetic tree in Figure 3, supplemented with four bacterial species found in high salt environments: *Salinibacter ruber* (Bacteroidetes), *Halorhodospira halophila* (Gammaproteobacteria), *Chromohalobacter salexigens* (Gammaproteobacteria) and *Halothermothrix orenii* (Firmicutes).

Although genome percent G+C compositions were not explicitly included as one of the factors in this analysis, there is a trend for microorganisms with lower G+C (denoted with lower label numbers in Figure 5) to be located further to the right along the horizontal axis. This trend is consistent with the known influence of G+C composition on usage

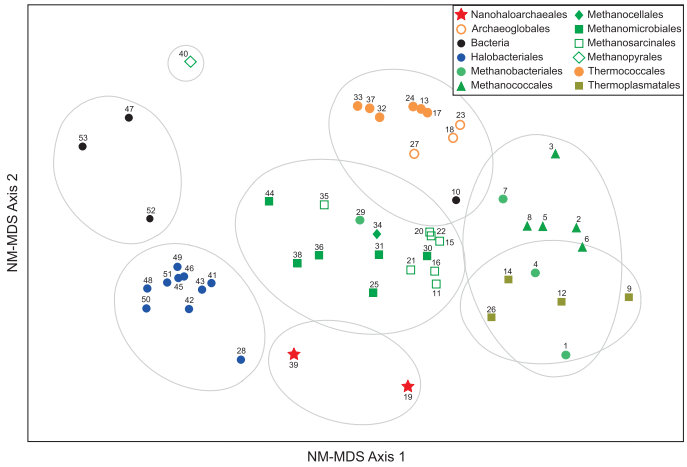


Figure 5 NM-MDS comparison of amino acid compositions. Euryarchaeal genomes were supplemented with four halophilic bacteria genomes. Symbols denote taxonomic classifications. Numbers rank genomes in increasing order of G + C content (1–10: 29–38%, 11–20: 38–43%, 21–30: 43–50%, 31–40: 50–60%, 41–53: 60–67%). Grey circles indicate hierarchical clustering, based on a 4% distance setting to define groups. A complete list of these genomes and their amino acid compositions is presented in Supplementary Table S6.

frequency for some amino acids because of codon bias (Liu *et al.*, 2010). In contrast, position along the vertical axis of Figure 5 was unrelated to percent G + C. Instead, amino acid composition differences captured along this axis appear to correlate more closely with common ancestry and/or shared environmental adaptations. The outlier positions of J07AB43 (#19) and J07AB56 (#39) along the vertical axis of Figure 5 clearly demonstrate their unusual amino acid compositions relative to other archaea. Similar outlier positions were observed for these two genomes when analyzed in the context of a much larger microbial genomic data set, including 1382 bacterial and 73 archaeal species (data not shown).

Inferred metabolic capabilities of the J07AB43 and J07AB56 genomes are consistent with a predominately aerobic, heterotrophic lifestyle. The absence of identifiable anaerobic terminal reductases suggests they are incapable of anaerobic respiration although the presence of lactate dehydrogenases suggests possible fermentative metabolism under microaerophilic conditions. Both genomes contain enzymes necessary to support glycolysis, as well as operons encoding key enzymes for glycogen synthesis and catabolism. Several of these enzymes, including a glycogen debranching enzyme and predicted alpha-1,6-glucosidase activity, are not present in any other known members of class Halobacteria. However, these enzyme activities are frequently found in archaea from classes Methanococci and Thermoplasmata that utilize starch as an internal storage molecule (König *et al.*, 1985, 1982). This suggests a possible common ancestral origin, with subsequent gene loss in the Halobacteria lineage.

In addition to the Embden–Meyerhoff pathway, genes supporting the entire pentose phosphate pathway were observed in both genomes, including both oxidative and non-oxidative branches. The presence of a complete pentose phosphate pathway has not been demonstrated previously in any other archaea, by either biochemical or bioinformatic methods (Verhees *et al.*, 2003). The key, rate-limiting enzyme for this pathway is glucose-6-phosphate dehydrogenase, which converts glucose-6-phosphate into 6-phosphoglucono-δ-lactone. Although both J07AB43 and J07AB56 appear to have complete genomic copies of this gene, the closest database relatives to their sequences are all bacterial, suggesting this functionality may have been acquired by ancient horizontal gene transfer. The nearest homolog of the glucose-6-phosphate dehydrogenases in J07AB43 and J07AB56 is from the genome of *Salinibacter ruber*, a common bacterial inhabitant of hypersaline environments believed to have experienced frequent horizontal gene exchange with archaea (Mongodin *et al.*, 2005).

Geographical distribution and diversity

Lineage-specific PCR primer, LT_1215R (Table 1) and general archaeal primer Arc21F were used to construct clone libraries from environmental DNA samples collected from both LT and CV, yielding 43 new 16S rRNA gene sequences. Additional 16S rRNA gene sequences, with >85% identity to J07AB43 and J07AB56, were identified in public databases. These published sequences originated in environmental samples from Africa, Asia and South America, as well as Australia and North America (Supplementary Table S7). The phylogenetic analysis of these 16S rRNA gene sequences reveals at least eight distinct clades with strong bootstrap support

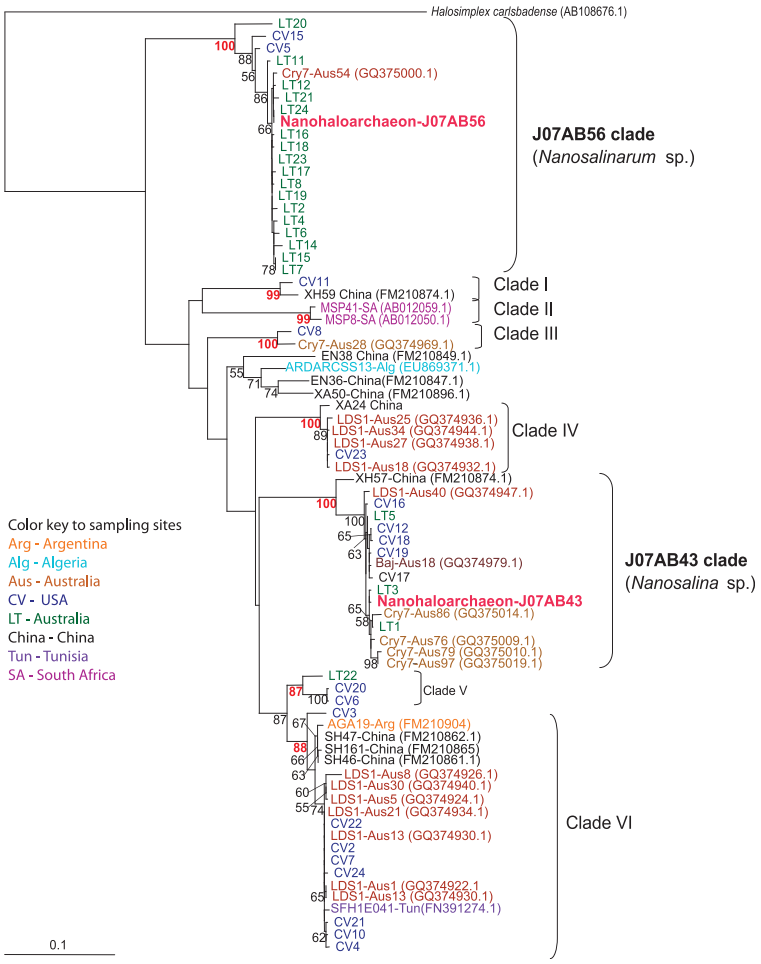


Figure 6 16S rRNA gene maximum likelihood tree of Nanohaloarchaea sequences recovered from worldwide hypersaline habitats. Tree is based on 709 nucleic acid positions in 77 sequences. Numbers at nodes represent bootstrap values (values <50% not shown). Scale bar shows average number of substitutions per site.

(bootstrap values >87%, Figure 6). Based on degree of sequence divergence, each clade most likely represents a new genus or higher taxonomic level. Classification of J07AB43 and J07AB56 into separate genera is strongly supported by tree topology, 16% sequence divergence in the 16S rRNA gene (Supplementary Table S5) and a 13% difference in genomic G + C content.

Discussion

This study has demonstrated that re-examination of a fairly simple, well studied environmental habitat using a combination of strategic environmental sampling, deep sequencing, and *de novo* metagenomic assembly can reveal significant new information. We have discovered and characterized nearly complete genomes representing a novel archaeal lineage prevalent in hypersaline systems worldwide, yet very different from all previously described members of class Halobacteria.

We propose the creation of a new class 'Nanohaloarchaea' within phylum Euryarchaeota to accommodate this new lineage. We further propose partitioning class Nanohaloarchaea to place J07AB43 and J07AB56 into distinct genera, *Candidatus* 'Nanosalina sp. J07AB43' and '*Candidatus*

Nanosalinarum sp. J07AB56'. Evidence supporting these proposals includes: (i) comprehensive euryarchaeotal phylogenetic analyses based on 16S rRNA genes and ribosomal proteins; (ii) lineage-specific features, including numerous genes without previously described close relatives; and (iii) significant intra-lineage diversity and abundance within geographically distinct hypersaline habitats worldwide. Evolutionary distinctness of J07AB43 and J07AB56 from other halophilic archaea is reinforced by taxonomic patterns of BLASTP matches for their predicted proteomes against GenBank nr, as well as amino acid composition-based clustering. The sister-grouping of class Halobacteria and class Nanohaloarchaea reflects probable derivation from an ancient common halophilic ancestor with a 'high salt-in' osmotic regulation strategy, followed by subsequent divergence along separate evolutionary paths.

Lineage-specific characteristics that distinguish '*Candidatus* Nanosalina sp.' and '*Candidatus* Nanosalinarum sp.' from most other extreme halophiles include their small physical size, compact genomes, single-copy rRNA operon, low G+C composition, unique proteome amino acid composition, absence of conserved gas vesicle genes and atypical predicted pathways associated with carbohydrate metabolism. Small compact genomes, as well as single-copy rRNA operons, have been proposed to minimize metabolic costs in habitats where neither broad metabolic repertoire nor high numbers of paralogous proteins are needed to accommodate rapid growth under fluctuating environmental conditions (Klappenbach *et al.*, 2000). Small cell size, which increases surface to volume ratio, could be an adaptation for optimizing nutrient uptake capacity. Alternatively it is possible that small physical size allows Nanohaloarchaea to remain suspended in oxygenated surface waters to support aerobic metabolism, thus eliminating the need for gas vesicles to provide buoyancy.

The low G+C compositions of the two Nanohaloarchaea genomes, especially J07AB43 (43%), are surprising considering their prevalence in high light habitats. In the absence of compensatory mechanisms, lower G+C would be expected to increase susceptibility to ultraviolet-induced DNA damage. One possible explanation is that the low G+C composition of J07AB43 is related to ecological lifestyle. Low G+C composition and genomic streamlining have been associated with decreased nitrogen requirements and a slow-growing, energy-conservative lifestyle in marine bacteria (Giovannoni *et al.*, 2005). However, the habitats from which these Nanohaloarchaea were isolated are not generally considered to be nutrient-limited (Oren, 2002b). Alternatively, it has been proposed that the low G+C composition of *H. walsbyi* (48%) compared with other halophiles is a specific adaptation to counteract the over-stabilizing effect of high magnesium concentrations on DNA structure (Bolhuis *et al.*, 2006). If extremely high environmental magnesium cannot be adequately excluded from the cell, lower genomic G+C helps maintain DNA structural flexibility and avoids difficulties in strand separation caused by elevated melting temperatures. These same principles could apply to J07AB43, providing a possible selective advantage under high magnesium conditions expected in evaporative high salt environments.

Nanohaloarchaea are estimated to represent at least 10–25% of the total archaeal community in surface water samples from LT, Australia and CV, California, USA. We believe these values are robust, based on agreement of three independent analysis techniques: amplification of environmental 16S rRNA gene sequences; statistical analysis of metagenomic sequencing reads assembled into near-complete draft genomes; and quantitative FISH of cells from natural water samples labeled with lineage-specific probes. Microscopic counts reveal that Nanohaloarchaea are present at cell concentrations exceeding 10^6 cells ml^{-1} in hypersaline habitats of Australia and North America. The sporadic identification of Nanohaloarchaea in other surveys of hypersaline communities worldwide suggests that Nanohaloarchaea represent a significant yet neglected fraction of the biomass and diversity in these habitats.

The inability of earlier studies to recognize the significant contribution of Nanohaloarchaea to hypersaline community composition is likely due to limitations of the tools routinely used to assess environmental microbial diversity, including laboratory culture, microscopy, amplification of 16S rRNA gene fragments, and sequence database similarity searches for unassembled metagenomic reads. The isolation of cultured strains from environmental habitats is known to exclude many organisms that are highly successful in their native habitats. It is therefore not surprising the 96 hypersaline archaeal isolates described to date do not include any Nanohaloarchaea. Repeated efforts to culture these microorganisms in our own laboratory have also been unsuccessful. Furthermore, cultivation-independent microbial diversity studies based on 16S rRNA gene amplification are known to suffer from primer bias (Sipos *et al.*, 2007). Mismatches between Nanohaloarchaea and many commonly used universal primers may have impeded detection in earlier studies. Primers likely to have been particularly problematic are highlighted in Table 1 (Amann *et al.*, 1990, 1995; Lane, 1991; DeLong, 1992; DasSarma and Fleischman, 1995; Ihara *et al.*, 1997; Brunk and Eis, 1998; Daims *et al.*, 1999; Grant *et al.*, 1999; Baker *et al.*, 2003; Raes and Bork, 2008). The exceptionally small size of Nanohaloarchaea compared with other halophilic microorganisms makes them difficult to visualize by microscopy in the absence of selective enrichment techniques or group-specific probes, and can prevent recovery during sample concentration procedures targeting larger microorganisms or smaller viruses (Rodriguez-

Brito *et al.*, 2010). Similar issues have been noted for other nano-sized archaea, identified solely by 16S rRNA gene sequencing (Casanueva *et al.*, 2008; Gareeb and Setani, 2009).

The presence of ultrasmall, uncultivated novel archaeal lineages in natural environments may be a common occurrence. Nanohaloarchaea represent the third nano-sized archaeal lineage to be described. However, unlike the thermophilic *Nanoarchaeum equitans* (Huber *et al.*, 2002) or the acidophilic ARMAN lineages (Baker *et al.*, 2006, 2010), members of the Nanohaloarchaea appear to be free-living based on microscopic observations. The larger genomes of Nanohaloarchaea (approximately 1.2 Mbp) relative to other symbiotic/parasitic nano-sized archaea (ARMAN, <1 Mbp; *Nanoarchaeum equitans*, <0.5 Mbp) are consistent with a possible non-host associated lifestyle for this group. It is interesting to contemplate the environmental pressures selecting for the evolution of ultrasmall microorganisms with small genomes, and to consider the extent of an ultrasmall microbial biosphere. The realization that ultrasmall populations can comprise a significant fraction of the total microbial community, yet have eluded previous detection, suggests that historical estimates of microbial biomass and numerical abundance in natural environments may be substantially underestimated. This is particularly relevant in non-extreme habitats where the existence of ultrasmall microbial populations have not yet been described or investigated.

Routine metagenomic analysis methods currently rely on the expectation that undiscovered microorganisms will have a certain degree of similarity to those already known, creating a potential bias against novel discoveries. Although this study exposes limitations of commonly used microbial diversity assessment tools in the context of detecting novel archaea in hypersaline lakes, these limitations apply even more emphatically to other more complex microbial communities, which often contain elaborate mixed consortia of Bacteria, Archaea, Eukarya and viruses. This study reinforces the utility of community genomics and *de novo* sequence assembly as important methods for the detection and analysis of biological diversity.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We thank Sue Welch and Dawn Cardace for sample collection assistance at Lake Tyrrell; Mike Dyall-Smith for generous access to reagents and laboratory equipment; Cheetham Salt Works (Lake Tyrrell, Australia) and South Bay Salt Works, Chula Vista (San Diego, CA) for permission to collect samples; Brian Collins (USFWS) for help with sample collection at South Bay Salt Works; Matt Lewis and the J Craig Venter Institute for library construction and sequencing; Nerida Wilson for assistance with phylogenetic trees; and the US Department of Energy Joint Genomes Institute for genome annotation support via the Integrated Microbial Genome Expert Review (IMG-ER) resource. We also thank Farooq Azam (SIO/UCSD) for kindly permitting use of the Nikon confocal microscope purchased with support from the Gordon and Betty Moore Foundation. Funding for this work was provided by NSF award number 0626526 (JFB, KBH, EEA) and NIH award R21HG005107-02 (EEA). JAU was supported by a Fulbright-Conicyt fellowship. CBA is supported by an Action Thématique et Incitative sur Programme of the French Centre National de la Recherche Scientifique (CNRS). Work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No DE-AC02-05CH11231.

References

Allen EE, Banfield JF. (2005). Community genomics in microbial ecology and evolution. *Nat Rev Microbiol* **3**: 489–498.

Amann RI, Binder BJ, Olson RJ, Chisholm SW, Devereux R, Stahl DA. (1990). Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl Environ Microbiol* **56**: 1919–1925.

Amann RI, Ludwig W, Schleifer KH. (1995). Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol Rev* **59**: 143–169.

Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD *et al.* (2010). Enigmatic, ultrasmall, uncultivated archaea. *Proc Natl Acad Sci USA* **107**: 8806–8811.

Baker BJ, Tyson GW, Webb RI, Flanagan J, Hugenholtz P, Allen EE *et al.* (2006). Lineages of acidophilic archaea revealed by community genomic analysis. *Science* **314**: 1933–1935.

Baker GC, Smith JJ, Cowan DA. (2003). Review and reanalysis of domain-specific 16S primers. *J Microbiol Methods* **55**: 541–555.

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

Bik EM, Long CD, Armitage GC, Loomer P, Emerson J, Mongodin EF *et al.* (2010). Bacterial diversity in the oral cavity of 10 healthy individuals. *ISME J* **4**: 962–974.

Bolhuis A, Kwan D, Thomas JR. (2008). Halophilic adaptations of proteins. In: Siddiqui KS, Thomas T (eds). *Protein Adaptation in Extremophiles*. Nova Biomedical Books: New York, pp 71–104.

Bolhuis H, Palm P, Wende A, Falb M, Rampp M, Rodriguez-Valera F *et al.* (2006). The genome of the square archaeon Haloquadratum walsbyi: life at the limits of water activity. *BMC Genomics* **7**: 169.

Branciamore S, Gallori E, Di Giulio M. (2008). The basal phylogenetic position of Nanoarchaeum equitans (Nanoarchaeota). *Front Biosci* **13**: 6886–6892.

Brochier C, Bapteste E, Moreira D, Philippe H. (2002). Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet* **18**: 1–5.

Brunk CF, Eis N. (1998). Quantitative measure of small-subunit rRNA gene sequences of the kingdom korarchaeota. *Appl Environ Microbiol* **64**: 5064–5066.

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

Casanueva A, Galada N, Baker GC, Grant WD, Heaphy S, Jones B *et al.* (2008). Nanoarchaeal 16S rRNA gene sequences are widely dispersed in hyperthermophilic and mesophilic halophilic environments. *Extremophiles* **12**: 651–656.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**: 1283–1287.

Clarke K, Gorley R. (2006). *Primer v6: User Manual/ Tutorial*. PRIMER-E: Plymouth, UK.

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

Daims H, Bruhl A, Amann R, Schleifer KH, Wagner M. (1999). The domain-specific probe EUB338 is insufficient for the detection of all Bacteria: development and evaluation of a more comprehensive probe set. *Syst Appl Microbiol* **22**: 434–444.

DasSarma S, Fleischman EM. (1995). *Archaea: A Laboratory Manual - Halophiles*, Vol 1. Cold Spring Harbor Laboratory Press: Plainview, NY.

DeLong EF. (1992). Archaea in coastal marine environments. *Proc Natl Acad Sci USA* **89**: 5685–5689.

Demergasso C, Casamayor EO, Chong G, Galleguillos P, Escudero L, Pedros-Alio C. (2004). Distribution of prokaryotic genetic diversity in athalassohaline lakes of the Atacama Desert, Northern Chile. *FEMS Microbiol Ecol* **48**: 57–69.

Fuchs BM, Glockner FO, Wulf J, Amann R. (2000). Unlabeled helper oligonucleotides increase the *in situ* accessibility to 16S rRNA of fluorescently labeled oligonucleotide probes. *Appl Environ Microbiol* **66**: 3603–3607.

Fukuchi S, Yoshimune K, Wakayama M, Moriguchi M, Nishikawa K. (2003). Unique amino acid composition of proteins in halophilic bacteria. *J Mol Biol* **327**: 347–357.

Gareeb A, Setani M. (2009). Assessment of alkaliphilic haloarchaeal diversity in Sua pan evaporator ponds in Botswana. *Afr J Biotechnol* **8**: 259–267.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D *et al.* (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**: 1242–1245.

Goldberg SM, Johnson J, Busam D, Feldblyum T, Ferriera S, Friedman R *et al.* (2006). A Sanger/ pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc Natl Acad Sci USA* **103**: 11240–11245.

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

Guindon S, Gascuel O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696–704.

Hallsworth JE, Yakimov MM, Golyshin PN, Gillion JL, D'Auria G, de Lima Alves F *et al.* (2007). Limits of life in MgCl2-containing environments: chaotropicity defines the window. *Environ Microbiol* **9**: 801–813.

Hrdy I, Hirt RP, Dolezal P, Bardonova L, Foster PG, Tachezy J *et al.* (2004). Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. *Nature* **432**: 618–622.

Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO. (2002). A new phylum of archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**: 63–67.

Hugenholtz P, Tyson GW. (2008). Microbiology: metagenomics. *Nature* **455**: 481–483.

Ihara K, Umemura T, Katagiri I, Kitajima-Ihara T, Sugiyama Y, Kimura Y *et al.* (1999). Evolution of the archaeal rhodopsins: evolution rate changes by gene duplication and functional differentiation. *J Mol Biol* **285**: 163–174.

Ihara K, Watanabe S, Tamura T. (1997). Haloarcula argentinensis sp. nov. and Haloarcula mukohataei sp. nov., two new extremely halophilic archaea collected in Argentina. *Int J Syst Bacteriol* **47**: 73–77.

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

Jobb G, von Haeseler A, Strimmer K. (2004). TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol* **4**: 18.

Klappenbach JA, Dunbar JM, Schmidt TM. (2000). rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microbiol* **66**: 1328–1333.

König H, Nusser E, Stetter KO. (1985). Glycogen in methanolobus and methanococcus. *FEMS Microbiol Lett* **28**: 265–269.

König H, Skorko R, Zillig W, Reiter W-D. (1982). Glycogen in thermoacidophilic archaebacteria of the genera Sulfolobus, Thermoproteus, Desulfurococcus, and Thermococcus. *Arch Microbiol* **132**: 297–303.

Lane DJ. (1991). 16S/23S rRNA sequencing. In: Stackebrandt E, Goodfellow M (eds). *Nucleic Acid Techniques in Bacterial Systematics*. Wiley: Chichester; New York, pp 115–175.

Lasken RS. (2007). Single-cell genomic sequencing using multiple displacement amplification. *Curr Opin Microbiol* **10**: 510–516.

Liu X, Zhang J, Ni F, Dong X, Han B, Han D *et al.* (2010). Genome wide exploration of the origin and evolution of amino acids. *BMC Evol Biol* **10**: 77.

Lo I, Denef VJ, Verberkmoes NC, Shah MB, Goltsman D, DiBartolo G *et al.* (2007). Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446**: 537–541.

Lopez-Garcia P, Moreira D, Lopez-Lopez A, Rodriguez-Valera F. (2001). A novel haloarchaeal-related lineage is widely distributed in deep oceanic regions. *Environ Microbiol* **3**: 72–78.

Loy A, Arnold R, Tischler P, Rattei T, Wagner M, Horn M. (2008). probeCheck—a central resource for evaluating oligonucleotide probe coverage and specificity. *Environ Microbiol* **10**: 2894–2898.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363–1371.

Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y *et al.* (2009a). The integrated microbial genomes system: an expanding comparative analysis resource. *Nucleic Acids Res* **38**: D382–D390.

Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC. (2009b). IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* **25**: 2271–2278.

Matte-Tailliez O, Brochier C, Forterre P, Philippe H. (2002). Archaeal phylogeny based on ribosomal proteins. *Mol Biol Evol* **19**: 631–639.

Maturrano L, Santos F, Rossello-Mora R, Anton J. (2006). Microbial diversity in Maras salterns, a hypersaline environment in the Peruvian Andes. *Appl Environ Microbiol* **72**: 3887–3895.

Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC *et al.* (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* **4**: 495–500.

Mongodin EF, Nelson KE, Daugherty S, Deboy RT, Wister J, Khouri H *et al.* (2005). The genome of Salinibacter ruber: convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proc Natl Acad Sci USA* **102**: 18147–18152.

Mutlu MB, Martinez-Garcia M, Santos F, Pena A, Guven K, Anton J. (2008). Prokaryotic diversity in Tuz Lake, a hypersaline environment in Inland Turkey. *FEMS Microbiol Ecol* **65**: 474–483.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ *et al.* (2000). A whole-genome assembly of Drosophila. *Science* **287**: 2196–2204.

Ochsenreiter T, Pfeifer F, Schleper C. (2002). Diversity of archaea in hypersaline environments characterized by molecular-phylogenetic and cultivation studies. *Extremophiles* **6**: 267–274.

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

Oren A. (1999). Bioenergetic aspects of halophilism. *Microbiol Mol Biol Rev* **63**: 334–348.

Oren A. (2002a). Diversity of halophilic microorganisms: environments, phylogeny, physiology, and applications. *J Ind Microbiol Biotechnol* **28**: 56–63.

Oren A. (2002b). *Halophilic Microorganisms and Their Environments*. Kluwer Academic: Dordrecht; Boston, xxi, 575pp.

Oren A. (2008). Microbial life at high salt concentrations: phylogenetic and metabolic diversity. *Saline Syst* **4**: 2.

Oren A, Arahal DR, Ventosa A. (2009). Emended descriptions of genera of the family Halobacteriaceae. *Int J Syst Evol Microbiol* **59**: 637–642.

Pagaling E, Wang H, Venables M, Wallace A, Grant WD, Cowan DA *et al.* (2009). Microbial biogeography of six salt lakes in Inner Mongolia, China, and a salt lake in Argentina. *Appl Environ Microbiol* **75**: 5750–5760.

Paul S, Bag SK, Das S, Harvill ET, Dutta C. (2008). Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. *Genome Biol* **9**: R70.

Pernthaler J, Glockner FO, Schonuber W. (2001). Fluorescence *in situ* Hybridization with rRNA-targeted oligonucleotide probes. In: Paul JH (ed), *Methods in Microbiology*, Vol 30. Academic Press: San Diego, pp 207–226.

Podell S, Gaasterland T. (2007). DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol* **8**: R16.

Podell S, Gaasterland T, Allen EE. (2008). A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. *BMC Bioinformat* **9**: 419.

Puigbo P, Wolf YI, Koonin EV. (2009). Search for a 'Tree of Life' in the thicket of the phylogenetic forest. *J Biol* **8**: 59.

Raes J, Bork P. (2008). Molecular eco-systems biology: towards an understanding of community function. *Nat Rev Microbiol* **6**: 693–699.

Ram RJ, Verberkmoes NC, Thelen MP, Tyson GW, Baker BJ, Blake II RC *et al.* (2005). Community proteomics of a natural microbial biofilm. *Science* **308**: 1915–1920.

Ramette A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiol Ecol* **62**: 142–160.

Rannala B, Yang Z. (2008). Phylogenetic inference using whole genomes. *Annu Rev Genomics Hum Genet* **9**: 217–231.

Rhodes ME, Fitz-Gibbon ST, Oren A, House CH. (2010). Amino acid signatures of salinity on an environmental scale with a focus on the Dead Sea. *Environ Microbiol* **12**: 2613–2623.

Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M *et al.* (2010). Viral and microbial community dynamics in four aquatic environments. *ISME J* **4**: 739–751.

Rokas A, Williams BL, King N, Carroll SB. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**: 798–804.

Sabet S, Diallo L, Hays L, Jung W, Dillon JG. (2009). Characterization of halophiles isolated from solar salterns in Baja California, Mexico. *Extremophiles* **13**: 643–656.

Sime-Ngando T, Lucas S, Robin A, Tucker KP, Colombet J, Bettarel Y *et al.* (2010). Diversity of virus-host systems in hypersaline Lake Retba, Senegal. *Environ Microbiol*, doi: 10.1111/j.1462-2920.2010.02323.x.

Sipos R, Szekely AJ, Palatinszky M, Revesz S, Marialigeti K, Nikolausz M. (2007). Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol Ecol* **60**: 341–350.

Susko E, Roger AJ. (2007). On reduced amino acid alphabets for phylogenetic inference. *Mol Biol Evol* **24**: 2139–2150.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.

Verhees CH, Kengen SW, Tuininga JE, Schut GJ, Adams MW, De Vos WM *et al.* (2003). The unique features of glycolytic pathways in Archaea. *Biochem J* **375**: 231–246.

Wall DP, Deluca T. (2007). Ortholog detection using the reciprocal smallest distance algorithm. *Methods Mol Biol* **396**: 95–110.

Walsby AE. (1994). Gas vesicles. *Microbiol Rev* **58**: 94–144.

Wilmes P, Simmons SL, Denef VJ, Banfield JF. (2009). The dynamic genetic repertoire of microbial communities. *FEMS Microbiol Rev* **33**: 109–132.

Wooley JC, Godzik A, Friedberg I. (2010). A primer on metagenomics. *PLoS Comput Biol* **6**: e1000667.

Woyke T, Xie G, Copeland A, Gonzalez JM, Han C, Kiss H *et al.* (2009). Assembling the marine metagenome, one cell at a time. *PLoS One* **4**: e5299.

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN *et al.* (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**: 1056–1060.

Wu M, Eisen JA. (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* **9**: R151.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)

Chapter 2 is a full reprint of: De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline communnities. P. Narasingarao, S. Podell, J.A. Ugalde, C. Brochier-Armanet, J.B. Emerson, J.J. Brocks, K.B. Heidelberg, J.F. Banfield and E.E. Allen. *ISME Journal*, **6**,81-93. 2012 (doi: 10.1038/ismej.2011.78), with permission from all coauthors.

Chapter 3

Xenorhodopsins, an Enigmatic New Class of Microbial Rhodopsins Horizontally Transferred Between Archaea and Bacteria

Ugalde *et al. Biology Direct* 2011, **6**:52
http://www.biology-direct.com/content/6/1/52

BIOLOGY DIRECT

DISCOVERY NOTES Open Access

Xenorhodopsins, an enigmatic new class of microbial rhodopsins horizontally transferred between archaea and bacteria

Juan A Ugalde[1], Sheila Podell[1], Priya Narasingarao[1] and Eric E Allen[1,2*]

Abstract

Based on unique, coherent properties of phylogenetic analysis, key amino acid substitutions and structural modeling, we have identified a new class of unusual microbial rhodopsins related to the Anabaena sensory rhodopsin (ASR) protein, including multiple homologs not previously recognized. We propose the name xenorhodopsin for this class, reflecting a taxonomically diverse membership spanning five different Bacterial phyla as well as the Euryarchaeotal class Nanohaloarchaea. The patchy phylogenetic distribution of xenorhodopsin homologs is consistent with historical dissemination through horizontal gene transfer. Shared characteristics of xenorhodopsin-containing microbes include the absence of flagellar motility and isolation from high light habitats.
Reviewers: This article was reviewed by Dr. Michael Galperin and Dr. Rob Knight.

Findings

Microbial rhodopsins are a widespread family of photoactive proteins found in all three domains of life. Based on their functional roles, characterized rhodopsin proteins have been classified into three distinct groups: (i) Proton pumps (bacteriorhodopsins and proteorhodopsins), involved in energy generation, (ii) Chloride pumps (halorhodopsins), involved in the maintenance of osmotic balance, and (iii) Sensory rhodopsins, which direct positive and/or negative phototaxis. Microbial proton pumps have the widest ecological niche distribution, and are found throughout the Bacteria and Archaea in hypersaline, marine, and freshwater habitats [1]. Chloride pumps and sensory rhodopsins are mostly limited to halophilic Archaea of class Halobacteria [1], excepting the few characterized examples in the freshwater cyanobacterium *Anabaena (Nostoc)* sp. PCC 7120 [2,3] and eukaryotic green algae including *Chlamydomonas reinhardtii* [4].

The evolutionary history of microbial rhodopsins is complex, showing broad but patchy phylogenetic distribution within and across disparate lineages. It has been suggested that horizontal gene transfer (HGT) has disseminated photoreceptor and photosensory activities across large evolutionary distances [1]. One salient example is a putative sensory rhodopsin found in the bacterium *Anabaena (Nostoc)* sp. PCC 7120 (Anabaena sensory rhodopsin, ASR). It has been suggested that this protein was originally acquired from a halophilic archaeon by HGT, and may play a sensory role [1,2]. However, sensory function performance has not yet been demonstrated experimentally, and the ASR protein differs from previously described sensory rhodopsins in: (i) a distinct signaling cascade mechanism that employs a soluble transducer protein, rather than the methyl-accepting taxis transducers (HTR proteins) found in halophilic Archaea [2,5] and (ii) its divergent photochemistry, including unique light-induced *cis/trans* configuration dynamics of the retinal chromophore, providing a possible mechanism for sensing and differentiating specific light qualities [3,6].

In the current study, we report the discovery of several new ASR protein homologs with shared characteristics consistent with the designation of a new class of microbial rhodopsins. ASR homologs were found in *Nanosalina* sp. J07AB43 and *Nanosalinarum* sp. J07AB56, the first representatives of a newly described major lineage of Archaea (class Nanohaloarchaea) within phylum Euryarchaeota [7]. The *Nanosalina* sp. and *Nanosalinarum* sp. rhodopsin proteins are highly similar

* Correspondence: eallen@ucsd.edu
[1]Marine Biology Research Division, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093-0202, USA
Full list of author information is available at the end of the article

Ugalde *et al. Biology Direct* 2011, **6**:52
http://www.biology-direct.com/content/6/1/52

to each other (89% amino acid identity) and are present in both genomes as single copy genes. Surprisingly, these two Nanohaloarchaeal proteins most closely resemble rhodopsins in taxonomically distant *Cyanothece* sp. PCC 7424 and *Anabaena* (*Nostoc*) sp. PCC 7120, at 31 and 34% amino acid identity respectively. No homologs were identified in other members of the Euryarchaeota, although related proteins were detected at 30-31% amino acid identity in *Bacillus coahuilensis* m4-4 (phylum Firmicutes), a sporulating halophilic bacterium isolated from a desiccation lagoon [8], the psychrophilic bacterium *Hymenobacter roseosalivarius* AA-718 (phylum Bacteroidetes), and the halophilic bacterium *Haloplasma contractile* SSD-17B (phylum Tenericutes) [9,10].

Figure 1 shows a phylogenetic analysis using maximum likelihood and Bayesian inference methods for the ASR homologs, together with a set of representative protein sequences from all previously recognized functional microbial rhodopsin classes. Methods and experimental procedures are provided in Additional File 1. The phylogenetic tree also includes additional sequences we obtained by PCR amplification using primers specifically targeting Nanohaloarchaeal rhodopsin genes. These sequences were recovered from a hypersaline environment (South Bay Salt Works, Chula Vista, California, USA) that is geographically distant from the original isolation site of the Nanohaloarchaea genomes (Lake Tyrrell, Victoria, Australia). Tree topology shows robust clustering of all ASR homologs as a single clade, distinct from other rhodopsin types. We propose the name "xenorhodopsins" to describe this class of rhodopsin proteins, articulating the wide taxonomic diversity of its members.

The patchy distribution and topology of the xenorhodopsin clade is consistent with HGT events between

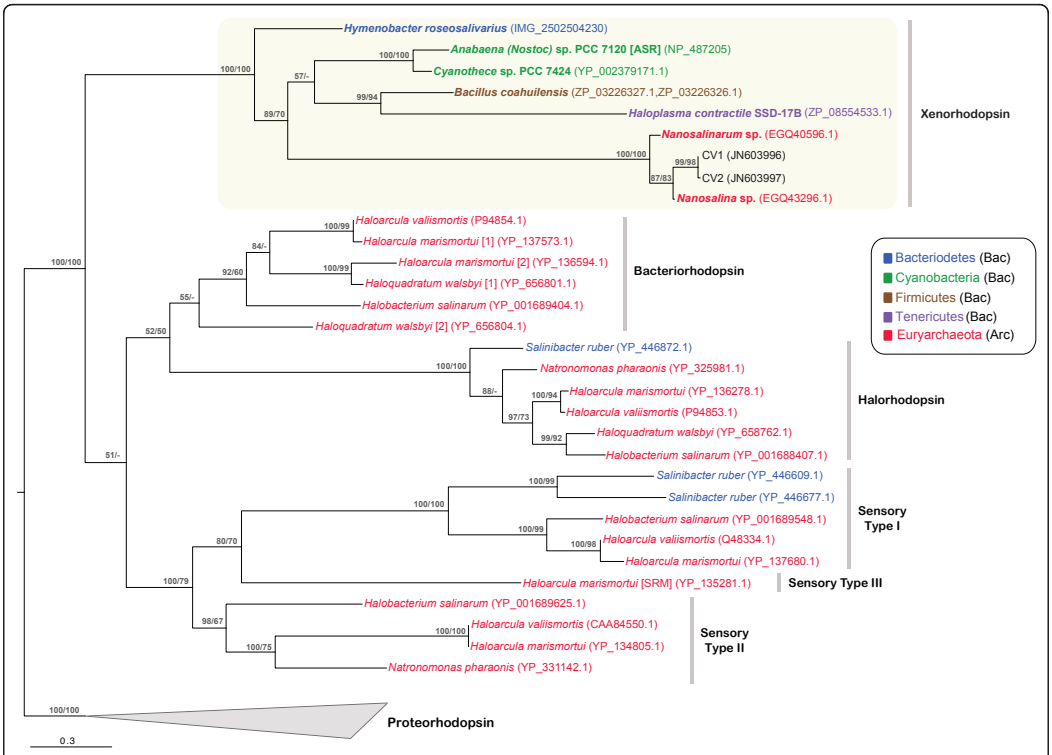


Figure 1 Phylogenetic tree of microbial rhodopsin proteins showing diversity of functional classes. Tree is based on a total of 34 sequences (205 amino acid positions) using maximum likelihood and Bayesian inference methods. Numbers at nodes represent posterior probablities inferred by MrBayes (first value) and maximum likelihood bootstrap values using RaxML (second value). Only values greater than 50% are shown. GenBank accession numbers are shown in parentheses for each protein except *H. roseosalivarius* (IMG-ER database gene object ID) [22]. Sequences CV1 and CV2 were recovered by PCR amplification of environmental DNA from a solar saltern in Chula Vista, California, USA, using Nanohaloarchaeal-specific xenorhodopsin primers (see Additional File 1).

domains and involving five disparate bacterial phyla. The large numbers of currently sequenced Firmicute (873), Bacteriodetes (169), Cyanobacteria (68), and Haloarchaea genomes (18) lacking xenorhodopsin homologs make it unlikely that the gene/species tree incongruencies shown in Figure 1 could be explained by independent gene loss among multiple species. Sufficiency of taxon sampling and information content in our 205-position trimmed amino acid sequence alignment (Supplementary File 2) are well supported by significant bootstrap values (Figure 1), and corroborated by complete topological agreement between trees constructed using Bayesian and maximum likelihood methods. Additionally, the new xenorhodopsin sequences identified here do not change overall tree topologies of other microbial rhodopsin sequences previously reported in the literature [11].

To supplement HGT analysis based on phylogenetic incongruencies, DNA signature patterns were analyzed for individual xenorhodopsin proteins relative to the genomes in which they were found, based on percent G +C, codon usage patterns, and Interpolated Variable Order Motifs [12] (Additional File 1: Table S1). By all of these criteria, xenorhodopsin genes in *Nanosalinarum* J07AB56, *Cyanothece* PCC 7424, *Nostoc* PCC 7120, *Hymenobacter roseosalivarius*, and *Haloplasma contractile* closely resemble other loci within their respective genomes. These data support the likelihood that the observed incongruencies between xenorhodopsin protein and species trees for these genomes represent ancient rather than recent HGT events, with subsequent amelioration of foreign DNA signatures over time. A different pattern was observed for xenorhodopsin proteins in the *Bacillus coahuilensis* and *Nanosalina* J07AB43 genomes, where atypical codon usage suggests that HGT events may have occurred more recently (Additional File 1: Table S1).

The absence of xenorhodopsin genes in all Euryarchaeaota other than members of class Nanohaloarchaea suggests that these genes were acquired subsequent to divergence of Nanohaloarchaeaota from other Euryarchaeotal classes. The high degree of similarity among xenorhodopsin proteins obtained from two different Nanohalorchaeal genera, as well as environmental sequences from a distant geographical location (North America versus Australia), is consistent with inheritance from a common ancestral source, coupled with strong selective pressure for amino acid sequence conservation. The discrepancy between ancestral inheritance and the atypical codon usage pattern observed in the *Nanosalina* J07AB43 protein may be explained by relatively recent secondary exchange with other Nanohaloarchaea, as multiple genera of this lineage are known to coexist in shared habitats [7].

The phylogenetic tree presented in Figure 1 includes only known, modern representatives of lineages that may have incorporated multiple HGT events between extinct ancestors and/or serial exchanges with unknown species whose genomes have not yet been sequenced. Although the complexity of these relationships precludes confident reconstruction of the exact timing, direction, and order of individual gene transfer events, cross-domain and cross-phylum gene acquisition through HGT provides the most parsimonious explanation for the data.

Amino acid alignments of residues known to determine function for previously characterized microbial rhodopsins are inconsistent with proton or chloride transporting activity for xenorhodopsins, suggesting a possible sensory role (see Additional File 2 for full alignment). Figure 2 shows that residues required to bind the retinal chromophore molecule are conserved across all xenorhodopsin group members. Ion transporting rhodopsins can be distinguished from sensory rhodopsins by comparing the residues that serve as the retinal Schiff base proton donor and proton acceptor during the photocycle [2,13]. These residues correspond to Asp98 (acceptor) and Asp109 (donor) in the *H. salinarum* bacteriorhodopsin (Helix C). Consistent with previously described sensory rhodopsins, ASR and all other xenorhodopsin homologs lack the canonical Asp residue at the donor position, a hallmark of proton translocating rhodopsins. Likewise, known sensory rhodopsins and xenorhodopsins both lack the Thr (acceptor) and Ala (donor) configuration diagnostic of chloride pumps (Figure 2).

Despite the insights provided by these results, it is not possible to predict functional activity based on sequence alignment alone. The structural sensitivity of microbial rhodopsins is highlighted by the ability to engineer aberrant functional properties in these proteins. A single amino acid substitution, Asp217 to Glu, has been shown to confer inward proton pumping activity to the ASR protein [14] and a single amino acid substitution is sufficient to convert a bacteriorhodopsin proton pump into a chloride pump [15].

One prominent difference between the xenorhodopsins and all other microbial rhodopsin proteins is a universal Pro to Asp substitution (Helix G), a substitution noted previously in the *Anabaena (Nostoc)* sp. PCC 7120 and *B. coahuilensis* homologs [8,16]. The shared position of this residue in all xenorhodopsins discovered to date suggests that it may provide a useful diagnostic for this protein class.

Sequence conservation and phylogenetic analysis of xenorhodopsin proteins is strongly supported by comparative 3-dimensional protein structure modeling. This similarity is illustrated in Additional File 3, showing a

Ugalde *et al. Biology Direct* 2011, **6**:52
http://www.biology-direct.com/content/6/1/52

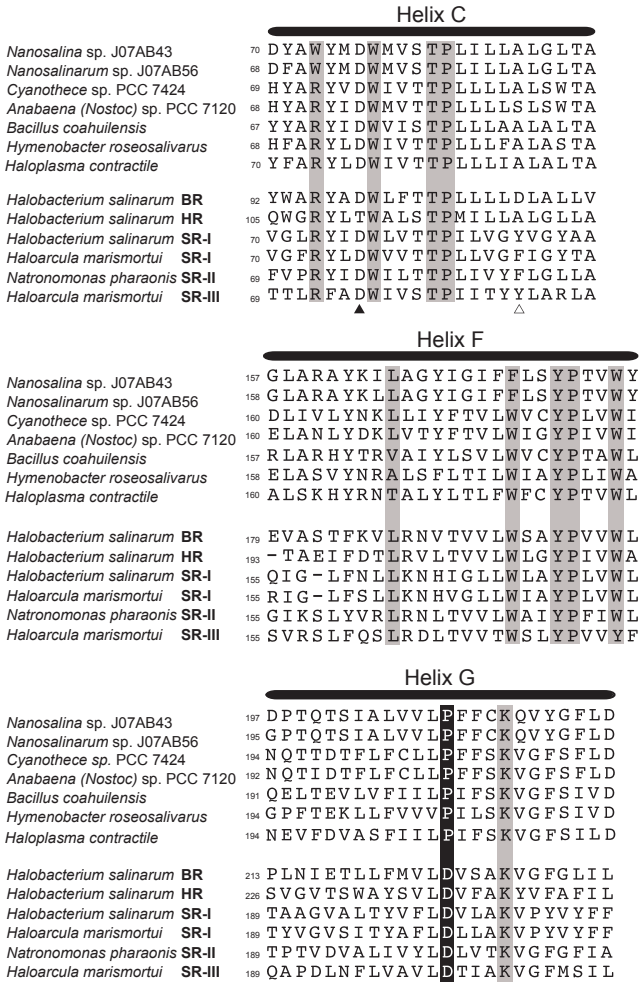


Figure 2 Protein alignment of key amino acid segments. Sequences include all xenorhodopsin homologs plus selected representatives of previously characterized microbial rhodopsin functional classes; bacteriorhodopsin (BR), halorhodopsin (HR) and sensory rhodopsins I (SR-I), II (SR-II) and III (SR-III). Shaded boxes indicate conserved residues involved in retinal binding [3]. The black box in Helix G shows a conserved Asp to Pro substitution in all xenorhodopsin proteins at this position. Retinal Schiff base proton acceptor (closed triangle) and proton donor (open triangle) residues are marked in Helix C.

SWISS-MODEL [17] prediction of the *Nanosalina* sp. rhodopsin structure using ASR as a template. The modeled structure demonstrates high congruence in residues that form the retinal binding pocket, as well as similar truncations in loop motifs (Additional File 3). The conserved primary and tertiary structure of xenorhodopsins combined with their distinct phylogenetic clustering supports their classification as a coherent, highly conserved group.

An important element of previously characterized sensory rhodopsins in halophilic Archaea is the presence of a signal transduction mechanism, most often genetically encoded in a genomic position adjacent to the rhodopsin gene [18]. In *Anabaena (Nostoc)* sp. PCC 7120, the proposed soluble transducer protein ASRT (*Anabaena* sensory rhodopsin transducer) is encoded by a gene in the same operon as ASR [18] (Figure 3). Consistent with its putative role in light-activated sensory transduction,

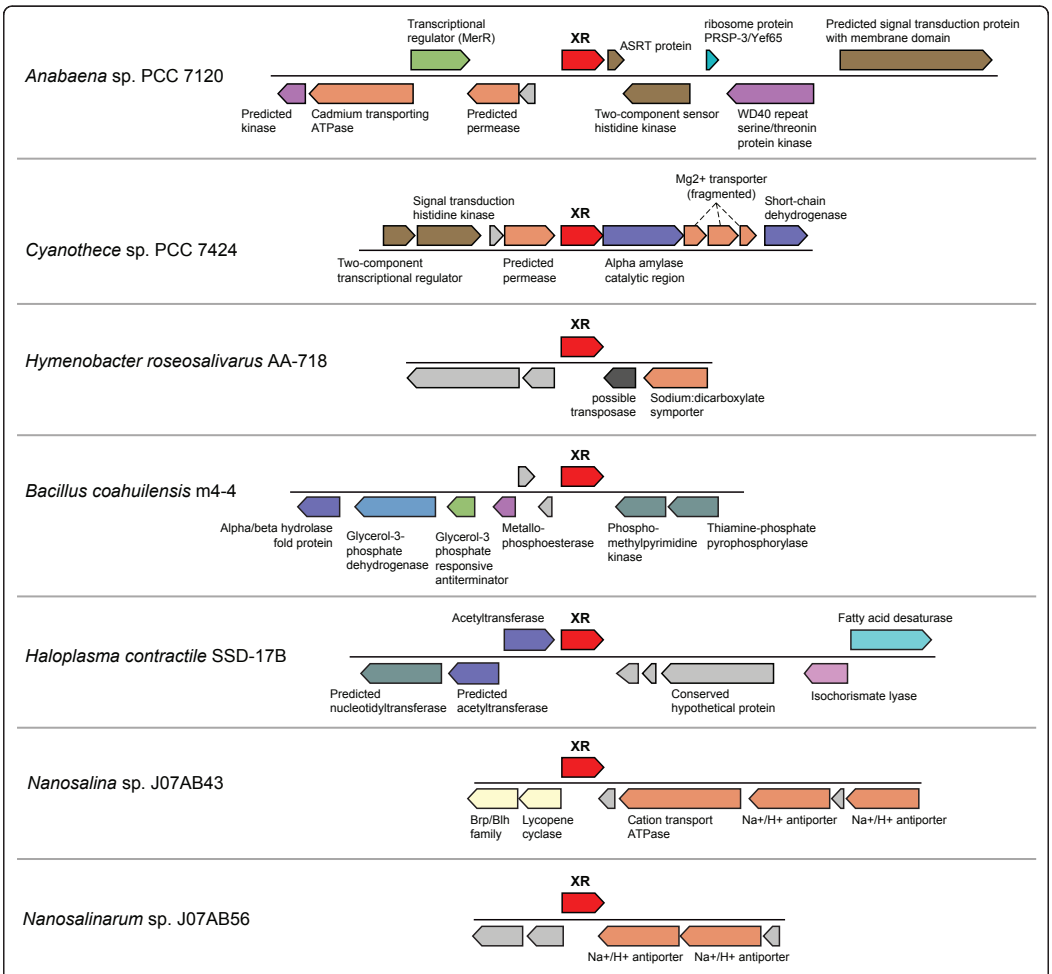


Figure 3 Genomic neighborhood of xenorhodopsin (XR) genes. Functional annotations were obtained from NCBI [23] and IMG [22]. Similar gene functions are color-coded. Hypothetical proteins are unlabeled and shown in gray.

the ASRT protein has been shown to bind DNA, specifically the promoter region of genes involved in the synthesis of light-harvesting accessory pigments [19]. However, no homologs of ASRT were identified in other genomes containing a xenorhodopsin gene, suggesting the ASR-ASRT association is a specific feature of *Anabaena (Nostoc)* sp. PCC 7120. Moreover, the identification of ASRT homologs in numerous bacterial and archaeal genomes that lack an ASR (xenorhodopsin) homolog suggests the ASRT protein family is not specific to photosensory signal transduction processes.

The lack of identifiable common transducer elements suggests possible plasticity in the transducer component (s) modulating possible xenorhodopsin-mediated photosensory activity. For example, *Cyanothece sp*. PCC 7424 has genes encoding a two-component regulatory system within the same genomic neighborhood as the xenorhodopsin gene (Figure 3). The two Nanohaloarcheal genomes (*Nanosalina* sp. and *Nanosalinarum* sp.) have genes encoding a putative Na^+/H^+ antiporter system adjacent to the rhodopsin gene. The high sequence identity shared between these transporter sequences along with their conserved genomic location is atypical for these two archaeal genomes, representing different genera, which are generally non-syntenic [7]. It is tempting to speculate that genes in this local region of

Ugalde *et al. Biology Direct* 2011, **6**:52
http://www.biology-direct.com/content/6/1/52

conservation could be related to rhodopsin function in these organisms.

Despite highly diverse taxonomic origins, the seven species possessing a xenorhodopsin protein share a number of common characteristics, including the absence of flagellar motility, relatively low genomic percent G+C content and isolation from habitats with a high incidence of UV light (Additional File 1: Table S2). The lack of flagellar motility is noteworthy because it eliminates the potential usefulness of previously characterized sensory rhodopsin classes which act by influencing the rotational state of the flagellar motor for phototaxis. The particularly low G+C compositions of *Nanosalina* sp. (43%), *Anabaena (Nostoc)* sp. PCC 7120 (41%), *Cyanothece* sp. PCC 7424 (38%), *Bacillus coahuilensis* (38%) and *H. contractile* (33.6%) are atypical for unicellular inhabitants of high light environments, rendering them especially sensitive to potential UV damage via the formation of thymidine dimers. The isolation of *H. contractile* from deep marine sediments, where light is not a factor, may be an anomaly, since closely related 16S rRNA gene sequences have also been found in high-light solar saltern environments [9].

Consistent with these observations, one intriguing hypothesis is that xenorhodopsins may play a role in pre-emptive photoprotection by inducing light-dependent changes in the expression of photoprotective pigments, a role proposed for the ASR protein due to its photochromic properties [3,6]. Alternatively, these proteins could be linked to expression of DNA repair mechanisms. However, these speculations must be tempered by the caveat that no sensory or ion transport function has yet been experimentally validated for ASR, or any other xenohodopsin protein. Future work on the biochemistry, photochemistry, and molecular genetic characterization of the xenorhodopsin class of proteins will undoubtedly provide fascinating insights into their physiological function in light-induced biological processes.

Reviewers' comments

Reviewer 1

Dr. Michael Y. Galperin, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

I agree with the authors' conclusion that Anabaena sensory rhodopsin (ASR) and closely related proteins form a separate family of rhodopsins. However, I believe that the current version of the paper would need a substantial revision to become acceptable for Biology Direct.

The notion that ASR comprises a new type of sensory rhodopsins is not new and should not be presented as such. Spudich and colleagues described the uniqueness of ASR in their early papers [2,16] and unequivocally stated that ASR belongs to a separate family [6]. This does not diminish the contribution of this work, which describes six new members of that family, but the text of the Abstract and the tone of the whole paper must be changed.

Author's response

We thank the reviewer for bringing to our attention these deficiencies in our original summary of previous work recognizing the uniqueness of the ASR protein. We have modified the manuscript to address these issues by changing the title, the abstract, and the interpretational emphasis of our text. We believe these revisions clarify the significance of our findings in discovering that the ASR protein is not a single, isolated anomaly, but rather part of a large, cohesive family of related proteins with an unusual taxonomic distribution. To further emphasize this point, we propose the name "xenorhodopsin" to describe the members of this group, rather than calling them ASR-like (or Sensory Rhodopsin-IV) proteins.

Although the name "Anabaena sensory rhodopsin" is being widely used in the literature, it is important to note that there has been no experimental proof that this protein actually performs sensory function. Indeed, ASR has been shown not to function as a proton pump and it has been reasoned that it is unlikely to work as a chloride pump. Nevertheless, there remains a distinct possibility that ASR functions as a membrane pump for some other ion, for example, sodium. This proposal is hardly more speculative than the suggestion of the sensory function and is supported by at least three lines of evidence:

1) the adjacency of genes coding for ASR homologs and Na^+/H^+ antiporters, noted by the authors themselves.

2) the observation of Kawanabe *et al.* [14] that a single amino acid change converts ASR into an inward proton pump; and 3) the observation of De Souza *et al.* [20] that so-called ASR transducer is found in a variety of genomes that do not encode ASR and is likely to bind sugars. Further, the previously overlooked absence of the ASRT gene in the complete genome of *Cyanothece* sp. PCC7424 and its recently reported ability to bind DNA [19] strongly suggest that the putative ASR-ASRT signaling cascade is a specific feature of *Anabaena* sp. PCC7120. The authors correctly point out the absence of flagellar motility in the ASR-carrying organisms; this argument, however, is weakened by the chemotactic ability of both *Anabaena* sp. PCC7120 and *Cyanothece* sp. PCC7424, owing to the presence of 3 and 9 methyl-accepting chemotaxis sensors, respectively [21]. In the absence of direct experimental data, the authors should

Ugalde *et al. Biology Direct* 2011, **6**:52
http://www.biology-direct.com/content/6/1/52

discuss possible alternative functions of the ASR-like family and should be more careful in describing this new rhodopsin family as sensory rhodopsins.

Author's response

We have expanded the text to include a discussion of possible alternative functions for the xenorhodopsin family, including how lack of experimental evidence for ASR sensory function affects interpretation of conserved amino acid sequences, the importance of mutational experiments demonstrating gain of inward proton pumping function, and the apparent species-specific nature of the ASR/ASRT interactions.

I would also suggest moving the Supplementary Figure S1 (Genomic neighborhood of SR-IV genes) to the main text.

Author's response

The previously presented Supplementary Figure S1 is now Figure 3*in the main text.*

Reviewer 2

Dr. Rob Knight, Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO, USA

In this manuscript, the authors analyze a set of microbial rhodopsin sequences (including some that they amplified for this study from an environmental sample), and demonstrate that there is a new clade of sensory rhodopsins that is basal, with high bootstrap support, and that includes sequences from a surprisingly broad phylogenetic range (including one archaeal and three bacterial phyla). This distribution is interesting because previous studies of sensory rhodopsins have found them primarily in the Euryarchaeota and in the Bacteroidetes.

The methods are generally sound except that the taxonomy of the sister groups to the new clade is poorly resolved (i.e. non-significant bootstraps), and it would be reassuring if the split were confirmed using other phylogenetic methods besides likelihood (e.g. distance or Bayesian methods) before the new set of sequences was claimed as distinct.

Author's response

We have supplemented our original phylogenetic analysis with Bayesian and distance-based methods, and find that all agree in supporting identical tree topologies. We have revised Figure 1*and the text to clarify the fact that the topologies agree and that bootstrap values supporting branches relevant to the new clade are highly significant using all methods.*

Additionally, although the patchy phylogenetic distribution is suggestive of horizontal gene transfer, formal methods (of which several exist) should be used to confirm HGT as opposed to other factors that can lead to gene/species tree incongruence

Author's response

Although many methods of HGT detection have been proposed in the literature, their lack of consistency and potential unreliability in the face of complex, real world data have long been a matter of controversy and debate. Phylogenetic tree incongruency is currently considered the gold standard by which all other HGT prediction methods are judged, and this is the primary technique we have used to reach conclusions presented in the manuscript, which we believe are compelling.

To supplement the phylogenetic analyses, we have performed several additional HGT analyses using methods based on DNA signature patterns, included these results as Supplementary Table S1, and expanded discussion of HGT in the text to include interpretation of these additional results.

Additional material

Additional File 1: Supplementary Methods and Tables.

Additional File 2: Trimmed amino acid alignment file of microbial rhodopsin sequences.

Additional File 3: SWISS-MODEL 3-dimensional protein structure model of *Nanosalina* sp. xenorhodopsin using ASR as a template.

Acknowledgements and Funding

This work was supported by NSF award number 0626526 (Emerging Frontiers; Microbial Genome Sequencing Program) and NIH award R21HG005107-02 (NHGRI). Juan A. Ugalde was supported by a Fulbright-CONICYT graduate fellowship.

Author details

[1]Marine Biology Research Division, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093-0202, USA. [2]Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093, USA.

Authors' contributions

All authors conceived the study. JAU and SP performed sequence analysis. PN performed experiments. JAU, SP and EEA wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 27 June 2011 Accepted: 10 October 2011
Published: 10 October 2011

References

1. Sharma AK, Spudich JL, Doolittle WF: **Microbial rhodopsins: functional versatility and genetic mobility.** *Trends Microbiol* 2006, **14(11)**:463-469.
2. Jung KH, Trivedi VD, Spudich JL: **Demonstration of a sensory rhodopsin in eubacteria.** *Mol Microbiol* 2003, **47(6)**:1513-1522.
3. Spudich JL: **The multitalented microbial sensory rhodopsins.** *Trends Microbiol* 2006, **14(11)**:480-487.
4. Sineshchekov OA, Govorunova EG, Spudich JL: **Photosensory functions of channelrhodopsins in native algal cells.** *Photochem Photobiol* 2009, **85(2)**:556-563.
5. Vogeley L, Trivedi VD, Sineshchekov OA, Spudich EN, Spudich JL, Luecke H: **Crystal structure of the Anabaena sensory rhodopsin transducer.** *J Mol Biol* 2007, **367(3)**:741-751.

6. Sineshchekov OA, Trivedi VD, Sasaki J, Spudich JL: **Photochromicity of Anabaena sensory rhodopsin, an atypical microbial receptor with a cis-retinal light-adapted form.** *J Biol Chem* 2005, **280(15)**:14663-14668.
7. Narasingarao P, Podell S, Ugalde J, Brochier-Armanet C, Emerson J, Brocks J, Heidelberg KB, Banfield J, Allen EE: ***De novo* metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities.** *ISME J* Advanced Online Publication; 2011.
8. Alcaraz LD, Olmedo G, Bonilla G, Cerritos R, Hernandez G, Cruz A, Ramirez E, Putonti C, Jimenez B, Martinez E, *et al*: **The genome of *Bacillus coahuilensis* reveals adaptations essential for survival in the relic of an ancient marine environment.** *Proc Natl Acad Sci USA* 2008, **105(15)**:5803-5808.
9. Antunes A, Rainey FA, Wanner G, Taborda M, Patzold J, Nobre MF, da Costa MS, Huber R: **A new lineage of halophilic, wall-less, contractile bacteria from a brine-filled deep of the Red Sea.** *J Bacteriol* 2008, **190(10)**:3580-3587.
10. Antunes A, Alam I, El Dorry H, Siam R, Robertson A, Bajic VB, Stingl U: **Genome sequence of *Haloplasma contractile*, an unusual contractile bacterium from a deep-sea anoxic brine lake.** *J Bacteriol* 2011, **193(17)**:4551-4552.
11. Sharma AK, Spudich JL, Doolittle WF: **Microbial rhodopsins: functional versatility and genetic mobility.** *Trends Microbiol* 2006, **14(11)**:463-469.
12. Vernikos GS, Parkhill J: **Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands.** *Bioinformatics* 2006, **22(18)**:2196-2203.
13. Klare JP, Chizhov I, Engelhard M: **Microbial rhodopsins: scaffolds for ion pumps, channels, and sensors.** *Results Probl Cell Differ* 2008, **45**:73-122.
14. Kawanabe A, Furutani Y, Jung KH, Kandori H: **Engineering an inward proton transport from a bacterial sensor rhodopsin.** *J Am Chem Soc* 2009, **131(45)**:16439-16444.
15. Sasaki J, Brown LS, Chon YS, Kandori H, Maeda A, Needleman R, Lanyi JK: **Conversion of bacteriorhodopsin into a chloride ion pump.** *Science* 1995, **269(5220)**:73-75.
16. Vogeley L, Sineshchekov OA, Trivedi VD, Sasaki J, Spudich JL, Luecke H: **Anabaena sensory rhodopsin: a photochromic color sensor at 2.0 A.** *Science* 2004, **306(5700)**:1390-1393.
17. Kiefer F, Arnold K, Kunzli M, Bordoli L, Schwede T: **The SWISS-MODEL Repository and associated resources.** *Nucleic Acids Res* 2009, **37(Database)**: D387-392.
18. Jung KH: **The distinct signaling mechanisms of microbial sensory rhodopsins in Archaea, Eubacteria and Eukarya.** *Photochem Photobiol* 2007, **83(1)**:63-69.
19. Wang S, Kim SY, Jung KH, Ladizhansky V, Brown LS: **A eukaryotic-like interaction of soluble cyanobacterial sensory rhodopsin transducer with DNA.** *J Mol Biol* 2011, **411(2)**:449-62.
20. De Souza RF, Iyer LM, Aravind L: **The *Anabaena* sensory rhodopsin transducer defines a novel superfamily of prokaryotic small-molecule binding domains.** *Biol Direct* 2009, **4**:25.
21. **Census of bacterial signal transduction proteins.** [http://www.ncbi.nlm.nih.gov/Complete_Genomes/SignalCensus.html].
22. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Anderson I, Lykidis A, Mavromatis K, *et al*: **The integrated microbial genomes system: an expanding comparative analysis resource.** *Nucleic Acids Res* 2010, **38(Database)**:D382-390.
23. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2009, **37(Database)**:D26-31.

doi:10.1186/1745-6150-6-52
Cite this article as: Ugalde *et al.*: Xenorhodopsins, an enigmatic new class of microbial rhodopsins horizontally transferred between archaea and bacteria. *Biology Direct* 2011 **6**:52.

Chapter 3 is a full reprint of: Xenorhodopsins, an enigmatic new class of microbial rhodopsins horizontally transferred between archaea and bacteria. J.A. Ugalde, S. Podell, P. Narasingarao and E.E. Allen. *Biology Direct*, **6**,52. 2011 (doi: 10.1186/1745-6150-6-52), with permission from all coauthors.

Chapter 4

Assembly-driven Community Genomics of a Hypersaline microbial Ecosystem

OPEN ACCESS Freely available online

PLOS ONE

Assembly-Driven Community Genomics of a Hypersaline Microbial Ecosystem

Sheila Podell[1], Juan A. Ugalde[1], Priya Narasingarao[1], Jillian F. Banfield[2,3], Karla B. Heidelberg[4], Eric E. Allen[1,5]*

1 Marine Biology Research Division, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, United States of America, **2** Department of Earth and Planetary Sciences, University of California, Berkeley, California, United States of America, **3** Department of Environmental Science, Policy, and Management, University of California, Berkeley, California, United States of America, **4** Department of Biological Sciences, University of Southern California, Los Angeles, California, United States of America, **5** Division of Biological Sciences, University of California San Diego, La Jolla, California, United States of America

Abstract

Microbial populations inhabiting a natural hypersaline lake ecosystem in Lake Tyrrell, Victoria, Australia, have been characterized using deep metagenomic sampling, iterative *de novo* assembly, and multidimensional phylogenetic binning. Composite genomes representing habitat-specific microbial populations were reconstructed for eleven different archaea and one bacterium, comprising between 0.6 and 14.1% of the planktonic community. Eight of the eleven archaeal genomes were from microbial species without previously cultured representatives. These new genomes provide habitat-specific reference sequences enabling detailed, lineage-specific compartmentalization of predicted functional capabilities and cellular properties associated with both dominant and less abundant community members, including organisms previously known only by their 16S rRNA sequences. Together, these data provide a comprehensive, culture-independent genomic blueprint for ecosystem-wide analysis of protein functions, population structure, and lifestyles of co-existing, co-evolving microbial groups within the same natural habitat. The ''assembly-driven'' community genomic approach demonstrated in this study advances our ability to push beyond single gene investigations, and promotes genome-scale reconstructions as a tangible goal in the quest to define the metabolic, ecological, and evolutionary dynamics that underpin environmental microbial diversity.

Citation: Podell S, Ugalde JA, Narasingarao P, Banfield JF, Heidelberg KB, et al. (2013) Assembly-Driven Community Genomics of a Hypersaline Microbial Ecosystem. PLoS ONE 8(4): e61692. doi:10.1371/journal.pone.0061692

Editor: Melanie R. Mormile, Missouri University of Science and Technology, United States of America

Received December 21, 2012; **Accepted** March 13, 2013; **Published** April 18, 2013

Funding: Funding for this work was provided by NSF award number 0626526 (JFB, KBH, EEA) and NIH award R21HG005107-02 (EEA). JAU was supported by a Fulbright- Conicyt fellowship. Work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No DE-AC02- 05CH11231. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: eallen@ucsd.edu

Introduction

Microbial diversity studies based on 16S rRNA gene amplification have identified large numbers of uncultured, uncharacterized organisms whose metabolic capabilities, lifestyle strategies, and ecosystem contributions remain largely unknown. Conversely, the subset of cultured microbial species from any particular habitat often fails to include even some of the most abundant members of the community. Efforts to bring these unknown organisms into laboratory culture are confounded by our limited understanding of the metabolic specializations of environmental microorganisms, the interdependencies of intra-/inter-species interactions, and the physicochemical conditions that promote or diminish microbial survival and population structure in natural environments.

Direct metagenomic sequencing of environmental samples can potentially provide functional information missing from 16S rRNA gene surveys and circumvent the constrained diversity found in representative cultured isolates. Composite genomes have been assembled from several environmental data sets [1,2,3,4,5,6], however comprehensive characterization of the genetic diversity of most naturally occurring microbial communities remains a significant challenge. Environmental sampling of predicted metabolic functions as a simple ''bag of genes'' via metagenomic read-based analysis cannot fully capture the genetic and metabolic potential of individual populations, and may overlook the significance of community-wide processes involving cooperative interactions between multiple species [7,8,9].

Reference genomes from cultured isolates and/or single-cell projects can greatly assist in taxonomic assignment of genes encoded on short metagenomic DNA fragments. However, with the recent exception of the human microbiome project [10], the time, effort, and expense required to develop reference resources of sufficient breadth to adequately represent the full diversity of most ecosystems using these methods are currently prohibitive, and the vast majority of environmentally identified species remain uncharacterized.

The issue of inadequate database representation is particularly relevant for microbial communities in extreme hypersaline aquatic environments, which are dominated by archaeal populations [11]. These environments provide an attractive model for studying microbial ecology, because the demands of surviving such extreme conditions limit taxonomic diversity, yet cell densities frequently exceed 10^7–10^8 per mL [12]. The aquatic milieu allows convenient large-scale sampling and fractionation of discrete

populations in particular size ranges, simplifying many types of analysis. These ecosystems have been well-studied historically using culture dependent-methods, 16S rRNA gene surveys and, more recently, single-cell genomics and metagenomics (reviewed in [11]). Despite these advances, the number of available sequenced genomes relevant to microbial communities in this specific habitat remains very small, and is not representative of the *in situ* diversity present in a natural microbial assemblage.

The extreme hypersaline habitat of Lake Tyrrell, Australia has recently been used to demonstrate the utility of *de novo* metagenomic assembly for characterizing organisms previously known only by their 16S rRNA gene sequences, including representatives of a globally distributed new class of Archaea, the Nanohaloarchaea [13,14]. In the current study, we extend this previous work, combining cell size-fractionated sample collection, deep metagenomic sequencing, multidimensional phylogenetic binning, and iterative *de novo* assembly to reconstruct ten additional population genomes. These new genomes provide a comprehensive, culture-independent genomic blueprint for ecosystem-wide analysis of protein functions, population structure, and lifestyles linked to specific microbial strains co-existing and co-evolving within the same natural habitat.

Materials and Methods

Sample Collection, Library Construction and Sequencing

Surface water samples collected from Lake Tyrrell, Victoria, Australia at 0.3 m depth were passed through filters of decreasing porosities (20 μm>3 μm>0.8 μm>0.1 μm) to obtain fractions enriched by cellular size [13]. Physical properties of the collection site are summarized in **Table S1**. Sanger sequencing libraries were constructed at the J. Craig Venter Institute using DNA from 0.8 μm and 0.1 μm filters [15], and sequenced using both paired-end Sanger sequencing and Roche 454 Titanium pyrosequencing (**Table S2**). 16S rRNA gene clone libraries were constructed from the same DNA samples used for sequencing, using archaeal primer sequences Arc21F (5′-TTCCGGTTGATCCTGCCGGA-3′) and Arc529R (5′-ACCGCGGCKGCTGGC-3′) and bacterial primer sequences 27F (5′-AGAGTTTGATCCTGGCTCAG-3′) and 1391R (5′-GACGGGCRGTGWGTRCA-3′) [16].

Lake Tyrrell Metagenome Assembly

Assemblies were performed using Celera Assembler software version 5.4 [17]. Read sizes, library sources, and the assembled positions of reads in contigs and scaffolds were extracted from the Celera Assembler ACE output file into a local MySQL database using custom perl scripts. Numbers of scaffold nucleotides, percentages of reads obtained from different libraries, and local coverage depth for specific scaffold subregions were calculated from SQL database queries.

Figure S1 summarizes the bioinformatic assembly pipeline. All trimmed Sanger reads were combined into a composite pool for initial assembly. Scaffolds from this assembly were classified into groups using the phylogenetic binning procedures described below, then used to construct a custom reference library for PhymmBL version 3.2 [18], to assign unassembled 454 Titanium reads to taxonomic bins.

After an initial composite assembly of total community DNA, iterative rounds of *de novo* assembly were performed on taxonomic subgroups identified by scaffolds sharing common signatures based on multiple independent properties, to optimize assembly fidelity for each group individually. Each taxonomic subgroup was assembled independently using a previously described subtractive enrichment strategy based on iterative scaffold binning [13]. Scaffolds were re-binned and subsequently deconstructed into their component reads after each assembly iteration. Reads associated with scaffolds having properties characteristic of a subgroup other than the one currently being targeted were removed prior to the next round of assembly. To avoid over-pruning, singletons and reads associated with unclassified scaffolds were retained in successive rounds of assembly.

Taxonomic binning, subtractive enrichment, read deconstruction, and re-assembly steps were repeated for each taxonomic subgroup until no misassemblies were detected and no improvement was observed in completeness of conserved marker genes, maximum contig length, number and size of scaffold gaps, or uniformity of binning parameters for scaffolds >50 Kb. Assembly quality was confirmed by visual inspection using Hawkeye [19] to assess mate-pair consistency and read depth uniformity.

Archaeal genome assembly completeness was evaluated based on 53 transcription, translation, and replication genes nearly universally conserved in Archaea [20,21,22]. Bacterial draft genome completeness was assessed using the Core Gene Evaluation Script developed for the Human Microbiome Project [23]. Metagenomic sequence data has been deposited at DDBJ/EMBL/GenBank under the accession APHM00000000, NCBI BioProject number PRJNA59457. Assembled genome sequences have been deposited in the JGI-Integrated Microbial Genome resource [24] under taxon-oid numbers 2502082092 (J07HX64), 2506783034 (J07HB67), 2512875005 (J07HQW1), 2512875006 (J07HQW2), 2512875007 (J07HN4), 2512875008 (J07HN6), 2512875009 (J07HQX50), 2512875010 (J07HX5), 2512875011 (J07HR59), and 2513020022 (J07SB67).

Phylogenetic Binning and Scaffold Annotation

Raw metagenomic reads and assembled scaffolds containing 16S rRNA gene sequences were identified by BLASTN search against the GreenGenes reference database [25], requiring a minimum alignment length of 200 nucleotides and e-value of 1e-7 or better. Scaffold genes were predicted and annotated using the Integrated Microbial Genomes Expert Review (IMG/ER and IMG/MER) systems [24]. Averaged amino acid frequencies for all predicted proteins on each scaffold were calculated using a custom perl script. Taxonomic associations of predicted protein matches to GenBank nr reference sequences were tallied using DarkHorse version 1.4 [26].

Non-metric multidimensional scaling (MDS) analysis was performed on scaffolds of 5000 nucleotides or longer containing <50% gap residues using Primer version 6.1.2 [27]. Scaffold input properties included nucleotide percent G+C; read depth; percent of reads from 0.1 μm filters; percentages of lysine, arginine, threonine, glutamic acid, aspartic acid, alanine, valine and isoleucine in predicted proteins; and percent of proteins with DarkHorse-filtered best matches to Eukaryota, Bacteria, Viruses, Nanohaloarchaea, and the genera *Haloquadratum*, *Halorabdus*, *Haloarcula*, *Halorubrum*, *Haloferax*, *Halogeometricum*, and *Salinibacter*. Scaffolds sharing a common signature based on these metrics were placed in the same taxonomic bin.

Unassigned scaffolds were searched against Lake Tyrrell-specific genome assemblies using BLASTN to identify potential variant sequences associated with strain level heterogeneity present in the natural population but not captured by targeted *de novo* assembly. Unassigned scaffolds matching a composite reference genome at 85% or higher average nucleotide identity (ANI) over >40% of their length were classified in the same "population group" as the matched genome [28]. Scaffolds matching at 95% or higher ANI were assigned to the same species. Total numbers of nucleotides for binned scaffolds in each population group, including species-

level classifications, were calculated using SQL queries from assembly-specific MySQL databases, and converted to a proportional treemap graph using the TreeMap package in R, version 2.14.1 [29].

Construction of Phylogenetic Trees

The Greengenes alignment tool NAST [25] was used to construct a reference alignment of 16S rRNA genes from assembled scaffolds, cultured isolate reference genomes, and closely related environmental sequences. Maximum likelihood reference trees were constructed using RaxML version 7.2.7 [30] and FastTree version 2.1.1 [31]. Partial 16S rRNA gene sequences from unamplified metagenomic reads and Lake Tyrrell PCR amplified clone libraries were inserted into reference trees using pplacer version 1.1 [32] and visualized using Archaeopteryx version 0.968 [33]. Amplified 16S rRNA sequences from Lake Tyrrell community DNA have been submitted to NCBI under accession numbers JX880413–JX81179 (archaeal) and JX881180–JX885105 (bacterial).

Clustering of Predicted Proteins

Predicted proteins were clustered into families using an unsupervised Markov Clustering algorithm (MCL software version 10–201), with BLASTP e-value cutoff 1e-5 and inflation parameter setting 1.4 [34]. Protein family diversity was estimated using MOTHUR version 1.23.1 [35]. Assembled genomes were clustered together based on their profiles of shared protein families using the modularity analysis function of Gephi, version 0.8.1 [36].

Results

Community Sequence Assembly

Metagenomic sequence assembly effectiveness for combined Sanger libraries was assessed statistically (**Table S3**), and visualized by comparing histograms of nucleotide composition (percent G+C) for unassembled reads versus assembled scaffolds and population genomes (**Figure 1**). Raw metagenomic sequencing reads prior to assembly have a broad, biphasic nucleotide distribution, reflecting their heterogeneous origin. The percent G+C distribution of assembled scaffolds is more tightly focused into discrete peaks because the assembly process consolidates multiple overlapping reads into longer, consensus sequences with uniform properties. The length-weighted nucleotide distribution for scaffolds thus reveals overall patterns that are hidden by random noise in the shorter read sequences.

Because the percent G+C content of individual microorganisms tends to be relatively uniform when averaged over long stretches of DNA, consolidated scaffold peaks in a length-weighted G+C histogram like **Figure 1** are useful in surveying diversity of dominant microbial populations within a mixed community. Prominent scaffold peaks at 43, 49, 56, 60–62, 64, and 67% GC suggested that the Lake Tyrrell microbial community contains at least 6 different abundant genomic populations. This observation was confirmed by the reconstruction of one or more composite genomes from each major peak (**Table 1**), including multiple archaeal populations with similar G+C compositions within broader peaks at 47–50%, 59–61%, and 63–64% G+C, and both archaeal and bacterial populations within the 67% G+C peak.

16S rRNA diversity

Assembled sequences contained 34 distinct 16S rRNA gene sequences of 450 nt or longer, including 27 longer than 700 nt (**Table S4**). One scaffold contained a full-length 16S rRNA sequence that was 97% identical to cultured isolates of the halophilic bacterium *Salinibacter ruber*. The remaining 16S rRNA sequences were all archaeal, based on both BLAST searches against the Greengenes database and phylogenetic placement relative to characterized 16S rRNA gene sequences in a maximum-likelihood phylogenetic tree (**Figure 2**). Assembled archaeal 16S rRNA genes were distributed among seven broad phylogenetic groups, including class Nanohaloarchaea and relatives of previously sequenced isolates from Halobacterial genera *Haloquadratum*, *Halonotius*, *Halorubrum*, *Halobaculum*, *Halorhabdus*, and *Haloarcula*. Nearly all assembled 16S rRNA gene sequences had closer matches among uncharacterized environmental clones than sequenced isolate genomes.

A composite phylogenetic tree comparing archaeal 16S rRNA sequences from assembled scaffolds with the shorter, unassembled fragments (>350 nt) present in raw reads, placed >99% (1187/1202) of the unassembled read sequences into branches that were either basal, adjacent or identical to sequences represented by assembled scaffolds (**Figure S2a**). Assignment of basal positions to some of the shorter sequences present in unassembled reads reflects the unavailability of sufficient information to accurately resolve the placement of these 16S gene fragments. Several low-abundance clusters found in raw reads were not detected among the assembled scaffolds. These sequences were placed on branches adjacent to *Halovivax ruber*, *Haladaptatus paucihalophilus* and *Halobacterium salinarum*.

A similar, but less complete pattern of extended archaeal microdiversity was observed in archaeal PCR products when compared with assembled scaffold sequences (**Figure S2b**). A number of lineages present in both assembled scaffolds and raw metagenomic reads were missing from the PCR-generated 16S rRNA sequences. This result is consistent with previously described cases of universal archaeal primer bias preventing detection of novel archaeal taxa via PCR amplification [13,37].

Eighty-five percent of the sequences amplified with archaeal primers matched assembled metagenomic scaffold sequences at 97% or greater sequence identity, suggesting membership in the same species. An additional 5% of the archaeal amplicons matched assembled sequences at 95–97% identity, most likely representing different species of the same genus. Eighty eight percent of the 16S rRNA amplicons obtained using bacterial primers matched cultured isolates of *Salinibacter ruber* at 97% or higher identity, confirming the dominance of this lineage among the bacterial community that was also observed in the assembled scaffolds.

Scaffold Binning and Targeted Genome Reconstruction

Eleven distinctive scaffold clusters were identified by applying the technique of Non-Metric Multidimensional Scaling to scaffold properties used for phylogenetic binning (**Figure S3**, **Table S5**). Each cluster was subjected to targeted iterative assembly yielding twelve genomes, eleven archaeal and one bacterial (**Table 1**). Each of these genomes represents the composite sampling of multiple individuals belonging to a genomically-similar population of closely related cells (species), approximating the dominant genotype extracted from a larger, polymorphic pool of closely related variants (strains). The treemap illustration presented in **Figure 3** shows the relative abundances of these populations in the context of all assembled scaffolds, organized according to taxonomically related population groups. This figure highlights the fact that each major population group contained multiple scaffold groups that could be identified as closely related to each other, but not necessarily assigned to specific genomes.

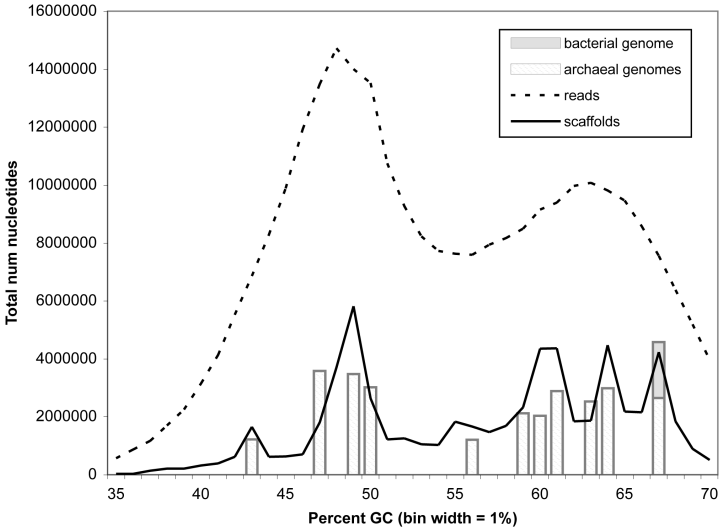


Figure 1. Length-weighted %G+C nucleotide composition of unassembled reads, assembled scaffolds, and composite population genomes. Genomes were constructed by targeted assembly of scaffolds with a uniform signature of phylogenetic binning properties, as described in Materials and Methods. Genome names, percent G+C, and other general properties of assembled genomes are shown in Table 1.
doi:10.1371/journal.pone.0061692.g001

Taxonomic Groups in Assembled Scaffolds

***Haloquadratum*-related populations J07HQW1, J07HQW2, and J07HQX50.** Microbial populations related to cultured isolates of *Haloquadratum walsbyi* comprised 38% of the assembled Lake Tyrrell community sequences. Three distinct population genomes were reconstructed, named J07HQW1, J07HQW2, and J07HQX50. Based on16S rRNA sequence identity, J07HQW1 (99%) was more closely related to *H. walsbyi* cultured isolates than J07HQW2 (97%) or J07HQX50 (93%). These relationships were confirmed by adjacency in a maximum-likelihood phylogenetic tree (Figure 2). Mean assembly depths of coverage for both J07HQW1 and J07HQW2 (8.8-fold) were more than three-fold higher than for J07HQX50 (2.5-fold), suggesting considerably greater environmental abundance (Figure S4).

Authenticity of assembled 16S rRNA gene sequences from groups J07HQW1, J07HQW2, and J07HQX50 were corroborated by the presence of identical sequences in independent PCR clone libraries, as well as near-exact matches (>99% identity) in

Table 1. Consensus population genome properties.

Genome name	Length (nt)	G+C pct	num scf	rRNA operons	tRNAs	predicted CDS	pct complete marker genes§
Haloquadratum walsbyi str J07HQW1	3,594,539	47	1	2	47	3,584	100
Haloquadratum walsbyi str J07HQW2	3,475,501	49	1	2	52	3,856	98
Haloquadratum sp. J07HQX50	3,019,909	50	2	1(2)*	39	2,872	91
Nanosalinarum sp. J07AB56	1,215,802	56	3	1	38	1,454	100
Nanosalina sp. J07AB43	1,227,157	43	7	1	59	1,739	83
Halonotius sp. J07HN4	2,888,659	61	2	1	52	3,230	100
Halonotius sp. J07HN6	2,529,000	63	6	1	47	2,914	100
uncultured archaeon sp. J07HX64	2,982,938	64	1	1	43	3,095	92
uncultured archaeon sp. J07HX5	2,040,945	60	1	1(2)*	24	2,139	53
Halobaculum sp. J07HB67	2,649,547	67	3	1	37	2,707	94
Halorubrum sp. J07HR59	2,120,805	59	7	1(3)*	26	1,841	83
Salinibacter sp. J07SB67	1,931,021	67	443	nd	13	1,641	39

§Marker gene detection details are shown in Table S6.
*Parenthetical values indicate cases where locally elevated depth of coverage suggests that assembly software may have compressed multiple 16S gene copies into a single locus.
doi:10.1371/journal.pone.0061692.t001

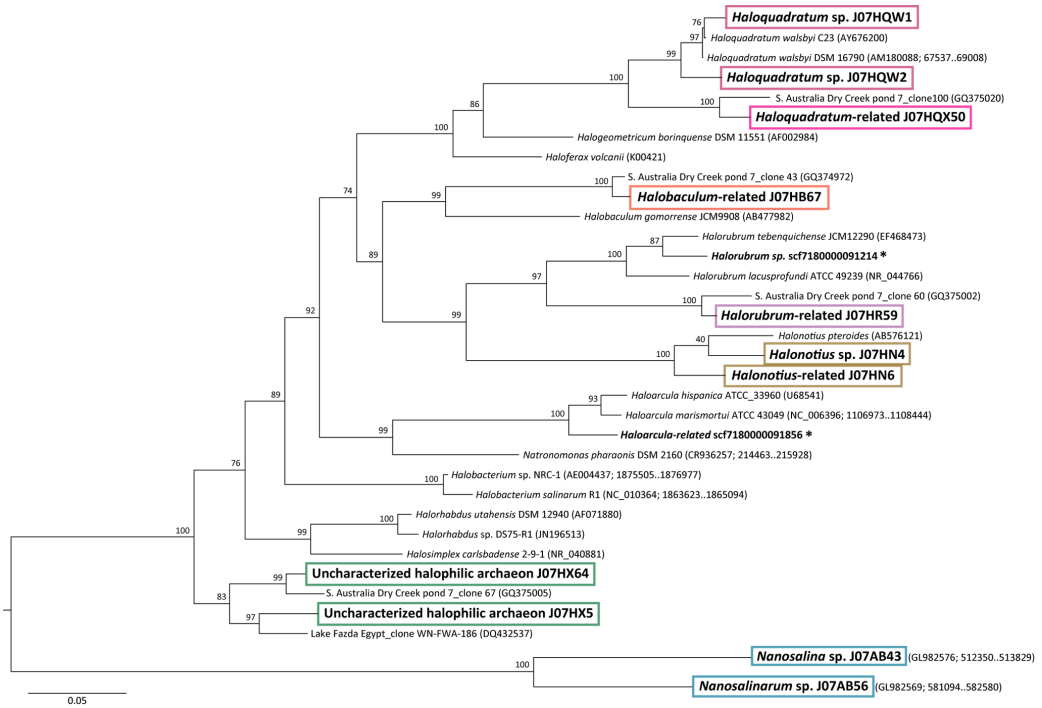


Figure 2. Phylogenetic distribution of archaeal 16S rRNA gene sequences in assembled scaffolds and population genomes. Names in bold indicate new 16S rRNA sequences identified in this study. Boxed names indicate sequences contained within Lake Tyrrell-specific population genomes. Asterisks indicate isolated individual sequences found on small scaffolds that were not associated with any assembled population genome.
doi:10.1371/journal.pone.0061692.g002

16S rRNA sequences amplified by other investigators studying a different Australian hypersaline habitat [38]. In that study, sequences most closely matching J07HQX50 (phylogroup 2) were suggested to represent a separate genus from *H. walsbyi* strains C23 and DSM 16790. BLASTP analysis of predicted proteins in all three *Haloquadratum*-related genomes against Genbank nr reinforced taxonomic the relationships observed with 16S rRNA genes (**Figure 4**). The J07HQW1, J07HQW2, and J07HQX50 genomes all included a significant number of core gene matches to *H. walsbyi* cultured isolates. However, the overall percentage of predicted proteins with best matches to previously sequenced *Haloquadratum* genomes was less than half in J07HQX50 (28%) compared to J07HQW1 (58%), consistent with evolutionary diversification as a separate genus.

For populations like J07HQX50, where no physical data is available, distribution of scaffold reads between libraries obtained from 0.1 versus 0.8 µm filters can be used to obtain a rough estimate of cell size. Although it is not possible to determine exact cell size from read library distributions,high and low ends of the microbial size range sampled in Lake Tyrrell can be bracketed based on microscopically observed diameters of approximately 2 µm for the square cells of cultured *Haloquadratum* isolates (80% on 0.8 µm filters) and 0.6 µm for coccus-shaped environmental Nanohaloarchaea (<10% on 0.8 µm filters) [13].

Eighty percent of reads from scaffolds in all three *Haloquadratum*-related genomes were isolated from 0.8 µm pore filters, making them the largest cells in the current study. Finding 20% of the reads on 0.1 µm pore filters was initially unanticipated, based on the diameter of cultured *Haloquadratum* isolates and their known propensity to form multicellular aggregates. However, cultured *Haloquadratum* cells contain especially fragile internal gas vesicles, susceptible to collapse under pressures experienced during cellular concentration by filtration [39,40]. In addition, nominal pore sizes reported for fiber-based filters are average values for a non-uniform size distribution that covers a wider range, explaining why some cells, especially those with flexible and/or asymmetric shapes, can routinely pass through filters with smaller than expected pore sizes.

Nanohaloarchaea populations J07AB43 and J07AB56. Sequences from archaeal class Nanohaloarchaea accounted for approximately 17% of the assembled microbial community, forming the second most abundant microbial group. Taxonomic binning of scaffolds from this group was facilitated by their significant divergence from other microbial groups in nucleotide G+C compositions, 16S rRNA gene sequences, predicted amino acid frequencies, and filter size distribution of reads [13]. Finding greater than 90% of the J07AB43 and J07AB56 reads in 0.1 µm pore filters agrees with previously reported cell diameters of approximately 0.6 microns, and suggests that they are the smallest cells whose genomes were assembled from the Lake Tyrrell environmental sequences.

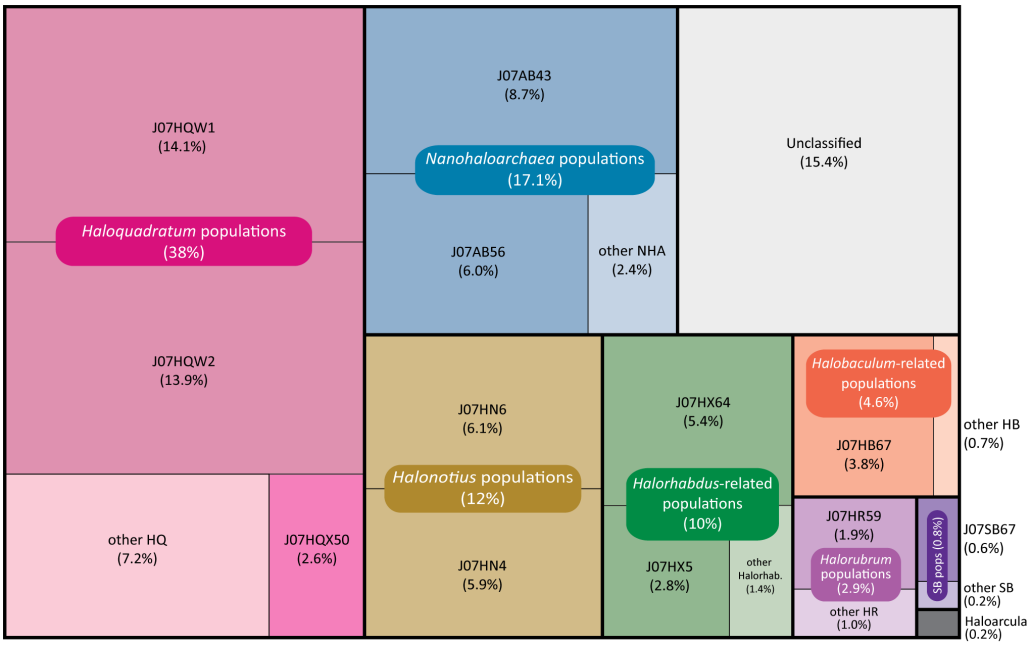


Figure 3. Relative abundance of microbial population groups. Colors correspond to taxonmically related microbial populations, including both assembled genome sequences and non-genomic scaffolds containing less abundant variant sequences. Percentage calculations include total number of assembled nucleotides in reads associated with each group, normalized for the group's average genome size. Percentage of unclassified sequences was calculated using an estimated genome size of 3 MB, the approximate abundance-weighted average for all other groups. Known viral and plasmid sequences, representing approximately 0.2% of assembled nucleotides, have been excluded from these calculations.
doi:10.1371/journal.pone.0061692.g003

***Halonotius*-related populations J07HN4 and J07HN6.** The next most abundant population group, comprising approximately 12% of the community, contained two population genomes, J07HN4 and J07HN6. 16S rRNA gene sequences from these populations were 95–97% identical to *Halonotius pterides*, a cultured isolate for which no genome sequence is currently available [38,41]. Despite differences in nucleotide composition between the two Lake Tyrrell *Halonotius*-like populations (63% versus 61% G+C), both shared similar amino acid composition profiles and taxonomic distributions of database matches for predicted proteins (Figure 4).

Based on scaffold read library distribution between 0.1 and 0.8 µm size fractions, *Halonotius*-like populations have the next smallest cells after Nanohaloarchaea in the Lake Tyrrell community. The percentage of 0.1 µm filter reads in J07HN6 (80%) was much higher than J07HN4 (50%) suggesting smaller cellular diameter in J07HN6. Although *Halonotius* cells have not been observed to undergo significant aggregation in culture, no data is currently available on whether this behavior might occur under natural conditions. Neither of the *Halonotius*-related genomes contain gas vesicle protein (gvp) synthesis genes, but both contain flagellar synthesis genes. Small flagellated cells and the absence of gas vesicles are consistent with light and electron micrograph observations of *H. pteroides* isolates in culture, which have cell diameters ranging between 0.7–1.5 µm and variable morphophologies including cocci, elongated rods and airfoil-like shapes [41].

***Halorhabdus*-related populations J07HX64 and J07HX5.** Approximately 10% of assembled scaffold sequences formed a group most closely related to the genus *Halorhabdus*. The J07HX5 and J07HX64 genomes differed by 4% G+C, with 16S rRNA genes that were 96% identical to each other. J07HX64 matched an environmental 16S rRNA gene cloned from an Australian salt crystallizer (GQ375005) at 98% identity [38]. The closest environmental match to J07HX5 was to a 16S rRNA gene cloned from an Egyptian hypersaline lake (DQ432537), at 96% identity [42].

Predicted proteins from J07HX5 and J07HX64 shared similar amino acid composition signatures (**Table S5**) and similar taxonomic patterns of reference database BLASTP matches (**Figure 4**). *Halorhabdus* was the single most frequently matched genus at 15%, although several other Haloarchaeal genera matched at frequencies of 5–8%. Percentages of 0.1 µm pore filter reads comprising the J07HX5 (21%) and J07HX64 (24%) genome scaffolds suggest an effective cell size similar to *Haloquadratum*.

***Halobaculum*-related population J07HB67.** Approximately 5% of the assembled Lake Tyrrell sequences were associated with a scaffold group named J07HB67. These scaffolds contain a 16S rRNA gene matching the genome of cultured isolate *Halobaculum gomorrense* at 92% identity. The J07HB67 16S rRNA gene is 99% identical to Australian salt crystallizer environmental clone GQ374998 (phylogroup 7) [38]. Approximately 33% of reads associated with J07HB67 populations were isolated from 0.1 µm pore filters, suggesting that cells from this population are larger than those of the *Halonotius* group, but smaller than *Haloquadratum*,

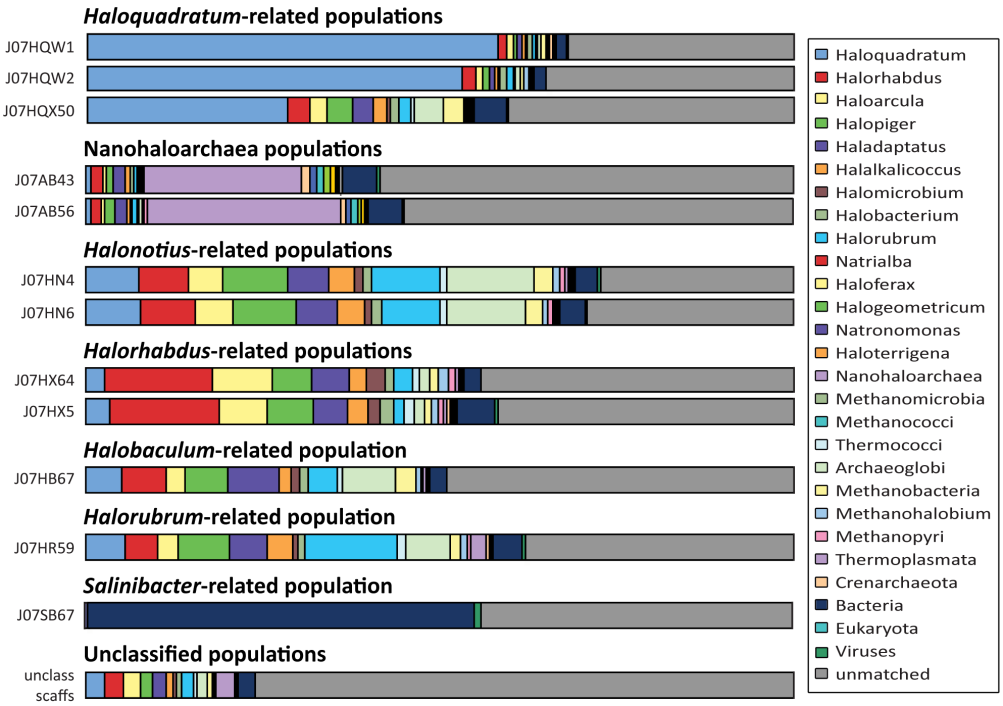


Figure 4. Phylogenetic distribution of protein BLAST matches for assembled population genomes and unclassified scaffolds. Taxonomic distribution of non-self matches versus the Genbank nr database were calculated using the DarkHorse algorithm at a filter threshold setting of 0.05, including only alignments covering at least 70% of both query and target sequences with an e-value of 1e-5 or better.
doi:10.1371/journal.pone.0061692.g004

J07HX5 and J07HX64. This finding is consistent with microscopic observations of *H. gomorrense*, whose rod-shaped cells measure 0.5–1 μm wide by 5–10 μm long [43].

***Halorubrum*-related populations.** Assembled scaffolds from at least two *Halorubrum*-related populations, representing approximately 3% of the Lake Tyrrell microbial community, were linked by a common pattern of filter size distribution, percent G+C, amino acid sequence composition, and taxonomic classification of BLASTP hits against GenBank nr, in which *Halorubrum* was the most frequently matched genus. Two different *Halorubrum*-related 16S rRNA sequences were observed in assembled scaffolds, 90% identical to each other. Only one of these scaffold groups (J07HR59), representing approximately 2% of the assembled microbial community, was sufficiently abundant for population genome assembly. The J07HR59 16S rRNA sequence matched an environmental clone (GQ374972) described as Halorubrum-related phylogroup 4 at 97.4% identity [38], but J07HR59 and GQ374972 form a separate, independent branch from previously cultured isolate *Halorubrum* genomes (Figure 2). The other *Halorubrum*-related Lake Tyrrell population, representing approximately 1.0% of the assembled community, cladded with previously cultured isolates, matching the *Halorubrum tebenquichense* 16S rRNA gene at 96% identity.

***Haloarcula* and other low abundance archaeal populations.** Several small scaffolds containing solely archaeal 16S rRNA gene sequences were identified from populations with minimal genomic sampling (Table S4). These included two 16S rRNA sequences similar to cultured isolates of genus *Haloarcula*, at 3–4X depth of coverage. However, other scaffolds identifiable as *Haloarcula*-related were assembled at a coverage of 1.2 fold or less. *Haloarcula*-related 16S rRNA genes may have been more completely assembled than other loci from the same population due to multiple co-assembling gene copies; sequenced *Haloarcula* isolate genomes typically contain three 16S rRNA copies. Based on an estimated genome size of 3.9 Mbp, *Haloarcula*-related populations comprised approximately 0.2% of the assembled community, consistent with the lower depth of coverage of non-16S rRNA containing scaffolds.

***Salinibacter* population J07SB67.** The only bacterial 16S rRNA sequence obtained from Lake Tyrrell metagenomic assembly matched cultured isolates of *Salinibacter ruber* at 98% identity, consistent with the observation that 3,480/3,958 (88%) of 16S rRNA sequences independently amplified using universal bacterial PCR primers matched cultured *Salinibacter* at 97% or higher identity. The assembled *Salinibacter* 16S rRNA gene was located on a small, 2,795 nucleotide scaffold, adjacent to a single predicted hypothetical protein. However, more than 400 additional scaffolds, ranging in size from 1,000–19,000 nucleotides, shared patterns of BLAST match taxonomy, nucleotide composition, and predicted amino acid composition consistent with assignment to a *Salinibacter*-related species.

Targeted assembly of the *Salinibacter*-related scaffold group yielded an incomplete genome of only 1.2 Mbp, versus 3.6 MB for previously sequenced *Salinibacter* isolates (33.3% genome coverage) [44]. Thirty-nine percent of highly conserved bacterial core proteins present in both cultured *Salinibacter* isolate genomes were recovered, consistent with total genome length. Depth of coverage for *Salinibacter*-related scaffolds averaged 1.5 fold, corresponding to a nucleotide abundance of approximately 0.6 percent of the microbial community.

Viral and "Plasmidome" community sampling. Despite the use of sample preparation methods designed to capture only cells between 0.1 and 3 μm in diameter, a group of 142 small scaffolds, representing approximately 0.2% of assembled nucleotides, contained DNA fragments that appear viral in origin. These fragments ranged in size from 1,000 to 25,000 nucleotides in length, with compositions varying between 35–71% G+C. Most of these putative viral scaffolds were reconstructed exclusively from 0.1 μm filter reads. These results are consistent with non-specific retention of viral particles on filter surfaces and/or recovery of phage genomes from infected cells during sample preparation. Predicted proteins in these scaffolds included BLAST matches to viral groups previously shown to be abundant in hypersaline waters, including BJ1-like Siphoviridae and PhiCh-like Myoviridae [45,46,47,48,49,50]. Recovered data were insufficient to determine whether or not these sequences were integrated as prophage in microbial genomes.

Forty scaffolds ranging in size from 1–50 kbp, comprising approximately 0.2% of assembled nucleotides, contained genes encoding p4 plasmid primase, suggesting that they may be archaeal plasmid sequences. Two additional scaffolds contained matches to the *Salinibacter ruber* plasmid protein init Rep_3. Nucleotide composition of putative plasmid scaffolds ranged from 49–66% G+C, at 1.1–12.8 fold depth of coverage, suggesting association with both dominant and rare community members. However, most putative plasmid scaffolds could not be confidently assigned to a specific host organism, and contained few predicted proteins similar to previously sequenced database representatives. Plasmid numbers in cultured halophilic Archaea and Bacteria vary between zero (e.g *Haloquadratum walsbyi* DSM 16790) and seven (e.g. *Haloarcula marismortui* ATCC 43049), with sizes ranging from <2 Kbp (*Halobacterium salinarum*, NC_002121) to >600 Kbp (*Haloferax volcanii* DS2, NC_013966). This extremely wide variability makes it difficult to determine the extent to which the plasmid scaffolds we observed represent partial versus complete sequences.

Unclassified Sequences. Approximately 15% of assembled scaffold sequences could not be confidently assigned to any of the groups described above. The low assembly coverage and short sequence lengths in these scaffolds most likely encompass not only less abundant members of the community, but also partial, incomplete fragments corresponding to polymorphic insertions, deletions, mutations, and rearrangements between related strains. Seventy-six percent of predicted protein sequences in the unclassified scaffold group failed to match any sequences in Genbank nr. Database matches were predominately archaeal in origin, including the same reference organisms as assembled consensus population genomes (Figure 4).

To estimate the extent to which unclassified scaffolds might represent uncaptured functional diversity within the community, all predicted proteins from the original composite assembly, including both classified and unclassified sequences, were screened for matches to PFAM, COG, and KEGG protein database patterns. At least one pattern was found in 31,696 of 62,918 predicted proteins. Even though unclassified scaffolds comprise 15% of total assembled nucleotides, they contained only 326 patterns absent from the classified data set, corresponding to 7.5% of the overall pool. Classified scaffolds contained 92.5% of all protein patterns detected (5,197 proteins). Protein domain patterns unique to the unclassified scaffolds included a large number of viral-related functional elements, as well as low complexity short repeats characteristic of incomplete protein fragments, suggesting that this group contains an over-representation of partial genes and viral fragments.

To eliminate potential bias due to the highly conserved nature of COG, KEGG, and PFAM patterns, unsupervised Markov algorithm clustering was also performed on all 62,918 predicted proteins in the initial combined assembly. Based on frequencies of these unsupervised clusters, Chao and Ace estimators indicate that assembled scaffolds include greater than >90% of the expected functional diversity in the sampled community. Classified scaffolds contained 4,432 of the 5,242 clusters observed, with only 810 clusters occurring uniquely in the unlassified scaffold set. Close agreement between the percentage of protein clusters (84.6%) and total nucleotides incorporated in assembled scaffolds (84.5%) supports use of the classified data set as a representative sample of functional diversity within the community.

Population Distribution of Community Functions

Markov algorithm clustering was applied to all 31,062 predicted proteins from the twelve Lake Tyrrell genomes, generating 6,591 protein families. Protein family clusters shared between different populations were plotted as connections in a network representation (**Figure 5**). Highly interconnected clusters, converging at the center of the diagram, include both universal housekeeping genes and habitat-specific adaptive capabilities. Functions broadly shared among all taxonomic groups suggest a common aerobic, heterotrophic lifestyle. Protein families conserved in all 11 archaeal populations also include UV damage repair endonucleases, peroxiredoxins and thioredoxins, halocyanins, Ca^{2+}/Na^{+} antiporters, and type IS605 OrfB family transposases.

Population-specific protein families located at the periphery of **Figure 5** capture functional novelty of both individual genomes and closely related taxonomic groups relative to the rest of the community. **Table 2** compares numbers of unique clusters found in each of the twelve consensus genomes. The population with the greatest number of unshared protein families is *Salinibacter*, the only bacterium in the group, even though the assembled genome was less than 40% complete. The two Nanohaloarchaeal genomes J07AB43 and J07AB56 also contained many unique clusters, both individually and shared between them.

Although each of the three *Haloquadratum* genomes had more than 350 unique clusters, these numbers were similar to other Lake Tyrrell Haloarchaeal populations when normalized for genome size. Numbers of novel clusters found in *Haloquadratum*-related populations suggest more diverse protein functions than other community members, but most likely also include a higher number of pseudogenes, as observed previously in *Haloquadratum* isolate genomes [51,52]. In contrast to *Haloquadratum*, *Halonotius* populations J07HN4 and J07HN6 contain surprisingly few novel protein clusters in each individual genome, despite 16S rRNA sequences that are more divergent from each other than the J07HQW1 and J07HQW2 genomes. Unique functional properties of the *Halonotius* group are captured instead in clusters shared between J07HN4 and J07HN6.

Many protein families shared between different community members contain only sequences whose function cannot be predicted from bioinformatic inference. Proteins of unknown function are more abundant among population-specific protein

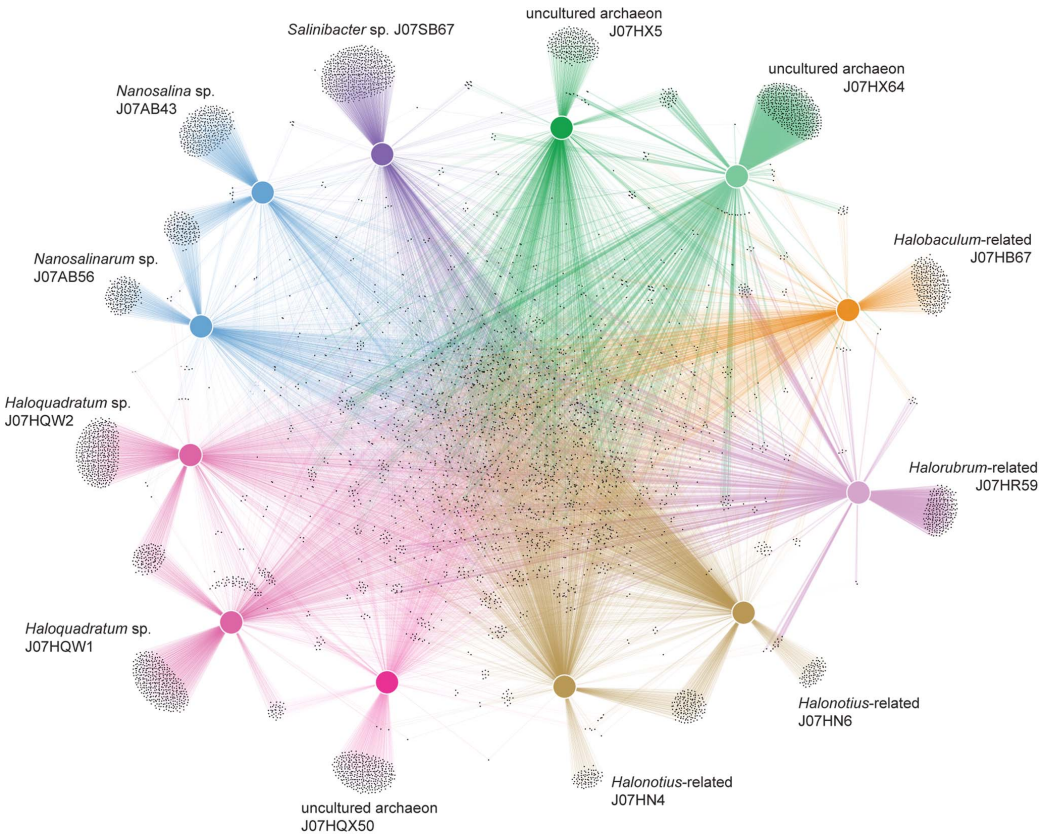


Figure 5. Metabolic connectivity graph showing community distribution of protein family clusters. Cohesive populations are shown as similarly colored nodes and vectors according to numbers of shared features, based on unsupervised protein family clustering of 12 habitat-specific genomes.
doi:10.1371/journal.pone.0061692.g005

families than in more widely distributed clusters. However, even confidently annotated proteins sometimes generate multiple clusters with similar descriptions, and may correspond to protein isoform variants with unknown but possibly significant differences in biological activity.

Discussion

In this study we have captured the taxonomic diversity, population abundance, and functional properties associated with both broad phylogenetic groups and individual microbial populations in a mixed, natural ecosystem community. Reconstruction of 12 habitat-specific population genomes from a single pool of metagenomic sequencing reads demonstrates the value of combining *de novo* assembly with iterative, multi-dimensional phylogenetic binning. This approach proved particularly useful in characterizing previously undescribed novel organisms, avoiding problematic issues of amplification primer bias and variable 16S rRNA gene copy number in divergent populations. Eight reconstructed genomes represented species with no previously cultured isolates, including populations comprising 2–14% of the sampled microbial community. Ten of the twelve genomes were nearly complete, in assemblies of seven or fewer scaffolds.

Each of these population genomes represents a composite sequence constructed from multiple, closely related individual cells, providing a set of core gene models and operon structures common to most members of the population. These genomes do not include peripheral pan-genomic content that is unique to individual strains. Regions of significant population divergence (intra-species heterogeneity) are incorporated as gaps in larger scaffolds and/or separate shorter overlapping scaffolds with lower read coverage. The composite sequences we have obtained by community metagenomic assembly cannot be expected to furnish the same level of detail and accuracy as the closed, finished genome of an individual isolate, yet their ability to deliver full length genes in cellular context has provided important new insights into community structure, novel taxa, and compartmentalized protein functional associations that could not be obtained from unassembled reads alone.

Although Sanger technology was the primary source of reads for this study, the subtractive taxonomic enrichment strategies we

Table 2. Population-unique protein family clusters.

Genome name	num unique clusters	total num genome clusters	pct. unique clusters
Salinibacter sp. J07SB67	581	1,639	35%
Nanosalina sp. J07AB43	366	1,678	22%
uncultured archaeon sp. J07HX64	441	3,047	14%
Nanosalinarum sp. J07AB56	184	1,410	13%
Halorubrum sp. J07HR59	232	1,839	13%
Haloquadratum sp. J07HQX50	351	2,872	12%
uncultured archaea sp. J07HX5	258	2,139	12%
Haloquadratum walsbyi str. J07HQW1	403	3,584	11%
Haloquadratum walsbyi str. J07HQW2	433	3,855	11%
Halobaculum sp. J07HB67	296	2,846	10%
Halonotius sp. J07HN6	90	2,913	3%
Halonotius sp. J07HN4	81	3,229	3%
total	**3,716**	**31,051**	**12%**

doi:10.1371/journal.pone.0061692.t002

have developed could also be applied to metagenomic assemblies using paired-end reads obtained by more contemporary platforms such as Illumina. Our *de novo* assembly procedures were especially effective in faciliting genome recovery for populations (species) with no closely related sequenced relatives. Assembly quality was improved as data complexity was reduced and the accuracy of read binning enhanced by iterative, scaffold-based read selection using multiple, independent parameters. These parameters included uniform nucleotide composition, depth of coverage, taxonomic distribution of BLASTP database matches, and amino acid composition of predicted proteins. Read distribution frequencies from overlapping libraries obtained using different filter pore sizes provided an additional source of independent information to help distinguish difficult-to-separate groups and verify assembly fidelity, as well as offering a novel opportunity to estimate physical cell size of uncharacterized organisms relative to other members of the community.

Archaea greatly outnumbered Bacteria in the Lake Tyrrell hypersaline ecosystem, as previously reported for other extreme hypersaline environments [14,53]. Although relatives of *Haloquadratum walsbyi* were the most abundant taxonomic group, comprising approximately 38% of the community, nearly 47% of the assembled sequences were derived from a combination of Nanohaloarchaea (17%) and relatives of the Haloarchaeal genera *Halonotius* (12%), *Halorhabdus* (10%), *Halobaculum* (4.6%), *Halorubrum* (2.9%), and *Haloarcula* (0.2%). Based on historical accounts of other hypersaline habitats [52,54,55], diversity within the *Haloquadratum*-related population was higher than expected, including at least three different species from two different genera.

The 62,918 environmental genes recovered from the assembled metagenomic sequences were estimated to encompass more than 90% of the functional diversity present in the community. The construction of multiple habitat-specific Lake Tyrrell population genomes has enabled genome-wide assignment of functional activities to specific individual organisms of known abundance in the community. These assignments provide new opportunities to begin comparing shared and novel protein families across related and divergent co-occurring populations adapted to the same environmental conditions with a level of organism-specific context that would not be possible with unassembled reads alone.

The relatively constrained metabolic repertoire of broadly shared protein functional families in the Lake Tyrrell community may be linked to physicochemical uniformity in the shallow, aquatic hypersaline environment from which organisms were sampled. The common evolutionary history of halophilic Archaea adapted to extreme salinity may also play a role. It has been speculated that abundances of different microbial populations under these conditions might be driven more by top down forcing dynamics, for example protozoan predation and/or viral infections, rather than nutrient availability [56]. The current study does not include seasonal fluctuations in temperature, salinity and nutrient inputs, which might reveal greater diversity over longer time scales. The availability of new habitat-specific reference genomes from the Lake Tyrrell ecosystem provides new reference data to track these populations over time and space at the level of both genes and genomes.

Functional genes and metabolic processes unique to individual populations may also provide information useful in designing cultivation methods for previously uncultured organisms, including the possibility of mixed co-cultures to accommodate natural symbiotic or co-dependent trophic relationships. The potential utility of this approach is illustrated by the observation that strains of *Haloquadratum walsbyi*, notoriously difficult to grow in isolate culture, form significantly larger colonies in the presence of *Salinibacter ruber* [57]. Although *Salinibacter*-related populations comprise only a small percentage of the ecosystem described here, *Haloquadratum* abundance could be driven by similar nutritional complementation provided by alternative members of the community.

The new genomes described in this study expand opportunities to identify novel phylogroups in other environments, providing new templates for fragment recruitment and assembly, as well as group-specific probes for *in situ* quantitation. Organisms previously identified by 16S rRNA gene sequences alone can now be prioritized as targets for more detailed investigations based on functional, as well as taxonomic information. Furthermore, the assembly of habitat-specific genomes provides an important foundation to decipher genotype-phenotype relationships based on metatranscriptomic and metaproteomic investigations in similar environments. The simultaneous interrogation and synthesis of composite data from multiple microbial populations in

natural ecosystems will provide the comprehensive level of genotypic and phenotypic data necessary to model synergistic activities of community members, while contributing to an enhanced understanding of the ecology and evolution of environmental microbial species.

Supporting Information

Table S1 Water chemistry of Lake Tyrrell sampling site. Located at 35°19′12.24S 142°48′00.45E.
(PDF)

Table S2 Summary of metagenomic sequencing libraries used in this study. Average read length is shown ± standard deviation.
(PDF)

Table S3 Assembly statistics for combined Sanger metagenomic libraries using Celera Assembler version 5.4. Assembly parameters used were as follows: utgErrorRate = 0.10; ovlErrorRate = 0.10; cnsErrorRate = 0.10; cgwErrorRate = 0.12; utgBubblePopping = 0; utgGenomeSize = 500000; merSize = 15; doFragmentCorrection = 0; doExtendClearRanges = 1; doResolveSurrogates = 1; Unitigger parameter –j = −20.
(PDF)

Table S4 Assembled 16S rRNA sequences and their closest database matches to environmental clones and cultured isolates. Matches were required to have BLAST alignments to previously identified 16S rRNA genes of 450 nt or longer, with e-value <1e-7 and 80% or greater sequence identity between query and subject. Part A shows 16S rRNA gene sequences obtained in targeted genomic assemblies. Part B shows additional 16S rRNA gene sequences observed in scaffolds obtained by composite assembly of all Sanger reads.
(PDF)

Table S5 Distinctive properties of major scaffold clusters. Percentages are based on taxonomic classifications of all predicted protein tophit matches to Genbank nr, as determined using the DarkHorse algorithm at a filter threshold setting of 0.05, including only alignments covering at least 70% of both query and target sequences with an e-value of 1e-5 or better.
(PDF)

Table S6 Estimated genome completeness. Based on presence/absence of 53 conserved genes in assembled archaeal composite population genomes.
(PDF)

Figure S1 Bioinformatic Analysis Pipeline.
(PDF)

Figure S2 Phylogenetic trees showing abundance of clustered archaeal 16S rRNA sequences from (A) unassembled reads and (B) PCR-amplified clone libraries. A maximum likelihood archaeal reference tree was constructed using FastTree [1], based on full-length 16S genes from isolate genomes and environmental clones from Genbank nt, supplemented with sequences obtained from Lake Tyrrell assembled scaffolds (highlighted in yellow). Additional partial 16S rRNA sequences from Lake Tyrrell were inserted into the reference tree using pplacer version v1.1 (model GTR, fig-eval-all) [2] and visualized using Archaeopteryx 0.968 [3]. Part A shows placement of unamplified raw metagenomic reads containing 16S gene sequences. Part B shows placement of PCR-amplified 16S rRNA clones. Numbers at nodes indicate confidence values estimated by FastTree for the reference tree. Red lines indicate branches where Lake Tyrrell sequences were observed. The thickness of each red line is proportional to the number of Lake Tyrrell sequences associated with that branch, ranging from one in the thinnest line to 74 in the thickest line.
(PDF)

Figure S3 Non-metric multidimensional scaling plot illustrating distinctive scaffold groups. Scaffolds >5 Kb from the composite Sanger assembly were subjected to non-metric multidimensional scaling analysis using Primer 1.6, with Euclidean distance, 25 random starts, Krustal fit scheme 1, and minimum stress value 0.01 for the 13 parameters shown in Table S5. Axes shown are arbitrary units of composite clustering, although the X axis appears to be dominated by nucleotide percent G+C. Scaffolds associated with major taxonomic groups are highlighted with colored symbols. Small grey dots indicate scaffolds that could not be unambiguously classified into major groups.
(PDF)

Figure S4 Rank abundance of assembled microbial populations based on depth of coverage.
(PDF)

Acknowledgments

We thank Sue Welch and Dawn Cardace for sample collection assistance at Lake Tyrrell; Mike Dyall-Smith for generous access to reagents and laboratory equipment; Cheetham Salt Works (Lake Tyrrell, Australia) for permission to collect samples; Matt Lewis and the J Craig Venter Institute for library construction and sequencing; and the US Department of Energy Joint Genomes Institute for genome annotation support via the Integrated Microbial Genome Expert Review (IMG-ER) resource.

Author Contributions

Conceived and designed the experiments: SP JAU JFB KBH EEA. Performed the experiments: SP JAU PN JFB KBH EEA. Analyzed the data: SP JAU PN EEA. Contributed reagents/materials/analysis tools: SP JAU EEA. Wrote the paper: SP JAU EEA. Designed the software used in analysis: SP JAU.

References

1. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37–43.
2. Allen EE, Banfield JF (2005) Community genomics in microbial ecology and evolution. Nat Rev Microbiol 3: 489–498.
3. Garcia Martin H, Ivanova N, Kunin V, Warnecke F, Barry KW, et al. (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nat Biotechnol 24: 1263–1269.
4. Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, et al. (2014\1) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature 480: 368–371.
5. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, et al. (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science 335: 587–590.
6. Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, et al. (2012) Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res.
7. Eisen JA (2007) Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. PLoS Biol 5: e82.
8. Holler T, Widdel F, Knittel K, Amann R, Kellermann MY, et al. (2011) Thermophilic anaerobic oxidation of methane by marine microbial consortia. Isme J 5: 1946–1956.
9. Brogden KA, Guthmiller JM, Taylor CE (2005) Human polymicrobial infections. Lancet 365: 253–255.
10. Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, et al. (2010) A catalog of reference genomes from the human microbiome. Science 328: 994–999.

11. Andrei AS, Banciu HL, Oren A (2012) Living with salt: metabolic and phylogenetic diversity of archaea inhabiting saline ecosystems. FEMS Microbiol Lett 330: 1–9.
12. Oren A (2002) Halophilic microorganisms and their environments. Dordrecht; Boston: Kluwer Academic. xxi, 575 p.
13. Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, et al. (2012) De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. Isme J 6: 81–93.
14. Ghai R, Pasic L, Fernandez AB, Martin-Cuadrado AB, Mizuno CM, et al. (2011) New abundant microbial groups in aquatic hypersaline environments. Sci Rep 1: 135.
15. Goldberg SM, Johnson J, Busam D, Feldblyum T, Ferriera S, et al. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci U S A 103: 11240–11245.
16. Bik EM, Long CD, Armitage GC, Loomer P, Emerson J, et al. (2010) Bacterial diversity in the oral cavity of 10 healthy individuals. Isme J 4: 962–974.
17. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, et al. (2000) A whole-genome assembly of Drosophila. Science 287: 2196–2204.
18. Brady A, Salzberg S (2011) PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nat Methods 8: 367.
19. Schatz MC, Phillippy AM, Shneiderman B, Salzberg SL (2007) Hawkeye: an interactive visual analytics tool for genome assemblies. Genome Biol 8: R34.
20. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311: 1283–1287.
21. Wu M, Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. Genome Biol 9: R151.
22. Puigbo P, Wolf YI, Koonin EV (2009) Search for a 'Tree of Life' in the thicket of the phylogenetic forest. J Biol 8: 59.
23. Mitreva M (2009) NIH Human Microbiome Project Data Analysis and Coordination Center. http://wwwhmpdaccorg/tools_protocols/tools_protocolsphp.
24. Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, et al. (2009) IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics 25: 2271–2278.
25. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, et al. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 72: 5069–5072.
26. Podell S, Gaasterland T (2007) DarkHorse: a method for genome-wide prediction of horizontal gene transfer. Genome Biol 8: R16.
27. Clarke K, Gorley R (2006) Primer v6: User Manual/Tutorial. Plymouth, UK: PRIMER-E.
28. Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A 102: 2567–2572.
29. RDevelopmentCoreTeam (2008) A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.
30. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.
31. Price MN, Dehal PS, Arkin AP (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 5: e9490.
32. Matsen FA, Kodner RB, Armbrust EV (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 11: 538.
33. Han MV, Zmasek CM (2009) phyloXML: XML for evolutionary biology and comparative genomics. BMC Bioinformatics 10: 356.
34. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575–1584.
35. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75: 7537–7541.
36. Bastian M, Heymann S, Gephi MJ (2009) An open source software for exploring and manipulating networks.
37. Casanueva A, Galada N, Baker GC, Grant WD, Heaphy S, et al. (2008) Nanoarchaeal 16S rRNA gene sequences are widely dispersed in hyperthermophilic and mesophilic halophilic environments. Extremophiles 12: 651–656.
38. Oh D, Porter K, Russ B, Burns D, Dyall-Smith M (2010) Diversity of Haloquadratum and other haloarchaea in three, geographically distant, Australian saltern crystallizer ponds. Extremophiles 14: 161–169.
39. Walsby AE (1994) Gas vesicles. Microbiol Rev 58: 94–144.
40. Kashyap S, Sundararajan A, Ju LK (1998) Flotation characteristics of cyanobacterium Anabaena flos-aquae for gas vesicle production. Biotechnol Bioeng 60: 636–641.
41. Burns DG, Janssen PH, Itoh T, Kamekura M, Echigo A, et al. (2010) Halonotius pteroides gen. nov., sp. nov., an extremely halophilic archaeon recovered from a saltern crystallizer. Int J Syst Evol Microbiol 60: 1196–1199.
42. Mesbah NM, Abou-El-Ela SH, Wiegel J (2007) Novel and unexpected prokaryotic diversity in water and sediments of the alkaline, hypersaline lakes of the Wadi An Natrun, Egypt. Microb Ecol 54: 598–617.
43. Oren A, Gurevich P, Gemmell RT, Teske A (1995) Halobaculum gomorrense gen. nov., sp. nov., a novel extremely halophilic archaeon from the Dead Sea. Int J Syst Bacteriol 45: 747–754.
44. Pena A, Teeling H, Huerta-Cepas J, Santos F, Yarza P, et al. (2010) Fine-scale evolution: genomic, phenotypic and ecological differentiation in two coexisting Salinibacter ruber strains. Isme J.
45. Garcia-Heredia I, Martin-Cuadrado AB, Mojica FJ, Santos F, Mira A, et al. (2012) Reconstructing viral genomes from the environment using fosmid clones: the case of haloviruses. PLoS One 7: e33802.
46. Atanasova NS, Roine E, Oren A, Bamford DH, Oksanen HM (2011) Global network of specific virus-host interactions in hypersaline environments. Environ Microbiol.
47. Bettarel Y, Bouvier T, Bouvier C, Carre C, Desnues A, et al. (2011) Ecological traits of planktonic viruses and prokaryotes along a full-salinity gradient. FEMS Microbiol Ecol 76: 360–372.
48. Santos F, Yarza P, Parro V, Briones C, Anton J (2010) The metavirome of a hypersaline environment. Environ Microbiol 12: 2965–2976.
49. Sime-Ngando T, Lucas S, Robin A, Tucker KP, Colombet J, et al. (2010) Diversity of virus-host systems in hypersaline Lake Retba, Senegal. Environ Microbiol.
50. Emerson JB, Thomas BC, Andrade K, Allen EE, Heidelberg KB, et al. (2012) Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. Appl Environ Microbiol 78: 6309–6320.
51. Bolhuis H, Palm P, Wende A, Falb M, Rampp M, et al. (2006) The genome of the square archaeon Haloquadratum walsbyi : life at the limits of water activity. BMC Genomics 7: 169.
52. Dyall-Smith ML, Pfeiffer F, Klee K, Palm P, Gross K, et al. (2011) Haloquadratum walsbyi: limited diversity in a global pond. PLoS One 6: e20968.
53. Oren A (2008) Microbial life at high salt concentrations: phylogenetic and metabolic diversity. Saline Systems 4: 2.
54. Legault BA, Lopez-Lopez A, Alba-Casado JC, Doolittle WF, Bolhuis H, et al. (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.
55. Cuadros-Orellana S, Martin-Cuadrado AB, Legault B, D'Auria G, Zhaxybayeva O, et al. (2007) Genomic plasticity in prokaryotes: the case of the square haloarchaeon. ISME J 1: 235–245.
56. Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF, et al. (2009) Explaining microbial population genomics through phage predation. Nat Rev Microbiol 7: 828–836.
57. Bolhuis H, Poele EM, Rodriguez-Valera F (2004) Isolation and cultivation of Walsby's square archaeon. Environ Microbiol 6: 1287–1291.

Chapter 4 is a full reprint of: Assembly-driven community genomics of a hypersaline microbial ecosystem. S. Podell, J.A. Ugalde, P. Narasingarao, J.F. Banfield, K.B. Heidelberg and E.E. Allen. *PLoS One*, **8**4:e61692. 2013 (doi: 10.1371/journal.pone.0061692), with permission from all coauthors.

Chapter 5

Deep-sequencing Approaches to Characterize the Fine-Scale Genetic Variation of the Lake Tyrrell Microbial Ecosystem

5.1 Abstract

The availability of habitat-specific genomes allows for a comprehensive characterization of the fine-scale genetic diversity that is present in a microbial community. In this chapter, the genetic diversity of the genomes for the Lake Tyrrell microbial community was studied using deeply sequenced metagenomic datasets.

Illumina metagenomic datasets were generated for four samples from the surface waters of Lake Tyrrell collected at different time points (January 23 & 25 and August 7 & 9, 2007). Reads were analyzed using the available reference genomes for this community and used to characterize both the abundance of individual populations and the fine-scale genetic diversity present in the community.

Results reveal that for some of the members of the community, such as *Haloquadratum*-related populations, a comparatively low level of strain diversity with almost no differences in the population composition between samples. For

other members, such as the *Nanohaloarchaea*, the strain diversity is higher with considerable differences in population composition between samples.

In addition, this analysis allowed the identification of candidate genes under positive selection in some of the populations of the Lake Tyrrell microbial ecosystem. Some of these functions include membrane proteins and secretory systems.

The approach presented in this study represents a broad overview of the fine-scale genetic diversity and generates a foundation for future studies of this community.

5.2 Introduction

The increasing number of microbial genomes sequenced over the last few years has been driven by the higher throughput and lower cost of DNA sequencing technologies. This progress has pushed the field of comparative genomics, allowing for comparisons of the genomes of closely related organisms, up to the strain level [37, 27, 41, 101], in search of insights into their evolutionary history and environmental adaptation [101, 142]. This has been a particularly powerful approach in medical microbiology where multiple cultivated strains are available, allowing for a deep coverage of microbial species of medical interest [35]. Some of this has been replicated in microorganisms isolated from the environment (non-clinical settings), including members of the *Vibrio* genus [20] and *Sulfolobus* [101].

The development of culture-independent approaches has allowed for removing some of the restrictions of culture-based comparative genomics by allowing us to capture directly the taxonomic, functional, and genetic diversity of the members of microbial community through direct sequencing of environmental DNA [10]. A main challenge in the analysis and interpretation of metagenomic data derives due to the genetic complexity of natural microbial communities. In these complex communities, metagenomic approaches often provide only a glimpse of the taxonomic and functional diversity and does not provide enough information to explore the full extent of genetic diversity, particularly at the strain-level.

In the case of low to moderate complexity microbial communities (2-30

dominant microbial species), the use of appropriate sequencing technologies and sampling approaches makes it possible to reconstruct the genomes for dominant community members [10]. In reality, these genomes do not represent a single clone, but are a composite of multiple related strains [96, 3]. This fine-scale genetic variation could have functional relevance for the members of the community, and can reveal information about their evolutionary history and signatures of environmental adaptation [63, 43, 88].

Several studies have approached the study of microbial communities with the goal of quantifying the level of fine-scale genetic variation present in community members using metagenomic approaches [139]. These studies can be divided into two categories based on the type of reference data used to quantify genetic heterogeneity. The first category of study obtains metagenomic sequence data from a microbial community and utilizes reference genomes from isolates to quantify the genetic variation present in the microbial community. Examples of this approach include the study of *Synechococcus* coastal populations [117] and in the human gut microbiome [107]. The limitation of this type of study is that the reference genomes used are not necessarily derived from the same environment as the metagenomic sample. In addition, it is possible to miss novel and abundant groups, by only focusing on genomes available from isolated species [96, 44].]. A second category of study involves the de novo assembly of metagenomic data and the reconstruction of habitat-specific population genomes from which the genetic heterogeneity present in the community can be quantified. Although this approach provides a more complete and less-biased picture of the genetic diversity, it has largely been limited to communities with low species diversity, such as acid mine drainage [4] or heavy-metal contaminated sites [43].

Chapters 2 and 4, described studies carried out on the Lake Tyrrell microbial community where the use of assembly-based metagenomics allowed the assembly of one bacterial and 15 archaeal population-level genomes [79, 96, 95]. These genomes have provided a set of habitat-specific reference genomes that can be used to study fine-scale genetic variability within this community.

Microscale genetic heterogeneity has been previously studied in hypersaline

ecosystems via comparative isolate genomics, such as in the case of *Salinibacter ruber* [93], or using metagenomic approaches [60, 90]. In both cases, this has been done using reference genomes from isolates that were not necessarily recovered from the same sampling location where the metagenomic data was recovered.

This chapter shows the results of a study to quantify the fine-scale genetic diversity that exists in the Lake Tyrrell microbial community. The study combines a deep-sequencing metagenomic approach with the availability of a set of habitat-specific genomes. Two different temporal samples, separated by seven months, were analyzed. The samples were collected in two different seasons, which can provide insights about the stability, both in abundances and in population structure, of this community over time.

5.3 Material and Methods

5.3.1 Sample Collection and Sequencing

Surface water samples from Lake Tyrrell were collected in 2007, during two different seasons, summer (January) and winter (August), with two days of difference in each season (January 23 & 25, August 7 & 9). Each water sample was filtered directly onto a Sterivex cartridge (Milipore, Bedford, MA, USA) (0.22 µm) using a peristaltic pump. For DNA extraction, each Sterivex was processed according to the following protocol:

- Addition of Proteinase K to a final concentration of 0.5 mg/ml^{-1} and SDS to a final concentration of 1%.
- Incubation at 55°C for 25 minutes, followed by incubation at 70°C for 5 minutes.
- Transfer of the lysate from the Sterivex to a clean Eppendorf tube.
- Nucleic acid extraction with two steps of phenol-chloroform extraction.

Construction of sequencing libraries for each of the four samples was performed at the UC San Diego IGM Genomics Center. Libraries were multiplexed

and sequenced on a single lane of Illumina HiSeq (Illumina, San Diego, CA), using the high-throughput mode, with pair-ended reads of 100 nucleotides in length

The demultiplexed reads were processed using Nesoni 0.117 (`http://www.vicbioinformatics.com/software.nesoni.shtml`) to remove adapters, trim low quality positions, and remove low-quality reads from the datasets. For trimming, a minimum quality score of 20 was used, and all reads shorter than 70 nucleotides (after trimming) were removed.

5.3.2 Read Mapping

The trimmed reads were mapped against a set of habitat-specific genomes (Table 5.1 generated by the assembly of metagenomic information of the Lake Tyrrell microbial community [79, 96, 95]. In addition, an archaeal isolate, *Candidatus* Halobonum tyrrellensis [124], obtained from samples collected in August of 2007, was included in this set of reference genomes. Each genome is labeled based on their phylogenetic classification on the original work, with the prefix J07 for the genomes representing January samples, and A07 for the genomes that represent August samples. Each metagenomic library (January 23, January 25, August 7 and August 9) was mapped independently to the reference genomes, using Bowtie 2.2.1 [59] with the *very-sensitive* alignment option and adjusting the N-ceiling function to (0,0.01) to reduce the number of ambiguous characters present in the alignment. Several tools were used for the analysis of the resulting files, including SAMtools 0.1.19[61], BEDtools 2.17 [98], and BCFtools 0.2.0 (`http://samtools.github.io/bcftools/bcftools.html`).

Coverage plots were generated with custom Python scripts, using the BAM files generated by read mapping.

To determine the differential coverage of each gene in the reference genomes, the RPKM (reads per kilobase per million reads mapped) values were calculated according to the following equation:

$$\text{RPKM} = \frac{\text{№ of mapped reads to the gene} * 10^9}{\text{№ of reads mapped in the experiment} * \text{Gene length}}$$

To facilitate the visualization of differences between samples from January

and August using the RPKM values, these were normalized using this formula:

$$\log_2(\frac{\text{January RPKM}}{\text{August RPKM}})$$

A two-tailed Fisher exact test (pvalue < 0.05) was used to determine which genes had differential recruitment of reads between the two seasons (January and August).

5.3.3 Taxonomic Classification of Mapped and Unmapped Reads

To compare the taxonomic diversity found in the mapped and unmapped reads, both sets were classified using Phylosift 1.0.1 [22], with the provided set of marker genes. These markers included all of the January 2007 genomes that were assembled previously from the Lake Tyrrell community [79, 96] (all the genomes with the prefix J07 in Table 5.1), but did not include the additional Lake Tyrrell genomes recently assembled from metagenomic samples from the August 2007 community, and described in [95].

5.3.4 Variation Analysis

The resulting BAM files with the information for the mapped reads were processed with Picard Tools 1.99 (`http://picard.sourceforge.net`), to sort and reorder the mapping information. GATK 2.7.2 ([23] was used to realign indel regions found by mapping against the reference genomes. These corrected files were processed using Freebayes v9.9.2-29-g9ed353c [38] (ploidy:1, minimum base quality: 20, minimum mapping quality: 30) to obtain a set of high-quality variations. Quantification of the type of variations (e.g. single nucleotide polymorphisms, SNPs) was done using SnpEff [18] and visualized using custom Python scripts.

To calculate the role of selective pressure in each gene, the ratio of non-synonymous to synonymous polymorphisms was calculated. Commonly this ratio is referred as dN/dS, but in this context we are looking at populations (we are not able to distinguish individual members of the population), so it will be referred as

pN/pS [107]. Calculation of the pN/pS values for each coding sequence were done using custom Python scripts, based on the approach used in Tai *et al.* [117]. For each gene, the pN/pS value was calculated as:

$$\frac{\text{pN}}{\text{pS}} = \frac{\frac{\text{Observed non synonymous mutations}}{\text{Number of non synonymous sites}}}{\frac{\text{Observed synonymous mutations}}{\text{Number of synonymous sites}}}$$

All of the visualizations of these results were done using custom Python scripts. To evaluate the differences in functional classifications, the Cluster of Orthologous Groups [118] annotation was used. The comparisons of abundance were done using an odds ratio test:

$$\frac{\frac{\text{№ of proteins under selection in category X}}{\text{№ of proteins not under selection in category X}}}{\frac{\text{Total № of proteins under selection, minus category X}}{\text{Total № of proteins not under selection, minus category X}}}$$

Statistical significance was evaluated using the 2X2 contingency table with a one-tailed Fisher exact test (pvalue < 0.05).

5.3.5 Statistical and Computer analysis

All the analyses were carried out on a large cluster instance (c3.8xlarge: 32 Intel Xeon E5-2680 v2 cores, 60 Gb RAM) using the Elastic Cloud Computing (EC2) infrastructure from Amazon Web Services (AWS). Plotting and calculations were carried out using Python scripts, with the standard packages and the Biopython [19], Numpy [81], Pandas [75], PyCogent [57] and Matplotlib [46] libraries. All statistical analyses were carried out using Python scripts and the Scipy libraries [81].

Table 5.1: List of the Lake Tyrrell habitat-specific genomes used for read mapping

Genome name (abbrv)	**Length**	**G+C pct**	**№ scaffolds**	**Reference**
Haloquadratum walsbyi J0HQW1	3,549,539	47	1	[96]
Haloquadratum walsbyi J0HQW2	3,475,501	49	1	[96]
Haloquadratum sp. J07HQX50	3,019,909	50	2	[96]
Nanosalinarum sp. J07AB56	1,215,802	56	3	[79]
Nanosalinarum sp. J07AB43	1,277,157	43	7	[79]
Halonotius sp. J07HN4	2,888,659	61	2	[96]
Halonotius sp. J07HN6	2,529,000	63	6	[96]
uncultured archaeon sp. J07HX64	2,982,938	64	1	[96]
uncultured archaeon sp. J07HX5	2,040,945	60	1	[96]
Halobaculum sp. J07HB67	2,649,547	67	3	[96]
Halorubrum sp. J07HR59	2,120,805	59	7	[96]
Salinibacter sp. J07SB67	1,931,021	67	443	[96]
Halorubrum sp. A07HR60	2,876,249	59	14	[95]
Halonotius sp. A07HN63	2,392,686	63	37	[95]
Halorubrum sp. A07HR67	2,890,468	67	16	[95]
uncultured archaeon A07HB70	2,389,822	71	15	[95]
Candidatus Halobonum tyrellensis G22	3,675,087	70	72	[124]

5.4 Results and Discussion

5.4.1 Overview of the Illumina datasets

In all four samples, between 71% and 74% of the original reads were retained after trimming and quality filtering (Table 5.2), representing an average of 6.9 billion bases per sample.

Preliminary visualization of the community composition was done by looking at the G+C content of each of the libraries. This allows capturing broader differences in community composition between samples [96, 39, 95]. Figure 5.1 shows the differences between the four libraries, highlighting the location of the reference genomes that will be used for read mapping and where they were recovered from this same community. The plot shows that the January community (in particular, the sample collected on January 25) is dominated by organisms with a low G+C content compared to the August community. This is similar to previous observations for the Lake Tyrrell microbial community [95]. The main driver of these differences was suggested to be the ionic composition of the water column, particularly the concentration of magnesium, which is higher in the January samples compared to the August ones (Table 5.3) and where microorganisms such as *Haloquadratum* (J07HQW1, J07HQW2, and J07HQX50) are more abundant in the January sample due to their tolerance to higher concentrations of magnesium [95].

Besides the differences between different months among the samples, we can observe differences within the January libraries with different G+C peaks in the January 23 versus the January 25 sample. In particular, the January 25 sample shows a higher peak at lower G+C, compared to the January 23 sample. Looking back into the chemical measurements done for these samples (Table 5.3) [95], we can see that the magnesium concentrations on the January 23 sample are lower compared to January 25. Weather records showed that there was an input of freshwater due to a storm prior to the sampling on January 23 [95], which suggests an effect on the overall concentrations of dissolved salts, including magnesium. After two days, mainly due to water evaporation, the concentration of dissolved salts in the water column increased, which could explain the difference in the

G+C plots and thus the changes in population abundances between the different samples.

Table 5.2: Summary of the total reads before and after trimming, for each of the four Illumina HiSeq libraries.

Library name	**Total reads**	**Read-pairs after QC**	**Unpaired reads after QC**	**Total bases (Mb)**
January 23	49,963,357	37,016,243	7,679,004	7,978.18
January 25	39,400,015	29,444,267	5,894,815	6,333.12
August 7	46,472,319	33,485,834	7,659,231	7,266.38
August 9	40,256,946	28,843,346	6,812,171	6,276.12

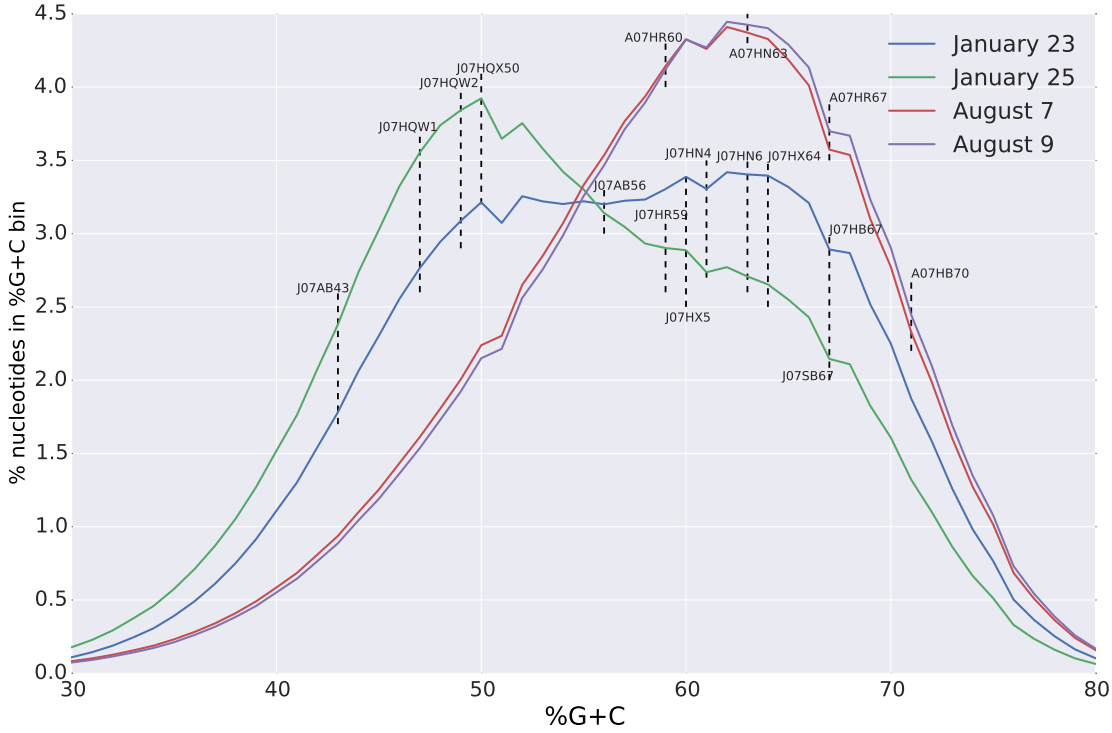


Figure 5.1: Percentage of nucleotides versus G+C content in each of the four sequenced libraries, where each G+C bin has a size of 1%. Dashed lines indicate the position, based on G+C content, for each of the reference genomes isolated from this community, and that will be used for read mapping (Table 5.1)

Table 5.3 *(next page)*: **Table 5.3:** Physical and chemical composition of the Lake Tyrrell water samples. Concentrations are given in units of $\mathrm{mmol\,L^{-1}}$

Sample	**Temp** ℃	**Total ionic strength**	**Na**	**K**	**Mg**	**Ca**	**Cl**	**Cl**	**SO42**
Jan 23	21.6	pH	5,721	4,338	32	298	10	5,345	123.6
Jan 25	27.9	7.09	5,950	4,163	43	419	11	5,291	170.5
Aug 6	9.9	7.00	4,403	3,724	19	126	15	4,298	50
Aug 8	11.5	7.01	4,060	3,557	18	117	14	3,830	47

5.4.2 Read Mapping using Habitat-Specific genomes

All the reads that passed the quality filters were mapped against the set of habitat-specific genomes (Table 5.1) from the Lake Tyrrell community. The results indicate differences in the number of reads that each genome recruited, both comparing between samples and between genomes (Table 5.3).

In the January samples, the genomes that recruited the most number of reads were those associated with *Haloquadratum* (J07HWQ1 and J07HWQ2), which agrees with previous observations regarding the abundance of these organisms in the Lake Tyrrell community [96]. The raw number of reads does not provide the best criteria to estimate the relative abundance of each genome in the overall dataset as these genomes vary in size. Rather, the depth of coverage for each genome (Table 5.4) provides a better metric for these estimations. The overall coverage, shows similar relative abundances compared to previous observations [96], but it is important to highlight that the number of mapped reads and the coverage do not necessarily reflect the real abundance of each organism in the community. Given the strict criteria used for mapping, it is likely that many sequences that could be recruited onto individual genomes could be missing. For example, the stringent criteria used could explain the low number of reads and low coverage for the genome of *Candidatus* Halobonum turrellensis (G22). This particular organism is an interesting case because it is the only genome that was obtained from an isolate (cultured from August 2007 samples) [124], but its abundance in the community appears to be very low compared to the other genomes that were recovered via metagenomic assembly. This is a clear example of the often observed phenomenon where cultured isolates from an environment are not representative of the most abundant members of a microbial community.

We can estimate how stringent the mapping criteria were by looking at the sequence identity of each read compared to the mapped genome and quantifying this for each library and genome (Figure 5.2). With the exception of the G22 genome, the majority of reads mapped at a 100% sequence identity, and it was never lower than 85% for any of the genomes and libraries. This analysis alone provides an accurate overview of the level of genetic heterogeneity that is

present in these populations. For example, for the *Haloquadratum* genomes, the majority of the reads mapped at 95% sequence identity, while in the case of the *Nanohaloarchaea* genomes, the sequence identity of the recruited reads goes down to 89%.

Table 5.3: Total number of recruited reads to each reference genome.

Genome	**Jan 23**	**Jan 25**	**Aug 07**	**Aug 09**
J07HWQ1	9,712,976	10,347,084	1,802,421	1,385,337
J07HWQ2	7,311,175	9,428,490	1,628,137	1,301,141
J07HQX50	922,138	1,041,326	501,477	330,307
J07AB56	565,197	266,831	194,445	165,278
J07AB43	760,203	486,360	63,209	64295
J07HN4	2,149,204	2,249,692	1,306,287	1,089,673
J07HN6	592,818	831,367	1,027,472	911,341
J07HX64	4,167,113	1,819,023	1,202,103	1,144,206
J07HX5	2,106,559	1,382,371	673,972	539,843
J07HB67	1,124,816	1,128,191	125,973	84,643
J07HR59	839,856	310,496	2,105,598	1,693,772
A07HB70	550,429	277,030	1,970,106	1,866,967
A07HR67	563,043	270,602	2,166,129	1,680,150
A07HN63	547,808	786,856	1,126,032	1,003,322
A07HR60	2,126,700	758,549	5,405,933	4,362,857
G22	62,983	39,696	72,261	66,778
J07SB	797,957	211,306	737,471	673,630
Unmapped	45,344,829	31,913,858	51,012,461	44,849,506

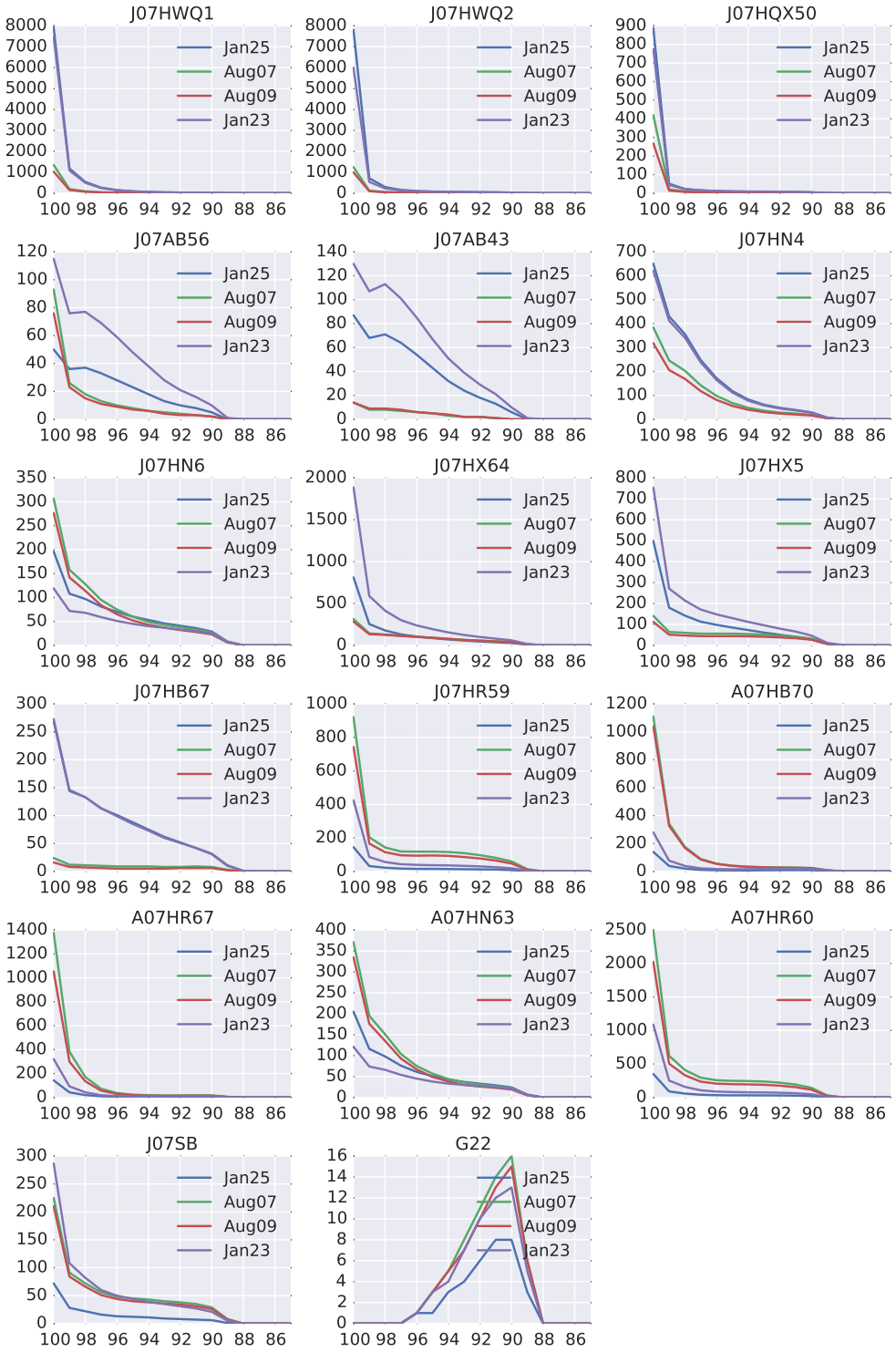


Figure 5.2: Total number of recruited reads, grouped by sequence identity. The X axis shows the identity of the read to the reference genome (%), while the **Y** axis shows the number of reads recruited at that identity (thousands of reads).

Table 5.4: Genomes coverage (expressed as X-fold) in each of the libraries.

Genome	**Jan 23**	**Jan 25**	**Aug 07**	**Aug 09**
J07HWQ1	274.9	292.7	51.0	39.2
J07HWQ2	200.2	258.0	44.5	35.6
J07HQX50	30	33.9	16.3	10.7
J07AB56	45.5	21.5	15.6	13.2
J07AB43	61.2	39.1	5.1	5.2
J07HN4	72.4	75.7	44.0	36.7
J07HN6	22.8	31.9	39.4	35.0
J07HX64	135.5	59.1	39.1	37.2
J07HX5	100.5	65.9	32.1	25.7
J07HB67	40.9	41.0	4.6	3.1
J07HR59	38.6	14.2	96.7	77.7
A07HB70	22.1	11.1	79.1	74.9
A07HR67	18.8	9.0	72.2	56.0
A07HN63	22.2	31.9	45.7	40.7
A07HR60	72.0	25.7	183.1	147.7
G22	1.6	1.0	1.9	1.7
J07SB	40.1	10.6	37.1	33.8

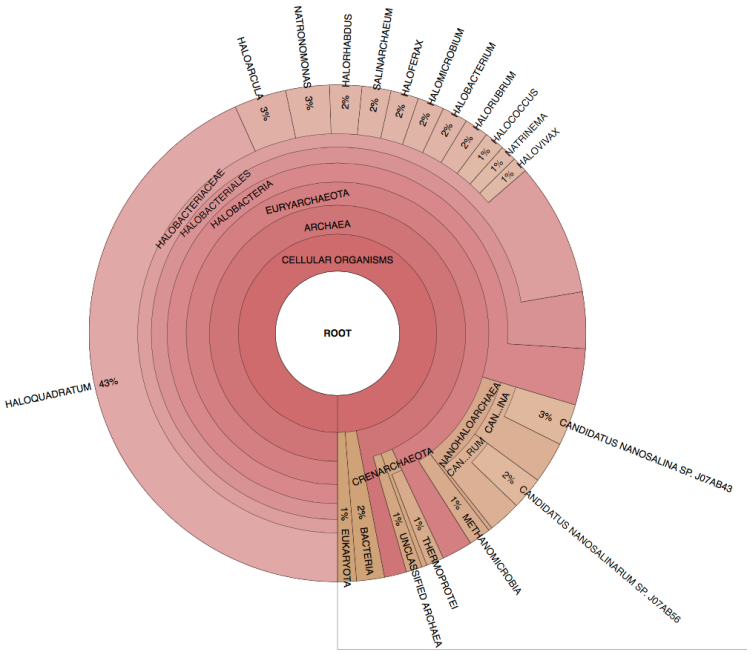
5.4.3 Taxonomic Classification of Mapped and Unmapped Reads

Before continuing with a more detailed analysis of the genetic heterogeneity present in the Lake Tyrrell microbial community, we need to address the results of the read mapping to the reference genomes. With the deep level of coverage of the community achieved with the four libraries (approximately 6 billion nucleotides in each library), this dataset represents not only the most abundant members of the community, as represented in the reference genomes used for mapping, but also allows the possibility of discovering novel organisms in the unmapped sequences [79, 2]. To provide an overview of the differences between the set of mapped and unmapped reads, the software Phylosift [22] was used to generate a taxonomic classification of both sets of reads. An example of these results is shown in Figure 5.3 for the sample collected on January 23. In the case of the mapped reads (Figure 5.3a), as expected the majority of the reads were classified as *Haloquadratum* followed by the *Nanohaloarchaea*. There are also hits to organisms that were not present in the reference genomes, such as *Natronema*, something that can be explained by the size of the reads (100 bp.), which could generate spurious hits against more distantly related species at highly conserved genomic regions. In contrast, the set of unmapped reads (Figure 5.3b) shows a broad diversity of taxonomic groups, including groups not present in the reference genomes, such as *Natronomonas*, but also groups that are present, such as *Salinibacter* and the *Nanohaloarchaea*. This suggests that the diversity of these groups is higher than expected and is not solely represented by what is present on the reference genomes used for mapping. This also highlights the potential for the recovery of novel genomic sequences, either from this group or from novel representatives from other archeal genera.

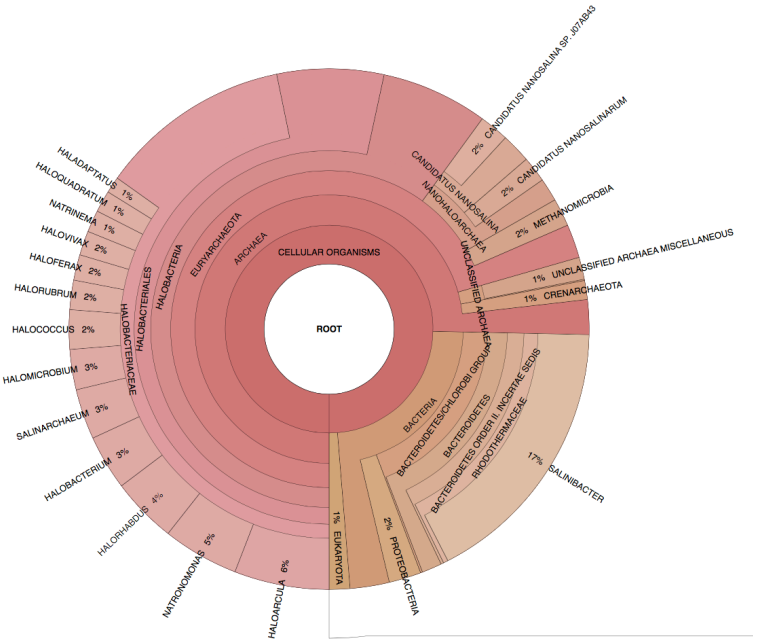
To evaluate the differences in taxonomic composition between the set of mapped and unmapped reads, the Phylosift results were evaluated using an edge principal component analysis (EPCA), which takes into account the phylogenetic composition (similar to Unifrac distances) [70]. Looking at the first two components (Figure 5.4) shows that based on the predicted taxonomic composition of

all the libraries, the reads separate between the mapped and unmapped groups in addition of a separation by sample (January versus August) in the case of the unmapped reads. Technical problems limited the analysis of two of the datasets from the mapped reads, so the only conclusions that can be made currently concern the separation between the unmapped and mapped reads.

As mentioned earlier, the taxonomic analysis suggests that indeed there are novel groups in the unmapped reads that are not represented in the reference genomes. This also includes the presence of viral sequences, which were not taken into account in the taxonomic analysis and likely compromise a large percentage of the sequences present in each of libraries [103, 34]. Further work should include viral markers in the taxonomic classification, and mapping the sequences against available viral genomes. Nevertheless, by using the current set of reference genomes, some of the most abundant members of this community are being explored, and this information can be used to explore the genetic diversity present in the community by using a set of already validated habitat-specific genomes.



(a) Mapped reads



(b) Unmapped reads

Figure 5.3: Taxonomic classification of the mapped and unmapped reads using Phylosift [22]

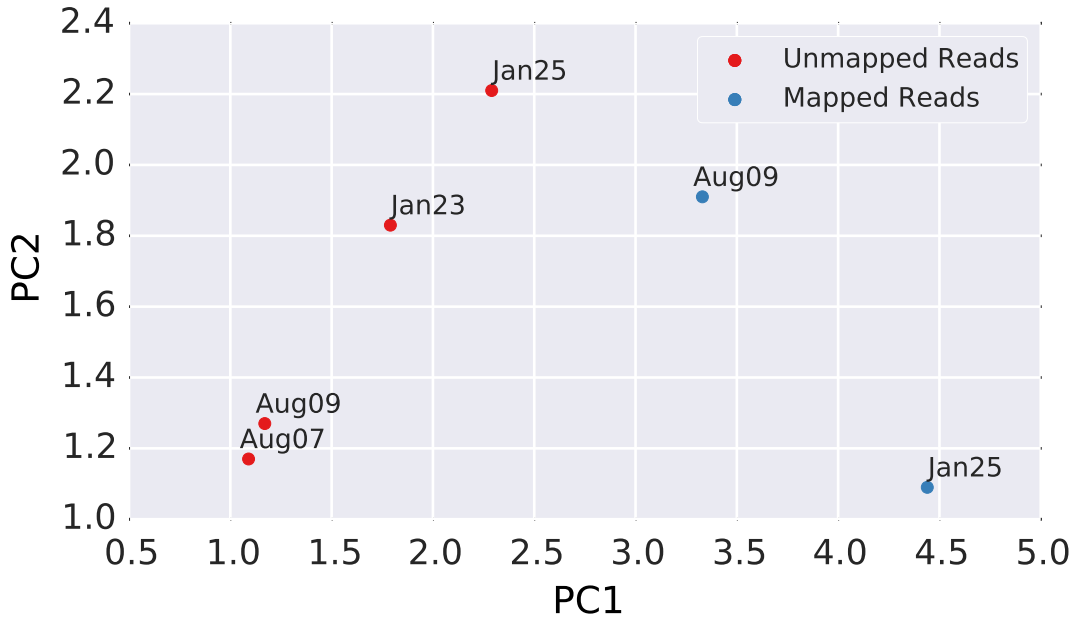


Figure 5.4: Edge principal component analysis (EPCA) of the taxonomic classification of each library.

5.4.4 Differential Coverage of Genomes and Genes

As mentioned in the previous sections, the number of mapped reads to each of the reference genomes and the genome coverage (Tables 5.3 and 5.4), suggests differences in the relative abundance of some of the populations between sampling times (January versus August). For example, the genomes of *Haloquadratum* microorganisms (J07HWQ1 and J07HWQ2) recruited more reads from the January samples than from the August samples. By comparison, the reference genomes that were assembled from the August samples [95], such as A07HR60, recruited more reads from the August libraries. However, just looking at the total coverage of a genome does not provide a complete picture because it is possible that some regions have higher coverage than other regions.

This differential coverage can be analyzed from two perspectives if we consider the comparison between January and August samples (combining the two individual dates for each season). From the analysis within each month, some regions may show depths of coverage that are lower than the rest of the genome. This suggests the presence of multiple strains within that particular organism where only some of the members of the population have that particular region [90, 60, 3]. This has been shown in the case of *Haloquadratum* [60] and *Salinibacter* [90], using reference genomes and mapping reads from different environments to show the presence of these metagenomic islands. The second perspective is to compare the coverage of each genome between the January and August samples to look for regions of differential coverage (higher in one sample versus the other).

To evaluate this differential coverage within and between samples, the identity and coverage of all the reads along each one of the reference genomes was evaluated. This was done for both January and August samples (Appendix B). This approach allows the identification of regions with low coverage, suggesting a region only present on a subset of the strains of the populations, as well as the differential coverage between January and August. In addition, changes in the relative abundance of each gene were incorporated with the goal of identifying differential covered genes (not only regions). Looking at the individual genes allows the identification of any possible functional processes that are more abundant in

one sample versus the other.

As expected, the coverage plots indicate that for some of the reference genomes, there are differences in the coverage along the genomic sequence (Appendix B). For example, in the case of the *Nanohaloarchaea* J07AB56 (Figure B.4), there are two regions that appear with a higher coverage in the August samples with genes that have a differentially higher coverage in this sample. Looking in more detail, within the first region, genes encoding for hypothetical proteins (several of them with homology to halophages), DNA-primases and replication proteins were found. This region is located within the largest scaffold of the J07AB56 genome, suggesting that is indeed part of the genome. A possible hypothesis is that this region constitutes a halovirus that is carried as a prophage, but was released on the August sample (maybe due to environmental changes) [97]. The second region has genes encoding for proteins that take part in detoxification mechanisms, such as ABC transporters for xenobiotics and oxidoreductases, involved in the degradation of xenobiotics. In addition, several hypothetical proteins can be found in this region.

A summary of the differential recruitment of the reads at the gene level is shown in Figure 5.5, which indicates that in the case of the reference genomes that were recovered from the January samples, the majority of genome genes indeed recruited to reads from the January Illumina libraries. Some exceptions to this are in the *Halonotius* J07HN6 genome, which has a large fraction of the genes recruiting more reads from the August libraries than from the January ones. It is possible that this organism is present and abundant in the August samples, but it was not previously assembled. Also the *Halonotius* J07HN6 and A07HN63 are very similar [95], which could also explain the higher differential coverage.

The idea of looking at the differential coverage of the genomes and the comparison between multiple samples is not new in the literature [28, 90], as it provides a broad picture of the possible strain variation that is present for each of the populations under study. To our knowledge, the idea of using RPKM values, commonly used in gene expression studies [77], and applying it to comparisons to metagenomic samples is indeed a novel one. Although a broad picture of the

differences was presented here, further work will focus in specific populations, such as the *Haloquadratum* genomes, and will also include the use of available isolate genomes [28].

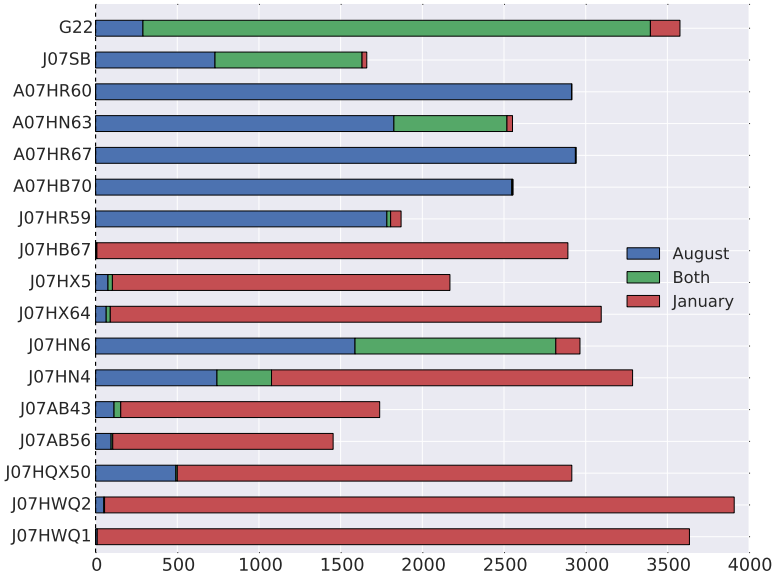


Figure 5.5: Number of genes that differentially recruited reads from either the January or August libraries, in each of the reference genomes. A two-tailed Fisher Exact test (p-value < 0.05) was used to determined the differences between samples. *Both*, indicates genes that were not found to be significantly more abundant in either of the samples.

5.4.5 Fine-scale Genetic Variation: Single Nucleotide Polymorphisms (SNPs)

The genetic heterogeneity that is present in each of the reference populations can be explored in more detail by leveraging the information provided by the reads and the high depth of coverage of each of the genomes (with the exception of G22). With this, it is possible to quantify the nucleotide variation in each genome (single nucleotide polymorphisms, SNPs) and to evaluate the effects of such variation at the functional level. Other types of variations, such as insertions, deletions, and structural variants, can be analyzed as well, but the current goal is to evaluate the evolutionary forces that are generating such variation. Under such an evaluation, the emphasis will be on the polymorphisms that act on protein-coding sequences.

For variant calling, only high-quality SNPs were considered in the analysis, filtering the alignment files priors to SNP calling with Freebayes. Only those sites with a quality score of 20 or more and a mapping quality score of 30 or more were selected in the analysis, similar to methods employed in other studies [107]

The number of SNPs/Kb on each of the genomes is shown on Table 5.5, indicating differences on the number of SNPs when comparing the different references. Within each of the genomes, there are small differences within each season sample (January 23 & January 25, August 7 & August 9), but there are differences between the January and August samples. The *Nanohaloarchaea* genomes exhibit the largest number of SNPs/Kb among all the genomes in the January libraries, but it is lower on the August libraries. Comparing the number of SNPs/Kb found on each library, we can visualize that all four of them are within similar ranges (Figure 5.7) with averages between 4-5 SNPs/Kb.

Although from the visual inspection of Table 5.5 it does not appear that there is any relationship between those genomes that recruited more reads and the number of SNPs, this needs to be evaluated. When comparing the depth of coverage of each reference genome versus the number of SNPs/Kb found in each of the libraries (Figure 5.6), there is no relationship between these two variables. This was confirmed by testing with the Spearman correlation coefficient for the

combined libraries (coefficient: 0.16, *pvalue*: 0.21). This strongly suggests that the differences in SNPs numbers are due to the genetic diversity found in each of these reference genomes, where organisms such as the *Nanohaloarchaea* J07AB43 and J07AB56 have higher rates of genetic diversity (more strain diversity) compared to organisms such as J07HWQ1 or J07HWQ2.

The effect of each variation within the genome can be evaluated in more detail by quantifying those SNPs that were intergenic or located within a coding region. In the latter case, this variation could be either synonymous (no amino acid change) or non-synonymous (change in the resulting amino acid). For each of the genomes, no differences between libraries in the percentage of intergenic, non-synonymous, and synonymous SNPs (Figure 5.8) were observed; the exception was the G22 genome, but the overall coverage and percentage of SNPs for this genome are low. The distribution of SNPs for all the genomes (Figure 5.9) shows that the majority of the SNPs are located in coding sequences and are synonymous. Also, it is interesting to observe the dispersion on the data, where for the case of the non-synonymous SNPs, the distribution goes between 20-30%, while in the case of intergenic SNPs, it moves between 15-37%. These can be observed more in detail on Table 5.6 where on one extreme there are genomes from the *Haloquadratum* group with over 30% of the SNPs within intergenic regions, and on the other side the *Nanohaloarchaea*, with 10% of the SNPs within intergenic regions. This can be explained by the differences in genome size, where *Haloquadratum* organisms have large genomes (over 3 Mb.), and the *Nanohalorachaea* have smaller genomes (1.8 Mb) with reduced intergenic space [79, 96]. Overall, the percentage of non-synonymous substitutions for all the genomes was similar, something expected given the characteristics of this type of change, which modifies the amino acid sequence and can led to non-functional organisms. Compared to values observed in other metagenomic studies (focusing only on a few species), the trends observed are similar, but with higher rates of synonymous substitutions compared to non-synonymous ones [114].

The sampling strategy allows us to ask two different questions in terms of the temporal variation between the samples. The first comparison is a variation

within each sampling month (January and August) between the samples collected two days apart. The second comparison is between the two different seasons (January vs August). By looking at the SNPs found on each of the samples, we can look whether the SNPs found are unique to each one of the libraries or there are similarities between time points. We can visualize the differences using Venn diagrams to compare the different sampling dates. The comparison between the January dates (23 versus 25) (Figure 5.10) shows that in all the reference genomes, the majority of the SNPs is shared between the two samples days. G22 is an exception due to its low coverage. A similar trend is observed in the comparison of the August 7 and August 9 libraries (Figure 5.11),however, in contrast to the January data, the*Nanohaloarchaea* genomes (J07AB56 and J07AB43) have differences between the two dates. These results suggest that different populations of these organisms may be found in the two samples. Further analysis (focusing only in these two populations) is needed to answer this question and evaluate whether these differences are indeed due to differences in population structure for these organisms [107, 111, 129]. Both dates (January versus August) were compared using the combined set of SNPs. Interestingly, the majority of the SNPs appear to be shared between the two samples with small differences in the case of the *Nanohaloarchaea* and with unique SNPs present in the January sample. This suggests that the populations found in January could be very similar to the ones present in August [26]. The same strains could be present, however the abundance of the members of the community varies. This is supported by the previous analysis of the reads that mapped to each reference genome. Exploring this phenomenon requires focusing on a few organisms, comparing not only these assembled genomes, but also genomes from other environments and/or isolates if available. In particular, measurements of nucleotide diversity, fixation index, and McDonald-Kreitman tests [107, 114] should be used to evaluate differences at the population level between the members of the community. Nevertheless, the initial conclusion of both populations being similar, with only differences in their abundances, is intriguing. The slow growth rate of some of the organisms found in hypersaline communities [28] (at least in laboratory studies) may suggest that samples of seven months apart is

not long enough to provide a temporal picture of the evolution of the population structure within this ecosystem. This contrasts with what has been observed for viral communities in hypersaline communities, where rapid temporal variation has been observed [103, 33].

Table 5.5: Count of number of SNPs per kilobase in each of the Illumina libraries for the reference genomes.

Genome	**Jan 23**	**Jan 25**	**Aug 07**	**Aug 09**
J07HWQ1	3.83	3.81	3.75	3.70
J07HWQ2	2.51	2.52	3.28	2.46
J07HQX50	1.44	1.44	1.42	1.39
J07AB56	10.10	9.71	4.95	5.21
J07AB43	10.75	10.64	7.30	7.19
J07HN4	9.09	9.09	9.01	9.01
J07HN6	2.87	2.61	2.85	2.81
J07HX64	9.52	9.43	9.35	9.26
J07HX5	9.26	9.17	9.09	9.01
J07HB67	9.09	9.17	6.45	5.24
J07HR59	1.36	1.15	1.54	1.51
A07HB70	7.19	7.04	7.30	7.30
A07HR67	5.85	5.71	5.92	5.92
A07HN63	1.72	1.98	2.12	2.10
A07HR60	3.79	3.69	3.91	3.89
G22	0.04	0.03	0.03	0.03
J07SB	6.06	5.65	6.10	6.10

Table 5.6: Percentage of the different type of SNPs on each genome

Genome	**Intergenic**	**Synonymous**	**Non-Synonymous**
J07HWQ1	36.90	38.56	24.55
J07HWQ2	37.31	34.48	28.22
J07HQX50	31.32	36.95	31.73
J07AB56	10.67	65.96	23.37
J07AB43	10.30	64.02	25.68
J07HN4	15.27	66.28	18.45
J07HN6	19.98	55.62	23.40
J07HX64	18.98	51.18	29.84
J07HX5	20.72	53.88	25.41
J07HB67	12.06	60.65	27.29
J07HR59	23.78	53.89	22.33
A07HB70	22.38	49.23	28.40
A07HR67	20.75	55.58	23.67
A07HN63	21.55	54.69	23.76
A07HR60	26.79	49.54	23.67
G22	15.86	65.34	17.80
J07SB	15.67	59.60	24.74

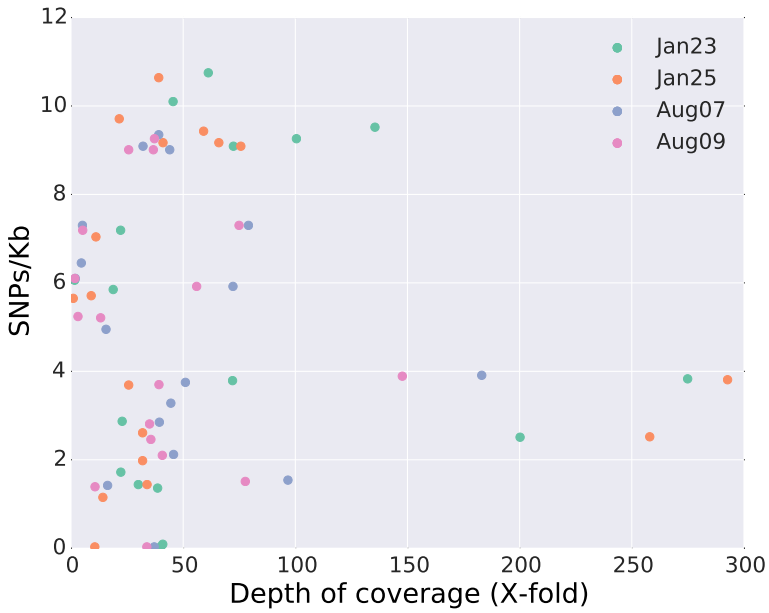


Figure 5.6: Scatterplot of depth of coverage versus SNPs/Kb for each of the reference genomes for the four libraries.

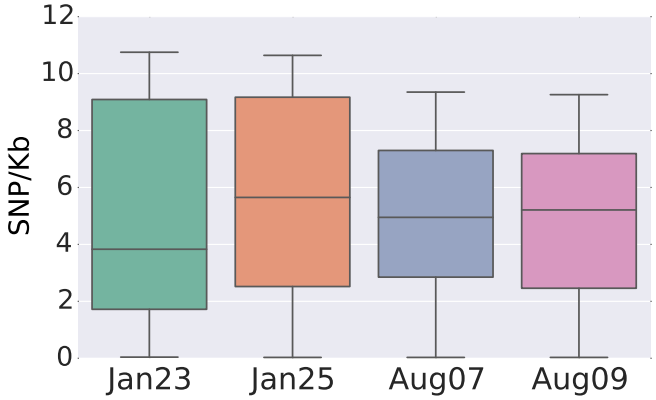


Figure 5.7: Boxplot summarizing the number of SNPs/Kb for each genome, in each of the sequence libraries.

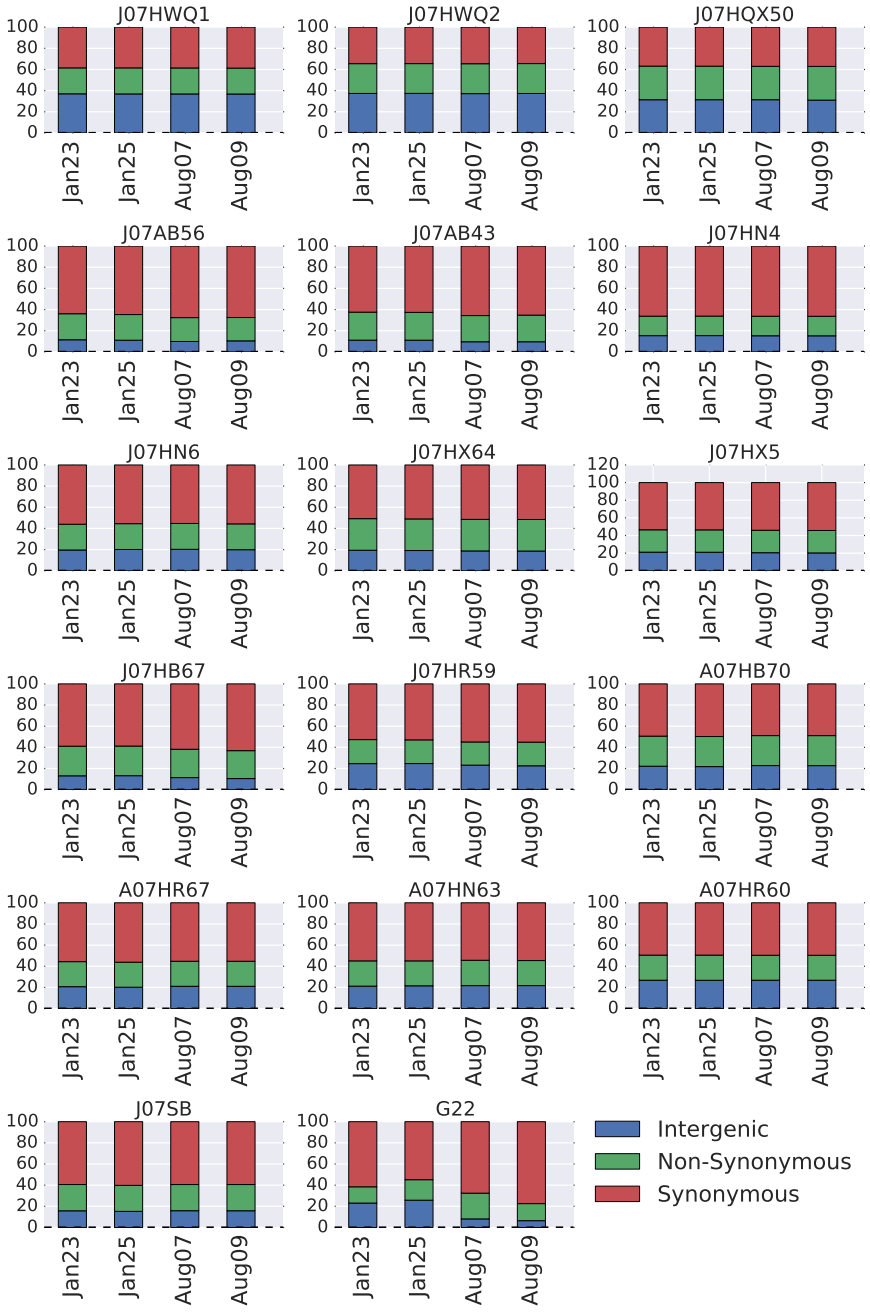


Figure 5.8: Percentage of intergenic, non-synonymous and synonymous SNPs in each genome, for all the sequence libraries.

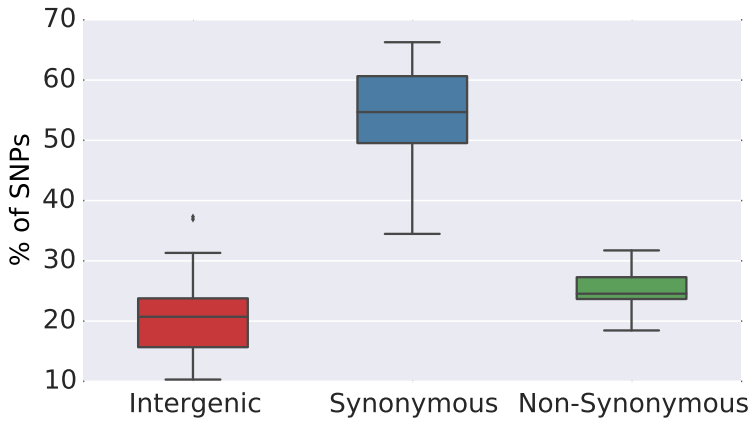


Figure 5.9: Boxplot summarizing the distribution of type of SNPs (intergenic, synonynous, non-synonymous) for all the genomes and samples.

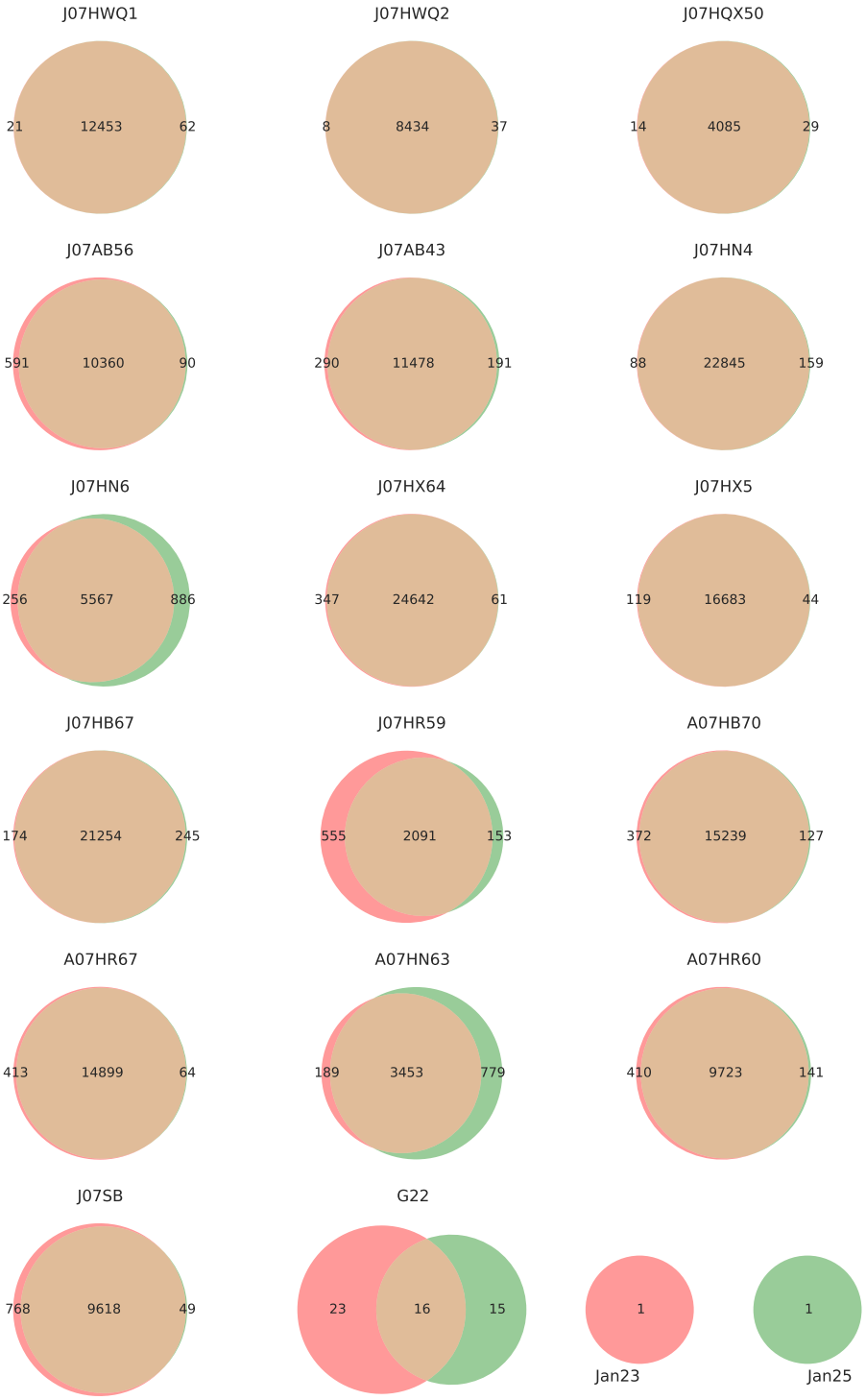


Figure 5.10: Venn diagram comparing the SNPs found in the January 23 versus January 25 libraries.

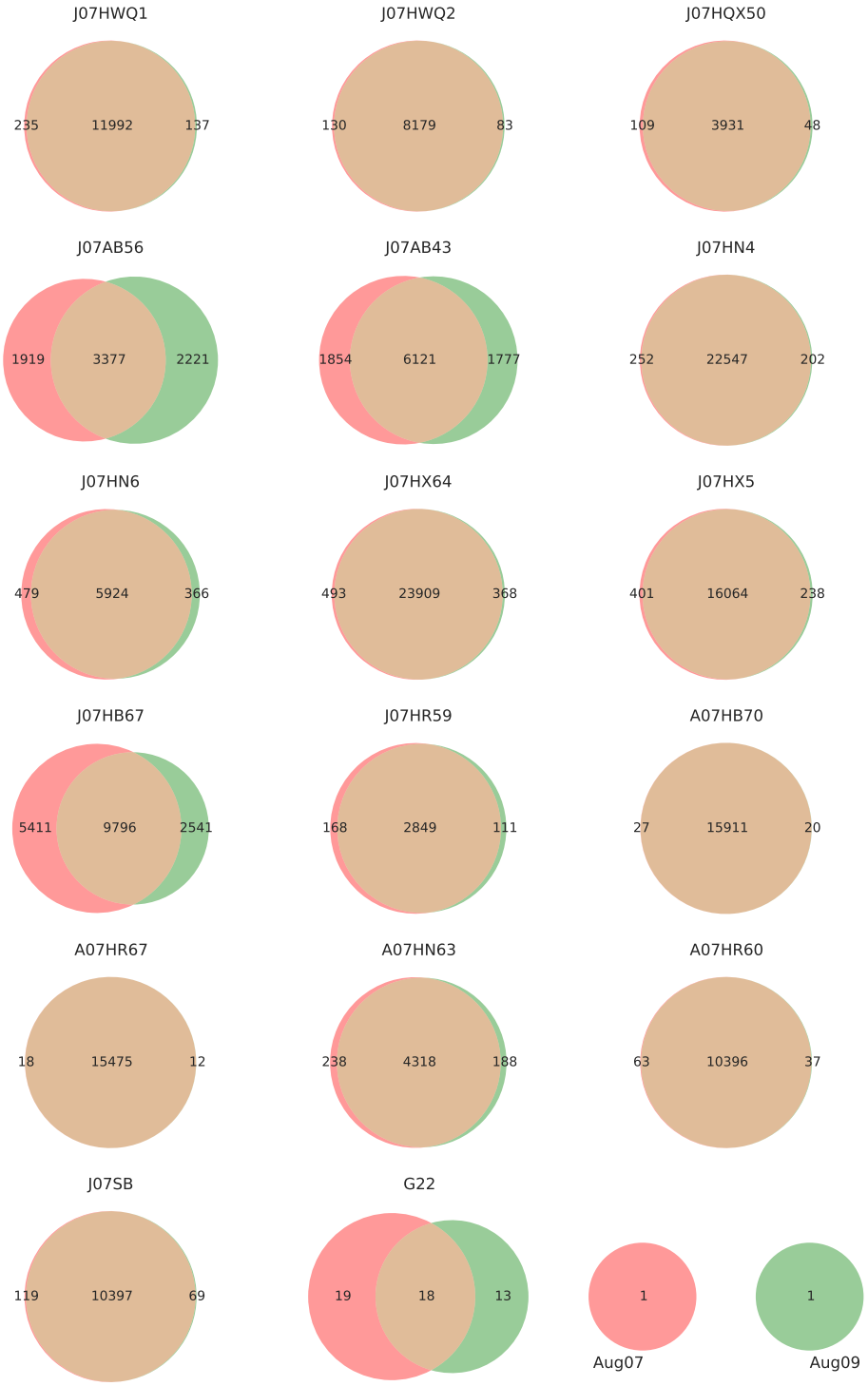


Figure 5.11: Venn diagram comparing the SNPs found in the August 7 versus August 9 libraries.

Figure 5.12: Venn diagram comparing the SNPs found in the January versus the August libraries

5.4.6 Genes Under Positive Selection

The results from the variant analyses allow for the evaluation of the role that natural selection is playing in the evolution of individual genes across the community and within individual population. One approach involves looking at the ratio between non-synonymous and synonymous substitution on each individual gene. This is defined as the dN/dS ratio [73], which compares genes between individual species. Similarly, the pN/pS ratio compares this effect at the population level, which is the case here, where we do not have individual isolates that are being under study, but a complex sample [30, 107]. Using the pN/pS ratio, we can identify genes that are under the effects of positive selection ($pN/pS > 1$), where natural selection is favoring diversification at the amino acid level [47]. On the other extreme, purifying selection ($pN/pS < 1$) occurs when fewer amino acid changes occur than the ones expected by chance [47].

There are multiple methods for detecting molecular adaptation based on pN/pS ratios [143]. Because short reads are being used to identify polymorphisms in the genomes, it is not possible to link multiple polymorphisms. This makes statistical approaches, such as Maximum Likelihood and Bayesian methods, difficult to implement for our data [143]. Instead, a pairwise method was used, quantifying the polymorphisms site by site, using a strategy based on previous work by Tai *et al.* [117] (Figure 5.13).

Average pN/pS values for each genome

First, the average pN/pS values were calculated for each of genomes to determine overall patterns of selection between samples. Figure 5.14 shows a comparison of the average pN/pS value for each genome between the months of January and August. Besides the extreme value of G22, no large differences are observed between the two months. This suggests that the overall values of pN/pS for the genes are similar between the two sampling dates.

pN/pS values and uniquely selected genes

The pN/pS values were calculated for all the coding regions in each of the reference genomes. To facilitate the comparisons, the analysis was limited to the comparisons between the January and August datasets, merging each individual time point within the same sampling month. An overview of each genome (Appendix D) shows that there does not appear to be hotspots of selection across the genomes. For example, in the genomes of J07HWQ1 (Figure 5.16) and J07HWQ (Figure 5.17), no striking differences between January and August were observed. Also, the genes under selection are located sparsely along the genome.

Comparing more in detail for January versus August for each reference genome, the results (Table 5.7) indicate that in both samples, the number of genes under positive selection (pN/pS > 1) is similar within each of the reference genomes. Comparing across the genomes, we can see that the percentage of genes under selection in each of the genomes varies between 1.5% up to 7% (without considering G22). This is similar to values observed in comparisons for isolate genomes of *Salinibacter* [93], as well in other similar metagenomic studies in *Synechococcus* populations [117]. A very interesting result, which is similar to what was observed by just looking at the number of shared SNPs between sampling months, is that for several of the genomes, the same genes are under selection in both January and August. Although more formal population analysis needs to be done to confirm this result, this strongly suggests that for some organisms, such as *Haloquadratum* (J07HWQ1, J07HWQ2 and J07HQX50) and *Halonotius* (J07HN4 and J07HN6), among others, the strain variation present in January is similar to that in August. [129]. For organisms such as the *Nanohaloarchaea* (J07AB56 and J07AB43), the differences suggest the opposite, that there are difference in the strain diversity present in January versus August. By focusing on the genes that are different between sampling months, we can look in more in detail at this variation, providing examples of some of the observed differences.

In the case of J07HWQ1, the single gene that is unique in the August sample codes for a predicted cobalamin biosynthesis protein, based on its COG annotation, and has a von Willebrand factor, type A domain. This protein family

has a wide range of functions and, in the case of the J07HWQ1 genome, is located next to predicted ATPase associated diverse cellular activities. The annotation of this protein does not suggest a particular role for the positive selected gene as it may fall into a wide variety of roles, including DNA repair, transcription, ribosomal and membrane transport, and cell adhesion, among other roles [134, 67].

For J07HWQ2, the gene that is under positive selection in January codifies for an hypothetical protein that is rich in serine residues. No function is associated with this sequence, but recent evidence from studies in Dictyostelid amoebae, suggest that these domains are frequently present and functional in eukaryote genomes [119]. In the August sample, the gene showing evidence of positive selection codes for an amino acid transporter.

The other example that will be explored here are the genomes of the *Nanohaloarchaea* (J07AB56 and J07AB43), where we observe differences in the genes that are under positive selection. In J07AB56, the January genes code for hypothetical proteins, transcriptional regulators, and chaperones, among other functions. Interestingly, a large number of proteins related to transmembrane functions was found, including transporters and transmembrane proteins. In contrast, the functions enriched in August, tend to be more hypothetical proteins and some proteins involved in beta-lactamase functions.

For J07AB43, the differences are hard to pinpoint to specific functional roles as in both cases over 70% of the functions are annotated as hypothetical proteins. An interesting protein that is under selection in the January sample encodes for a photolyase involved in light-driven DNA repair [131]. This could be related to the different environmental condition, particularly light intensity and temperature, between the two sampling dates.

Differences between pN/pS values between sampling dates

Even when the same genes appeared to be under positive selection for some of the analyzed genomes comparing the January and August samples, even in genes that are under selection in both samples, differences in their pN/pS values (Figure 5.15) can be observed. In some of the organisms, such as the *Haloquadratum*

genomes, only a few genes show different values in one of the samples. On the other hand, in the *Nanohaloarchaea*, there is a broad dispersion of the values, indicating differences in the nucleotide diversity of some of these genes, which could be translated into the presence of multiple strains for these organisms. This also supports the idea that the J07AB43 populations that are present in the January sample are different from the ones present in the August sample [129]. A similar pattern is observed in the J07HB67 genome and to a lower degree in the A07HN63 genome.

pN/pS values and functional classification

One of the limitations of the analysis of pN/pS values is that unless experimental evidence is available for some of the candidate genes, in most of the cases, the associated functions tend to be hypothetical proteins [117]. At the same time, this is one of the powerful things in the analysis as it can provide candidates for further experimental studies, at least where isolates are available [37].

To narrow some of the functions that could be under selection, the COG (cluster of orthologous groups) functional categories were used. All the genes that had a COG number assigned were analyzed, and COG categories that could be enriched in genes under selection were evaluated by comparing them with the classification of the complete genome (Appendix C). These results were validated by calculating the odds ratio (the difference between the categories with genes under selection versus the rest of the annotated cogs for the genome), with a one-tailed Fisher exact test (pvalue < 0.05). This analysis provided some functional information on some of the genomes and the genes that are under selection.

For example, in the case of J07HQX50, the cell wall/membrane/envelope biogenesis category has three genes that are under positive selection. Looking at these genes in more detail, all of them are involved in lipopolysaccharide synthesis, two of them with functions related to the surface layer modification, and the other as a sugar transferase. Considering that these genes are under selection both in January and August, these could suggest that J07HQX50 is under some type of selective pressure to modify its membrane composition, which could be due to

multiple factors, including possible phage predation [104].

Another interesting example can be found on J07AB56 (Table C.7), where in the January sample, the cell motility category is enriched in genes under positive selection. Looking in detail, there are four genes in this group, and all of them are encoding secretion system proteins. Three of them encode for VirB11 components of the type IV secretory pathway, which could play a role in DNA uptake or protein secretion [17]. In the August sample, a different category is enriched, intracellular trafficking, secretion, and vesicular transport, but a detailed look of the genes, shows that most of them encode for type IV secretory systems or protein export component, suggesting a similar role as in the January case, where the genes involved in protein secretion are under positive selection. It has been observed in some Bacteria that secreted proteins have rapid evolution rates [80], and some of these proteins could be involved with interaction with the outside environment, as they could be attached to the cell wall and include functions such as protection against grazing [71] and/or competition against other microbial groups [56].

Figure 5.13: Overview of the strategy used to quantify the SNPs differences on each gene, and calculate the ratio of non-synonymous to synonymous substitutions (pN/pS) on each of the reference genome. Figure based on [117].

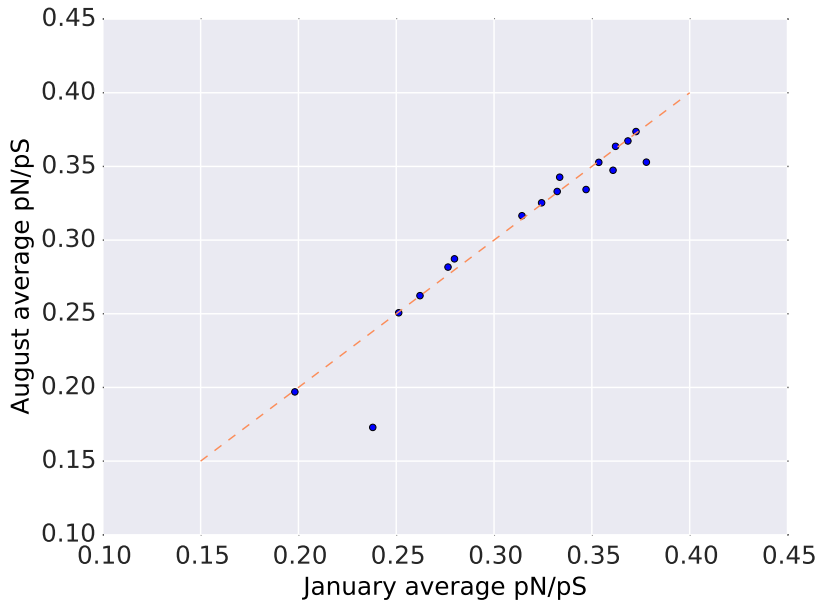


Figure 5.14: Scatterplot of average pN/pS ratios in January versus August, for each of the reference genomes used in the analysis. The extreme value in the right bottom of the plot, corresponds to G22.

Table 5.7: Count of Genes under positive selection (pN/pS > 1). Data where pS/pS = 0/0 or pS = 0, was not included.

Genome	**CDS**	**January**	**August**	**Unique (Jan/Aug)**
J07HWQ1	3,584	61 (1.7)	62 (1.7)	0/1
J07HWQ2	3,856	46 (1.2)	46 (1.19)	1/1
J07HQX50	2,872	19 (0.7)	19 (0.7)	0/0
J07AB56	1,411	96 (6.8)	79 (5.6)	37/20
J07AB43	1,678	96 (5.7)	87 (5.2)	25/16
J07HN4	3,230	198 (6.1)	196 (6.1)	3/1
J07HN6	2,914	59 (2.1)	64 (2.2)	3/8
J07HX64	3049	191 (6.2)	189 (6.2)	3/1
J07HX5	2,139	137 (6.4)	137 (6.4)	1/1
J07HB67	2,847	196 (6.8)	168 (5.9)	54/26
J07HR59	1,841	27 (1.4)	33 (1.8)	0/6
A07HB70	2,514	154 (6.1)	156 (6.2)	1/3
A07HR67	2,891	167 (5.8)	171 (5.9)	0/4
A07HN63	2,507	42 (1.7)	50 (2.0)	4/12
A07HR60	2,861	83 (2.9)	87 (3.1)	2/6
G22	3,525	0 (0)	0 (0)	1/1
J07SB	1,641	91 (5.6)	91 (5.5)	2/2

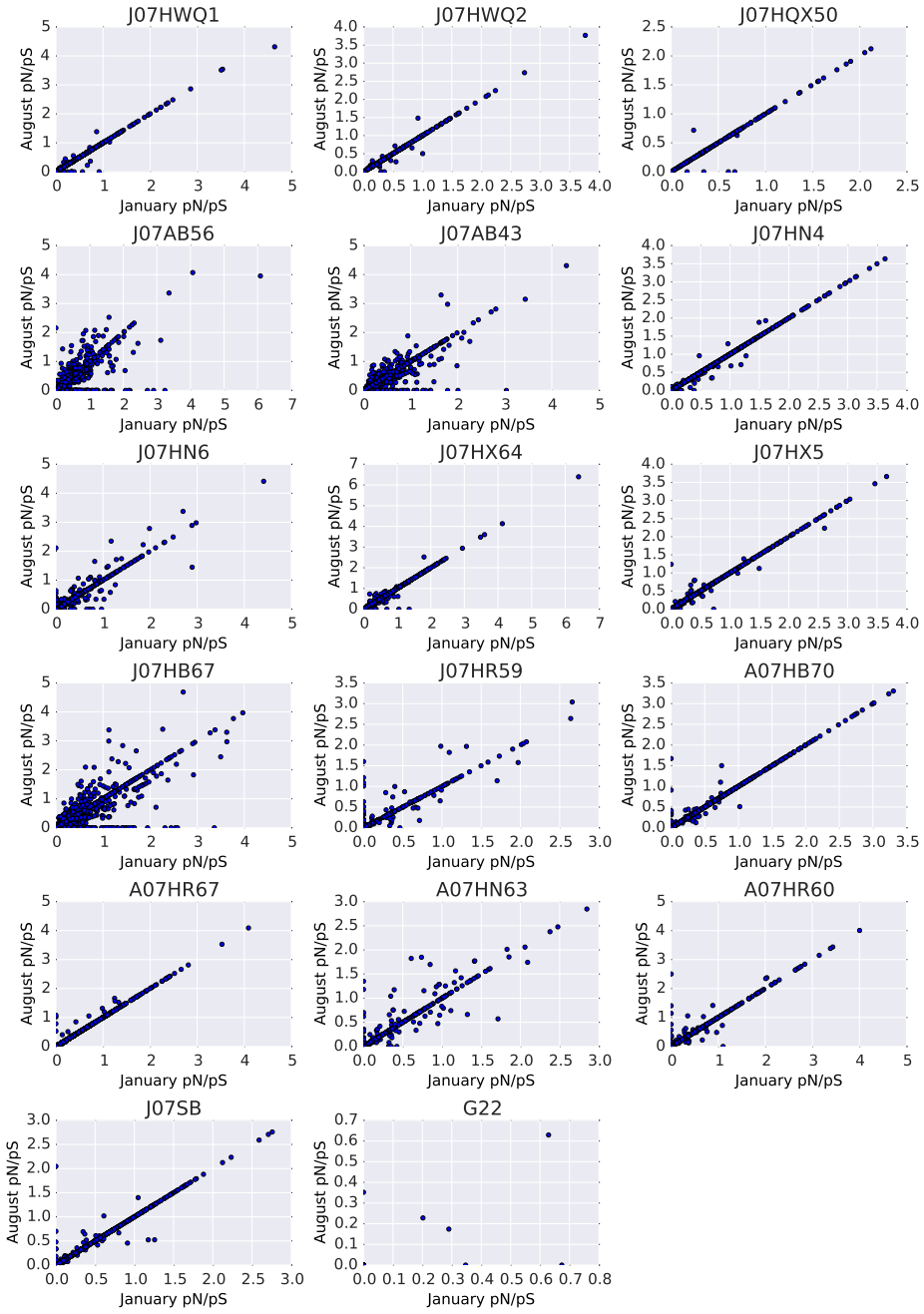


Figure 5.15: Scatterplot of the pN/pS values for each of the reference genomes, comparing the pN/pS values in January versus August.

Figure 5.16: pN/pS values for each gene in the J07HWQ1 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample.

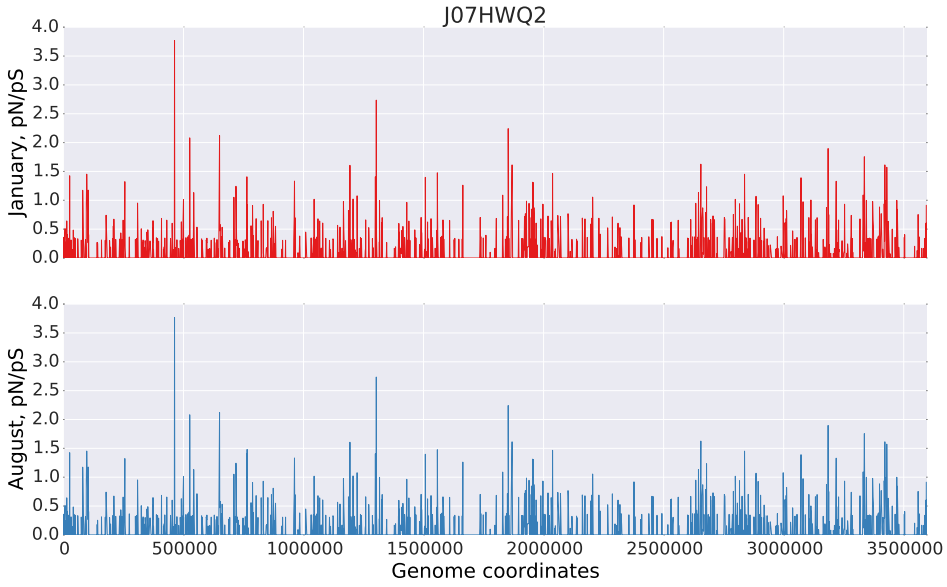


Figure 5.17: pN/pS values for each gene in the J07HWQ2 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample.

5.5 General Discussion

The main goal in this chapter was to provide a big-picture analysis of the relative abundance and fine-scale genetic diversity that is present in the Lake Tyrrell microbial community. The resulting data shows that the combination of high-throughput data with habitat-specific genomes assembled from metagenomic datasets from the same environment [79, 96, 95] provides an opportunity to evaluate the relative abundance of the members of the microbial community and allows the exploration of the genetic diversity that exists in the population.

One of the limitations of the current approach is the stringent parameters used in the analysis. Depending on the goal in mind, these parameters should be evaluated. A clear example of this is that during the analysis the G22 genome usually recruited a low number of reads (its coverage was no larger than 1.2X). This also highlights the issue of isolates and their real abundance in natural microbial communities. G22 was isolated from water samples collected in August of 2007 and chosen for genome sequencing based on similarities with the A07HB70 genome [124]. Based on the results presented here, it seems that not only is this a different species (maybe even more distant) than A07HB70, but also that its abundance in the Lake Tyrrell community is low.

Another element that needs to be taken in consideration is the viral component of the community. Recently, several assembled viral genomes from members of this community were described [33, 32]. Although the filtering strategy (direct water into a Sterivex) did not enrich for viruses (although large viruses could have been retained), using these genomes will allow us to evaluate the relative abundance of the viruses within the sequence dataset and provide an overview of the fine-scale genetic diversity that could be present on these phage genomes. It is possible that the low variation between the January and August samples, which was observed in the archaeal and bacterial genomes, is different in viral populations, as it has been suggested in other hypersaline ecosystems [103].

One of the most interesting results is the possibility that some of the populations that are present in the January community are the same that are present in August, such as the *Haloquadratum* groups, and what it changes over time is

the relative abundance of the genomes. To validate these observations, a more detailed population analysis needs to be performed [107, 58, 132], including detailed population metrics such as nucleotide diversity within and between populations.

A large percentage of the genes found to be under positive selection were hypothetical proteins, similar to what has been observed in similar studies [117, 43]. The potential of this type of analysis, both in isolate genomes and metagenomes, is that it provides a list of candidate genes for study with the idea that even if the candidate is a hypothetical protein, the evidence that is under selection can be used to prioritize it for further studies. Along this line, future work will include a more detailed analysis of each of the genes, including hypothetical proteins. For example, the incorporation of structure modelling may provide some additional information by mapping those sites that are under selection to the three-dimensional structure of the protein. In addition, this list can constitute possible markers to look for in further metagenomic surveys of this community.

One element that was not incorporated in the current analysis is the role of horizontal gene transfer and its relationship to natural selection [135]. It has been suggested that horizontally transferred genes evolve faster than duplicated genes in Bacteria [120]. This suggests that some of the genes that were under positive selection could have been acquired from other organisms.

Moving from away from the broad comparison, the attention can also be focused on individual species groups. One of the best candidates for this analysis is *Haloquadratum*. Besides the two genomes assembled from the Lake Tyrrell community, there are two genome sequences from isolates that are publicly available [8, 28]. A pan-genome analysis of the four genomes could help to explain some of the results for the positive selected genes. For example, it has been suggested that core genes have lower rates of selection compared to non-core genes [51, 105].

Finally, most of the effort was focused on the study of the positively selected genes as the results provide a small dataset to focus on. However, also looking at those genes under purifying selection could provide interesting information about the organisms. It should be expected that housekeeping genes will be under purifying selection as any changes could lead to lethal mutations for the organisms

[88, 107]. Also, in *Eschericha coli*, it has been shown that synonymous mutations alter the rates of gene expression, so it is not only the role of natural selection on changing the amino acid sequence of a protein, but also on how it affects the expression of the gene.

5.6 Conclusions

The results presented in this chapter provide the preliminary framework of analysis to quantify and evaluate the relative abundance and fine-scale genetic variation that is present in natural microbial communities. The genes and functions identified in this chapter as being under positive selection can be used in the future as markers for the study of microbial communities, particularly for future work in the Lake Tyrrell ecosystem.

The methods and results described here represent one of the few studies where habitat-specific genomes are used to quantify the genetic diversity present in a community. This work provides the first overview of this genetic diversity, generating the necessary data that will be required for further detailed analysis. In addition, the methods developed here can be applied to other microbial communities and are not limited to the study of the Lake Tyrrell hypersaline community.

5.7 Acknowledgments

I would like to thank Sheila Podell for discussions on the analysis and ideas for this Chapter. Also, funding from Fulbright-Conicyt and NSF 1149552 is acknowledged, and to Amazon Web Services for an education grant that allowed the use of the EC2 infrastructure.

References

[1] M. Albertsen, P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson, and P. H. Nielsen. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, May 2013.

[2] M. Albertsen, P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson, and P. H. Nielsen. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, 31(6):533–538, May 2013.

[3] E. E. Allen and J. F. Banfield. Community genomics in microbial ecology and evolution. *Nature Reviews Microbiology*, 3(6):489–498, June 2005.

[4] E. E. Allen, G. W. Tyson, R. J. Whitaker, J. C. Detter, P. M. Richardson, and J. F. Banfield. Genome dynamics in a natural archaeal population. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6):1883–1888, Feb. 2007.

[5] R. I. Amann, W. Ludwig, and K. H. Schleifer. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews*, 59(1):143–169, Mar. 1995.

[6] A.-Ş. Andrei, H. L. Banciu, and A. Oren. Living with salt: metabolic and phylogenetic diversity of archaea inhabiting saline ecosystems. *FEMS Microbiology Letters*, 330(1):1–9, May 2012.

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

[8] H. Bolhuis, P. Palm, A. Wende, M. Falb, M. Rampp, F. Rodriguez-Valera, F. Pfeiffer, and D. Oesterhelt. The genome of the square archaeon Haloquadratum walsbyi : life at the limits of water activity. *BMC Genomics*, 7(1):169, July 2006.

[9] A. Brady and S. Salzberg. PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nature Methods*, 8(5):367–367, May 2011.

[10] L. Bragg and G. W. Tyson. Metagenomics using next-generation sequencing. *Methods in molecular biology (Clifton, N.J.)*, 1096:183–201, 2014.

[11] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, S. B. Jovanovich, P. S. Krstic, S. Lindsay, X. S. Ling, C. H. Mastrangelo, A. Meller, J. S. Oliver, Y. V. Pershin, J. M. Ramsey, R. Riehn, G. V. Soni, V. Tabard-Cossa, M. Wanunu, M. Wiggin, and J. A. Schloss. The potential and challenges of nanopore sequencing. *Nature Biotechnology*, 26(10):1146–1153, Oct. 2008.

[12] L. S. Brown. Eubacterial Rhodopsins – Unique Photosensors and Diverse Ion Pumps. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, June 2013.

[13] J. Brulc, D. Antonopoulos, M. Berg Miller, M. Wilson, A. Yannarell, E. Dinsdale, R. Edwards, E. Frank, J. Emerson, P. Wacklin, P. Coutinho, B. Henrissat, K. Nelson, and B. White. Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proceedings of the National Academy of Sciences of the United States of America*, Jan. 2009.

[14] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. P. n. a, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, Apr. 2010.

[15] J. G. Caporaso, C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone, P. J. Turnbaugh, N. Fierer, and R. Knight. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences*, 108 Suppl 1:4516–4522, Mar. 2011.

[16] E. O. Casamayor, R. Massana, S. Benlloch, L. Øvreås, B. Díez, V. J. Goddard, J. M. Gasol, I. Joint, F. Rodriguez-Valera, and C. Pedros-Alio. Changes in archaeal, bacterial and eukaryal assemblages along a salinity gradient by comparison of genetic fingerprinting methods in a multipond solar saltern. *Environmental Microbiology*, 4(6):338–348, June 2002.

[17] V. Chandran, R. Fronzes, S. Duquerroy, N. Cronin, J. Navaza, and G. Waksman. Structure of the outer membrane complex of a type IV secretion system. *Nature*, 462(7276):1011–1015, Nov. 2009.

[18] P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, Apr. 2012.

[19] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25(11):1422–1423, May 2009.

[20] O. X. Cordero, H. Wildschutte, B. Kirkup, S. Proehl, L. Ngo, F. Hussain, F. Le Roux, T. Mincer, and M. F. Polz. Ecological Populations of Bacteria Act as Socially Cohesive Units of Antibiotic Production and Resistance. *Science*, 337(6099):1228–1231, Sept. 2012.

[21] T. P. Curtis, W. T. Sloan, and J. W. Scannell. Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences of the United States of America*, 99(16):10494–10499, 2002.

[22] A. E. Darling, G. Jospin, E. Lowe, F. A. Matsen, IV, H. M. Bik, and J. A. Eisen. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, 2:e243, Jan. 2014.

[23] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, Apr. 2011.

[24] S. C. Di Rienzi, I. Sharon, K. C. Wrighton, O. Koren, L. A. Hug, B. C. Thomas, J. K. Goodrich, J. T. Bell, T. D. Spector, J. F. Banfield, R. E. Ley, and R. Kolter. The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *eLife*, 2, Oct. 2013.

[25] E. A. Dinsdale, R. A. Edwards, D. Hall, F. Angly, M. Breitbart, J. M. Brulc, M. Furlan, C. Desnues, M. Haynes, L. Li, L. Mcdaniel, M. A. Moran, K. E. Nelson, C. Nilsson, R. Olson, J. Paul, B. R. Brito, Y. Ruan, B. K. Swan, R. Stevens, D. L. Valentine, R. V. Thurber, L. Wegley, B. A. White, and F. Rohwer. Functional metagenomic profiling of nine biomes. *Nature*, 452(7187):629–632, Apr. 2008.

[26] W. F. Doolittle. Population genomics: how bacterial species form and why they don't exist. *Current biology : CB*, 22(11):R451–3, June 2012.

[27] J. R. Doroghazi and W. W. Metcalf. Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics*, 14(1):611, Sept. 2013.

[28] M. Dyall-Smith, F. Pfeiffer, K. Klee, P. Palm, and K. Gross. Haloquadratum walsbyi: Limited Diversity in a Global Pond. *PLoS ONE*, 2011.

[29] M. Dyall-Smith, S.-L. Tang, and C. Bath. Haloarchaeal viruses: how diverse are they? *Research in Microbiology*, 154(4):309–313, May 2003.

[30] R. Egea, S. Casillas, and A. Barbadilla. Standard and generalized McDonald-Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Research*, 36(Web Server issue):W157–62, July 2008.

[31] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. deWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910):133–138, Jan. 2009.

[32] J. B. Emerson, K. Andrade, B. C. Thomas, A. Norman, E. E. Allen, K. B. Heidelberg, and J. F. Banfield. Virus-Host and CRISPR Dynamics in Archaea-Dominated Hypersaline Lake Tyrrell, Victoria, Australia. *Archaea (Vancouver, BC)*, 2013(2):1–12, 2013.

[33] J. B. Emerson, B. C. Thomas, K. Andrade, E. E. Allen, K. B. Heidelberg, and J. F. Banfield. Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. *Applied and Environmental Microbiology*, 78(17):6309–6320, Sept. 2012.

[34] J. B. Emerson, B. C. Thomas, K. Andrade, K. B. Heidelberg, and J. F. Banfield. New approaches indicate constant viral diversity despite shifts in assemblage structure in an Australian hypersaline lake. *Applied and Environmental Microbiology*, 79(21):6755–6764, Nov. 2013.

[35] W. G. Feero, A. E. Guttmacher, and D. A. Relman. Microbial Genomics and Infectious Diseases. *The New England journal of medicine*, 365(4):347–357, July 2011.

[36] J. J. Flowers, S. He, S. Malfatti, T. G. del Rio, S. G. Tringe, P. Hugenholtz, and K. D. Mcmahon. Comparative genomics of two 'Candidatus Accumulibacter' clades performing biological phosphorus removal. *The ISME Journal*, July 2013.

[37] W. F. Fricke, M. K. Mammel, P. F. McDermott, C. Tartera, D. G. White, J. E. LeClerc, J. Ravel, and T. A. Cebula. Comparative genomics of 28 Salmonella enterica isolates: evidence for CRISPR-mediated adaptive sublineage evolution. *Journal Of Bacteriology*, 193(14):3556–3568, July 2011.

[38] E. Garrison and G. Marth. Haplotype-based variant detection from short-read sequencing. July 2012.

[39] R. Ghai, C. M. Hernandez, A. Picazo, C. M. Mizuno, K. Ininbergs, B. Díez, R. Valas, C. L. Dupont, K. D. Mcmahon, A. Camacho, and F. Rodriguez-Valera. Metagenomes of Mediterranean Coastal Lagoons. *Scientific Reports*, 2:–, July 2012.

[40] R. Ghai, L. Pasić, A. B. Fernández, A.-B. Martin-Cuadrado, C. M. Mizuno, K. D. Mcmahon, R. T. Papke, R. Stepanauskas, B. Rodriguez-Brito, F. Rohwer, C. Sánchez-Porro, A. Ventosa, and F. Rodriguez-Valera. New abundant microbial groups in aquatic hypersaline environments. *Scientific Reports*, 1:135, 2011.

[41] Y. H. Grad, P. Godfrey, G. C. Cerquiera, P. Mariani-Kurkdjian, M. Gouali, E. Bingen, T. P. Shea, B. J. Haas, A. Griggs, and S. Young. Comparative genomics of recent Shiga toxin-producing Escherichia coli O104: H4: short-term evolution of an emerging pathogen. *mBio*, 4(1), 2013.

[42] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, 5(10):R245–9, Oct. 1998.

[43] C. L. Hemme, Y. Deng, T. J. Gentry, M. W. Fields, L. Wu, S. Barua, K. Barry, S. G. Tringe, D. B. Watson, Z. He, T. C. Hazen, J. M. Tiedje, E. M. Rubin, and J. Zhou. Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. *The ISME Journal*, 4(5):660–672, Feb. 2010.

[44] D. P. R. Herlemann, D. Lundin, M. Labrenz, K. Jürgens, Z. Zheng, H. Aspeborg, and A. F. Andersson. Metagenomic de novo assembly of an aquatic representative of the verrucomicrobial class Spartobacteria. *mBio*, 4(3):e00569–12, 2013.

[45] A. C. Howe, J. K. Jansson, S. A. Malfatti, S. G. Tringe, J. M. Tiedje, and C. T. Brown. Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences*, Mar. 2014.

[46] J. D. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, May 2007.

[47] L. D. Hurst. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in Genetics*, 18(9):486–487, Sept. 2002.

[48] B. L. Hurwitz and M. B. Sullivan. The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology. *PLoS ONE*, 8(2):e57355, Feb. 2013.

[49] E. Jaspers and J. Overmann. Ecological Significance of Microdiversity: Identical 16S rRNA Gene Sequences Can Be Found in Bacteria with Highly Divergent Genomes and Ecophysiologies. *Applied and Environmental Microbiology*, 70(8):4831–4839, Aug. 2004.

[50] C. Jones, E. Allen, J. Giska, S. Welch, D. Kirste, and J. Banfield. Iron formations at Lake Tyrrell, Victoria, Australia: Microbially-mediated redox chemistry. *Geochimica et Cosmochimica Acta*, 70(18):A297–A297, Aug. 2006.

[51] I. K. Jordan, I. B. Rogozin, Y. Wolf, and E. Koonin. Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. *Genome Research*, 12(6):962–968, May 2002.

[52] K. S. Kakirde, L. C. Parsley, and M. R. Liles. Size does matter: Application-driven approaches for soil metagenomics. *Soil Biology and Biochemistry*, 42(11):1911–1923, Nov. 2010.

[53] J. Kallmeyer, R. Pockalny, R. R. Adhikari, D. C. Smith, and S. D'Hondt. Global distribution of microbial abundance and biomass in subseafloor sediment. *Proceedings of the National Academy of Sciences of the United States of America*, 109(40):16213–16216, 2012.

[54] R. S. Kantor, K. C. Wrighton, K. M. Handley, I. Sharon, L. A. Hug, C. J. Castelle, B. C. Thomas, and J. F. Banfield. Small Genomes and Sparse Metabolisms of Sediment-Associated Bacteria from Four Candidate Phyla. *mBio*, 4(5):e00708–13–e00708–13, Aug. 2013.

[55] W. C. N. J. B. H. N. E. K. A. J. B. E. Karla B Heidelberg. Characterization of eukaryotic microbial diversity in hypersaline Lake Tyrrell, Australia. *Frontiers in Microbiology*, 4, 2013.

[56] B. C. Kirkup and M. A. Riley. Antibiotic-mediated antagonism leads to a bacterial game of rock–paper–scissors in vivo. *Nature*, 428(6981):412–414, Mar. 2004.

[57] R. Knight, P. Maxwell, A. Birmingham, J. Carnes, J. G. Caporaso, B. C. Easton, M. Eaton, M. Hamady, H. Lindsay, Z. Liu, C. Lozupone, D. McDonald, M. Robeson, R. Sammut, S. Smit, M. J. Wakefield, J. Widmann, S. Wikman, S. Wilson, H. Ying, and G. A. Huttley. PyCogent: a toolkit for making sense from sequence. *Genome Biology*, 8(8):R171, 2007.

[58] S. Kryazhimskiy and J. B. Plotkin. The population genetics of dN/dS. *PLoS genetics*, 4(12):e1000304, Dec. 2008.

[59] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, Mar. 2012.

[60] B. A. Legault, A. Lopez-Lopez, J. C. Alba-Casado, W. F. Doolittle, H. Bolhuis, F. Rodriguez-Valera, and R. T. Papke. 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, 2006.

[61] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, Aug. 2009.

[62] D. Lindell, J. D. Jaffe, Z. I. Johnson, G. M. Church, and S. W. Chisholm. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, 438(7064):86–89, Oct. 2005.

[63] I. Lo, V. J. Denef, N. C. Verberkmoes, M. B. Shah, D. Goltsman, G. Dibartolo, G. W. Tyson, E. E. Allen, R. J. Ram, J. C. Detter, P. Richardson, M. P. Thelen, R. L. Hettich, and J. F. Banfield. Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature*, 446(7135):537–541, Mar. 2007.

[64] D. T. Long, N. E. Fegan, W. B. Lyons, M. E. Hines, P. G. Macumber, and A. M. Giblin. Geochemistry of acid brines: Lake Tyrrell, Victoria, Australia. *Chemical Geology*, 96(1-2):33–52, Mar. 1992.

[65] M. López-Pérez, R. Ghai, M. Leon, Á. Rodríguez-Olmos, J. Copa-Patiño, J. Soliveri, C. Sánchez-Porro, A. Ventosa, and F. Rodriguez-Valera. Genomes of "Spiribacter", a streamlined, successful halophilic bacterium. *BMC Genomics*, 14(1):787, 2013.

[66] P. G. Macumber. Hydrological processes in the Tyrrell Basin, southeastern Australia. *Chemical Geology*, 96(1):1–18, 1992.

[67] K. S. Makarova and E. V. Koonin. Two new families of the FtsZ-tubulin protein superfamily implicated in membrane remodeling in diverse bacteria and archaea. *Biology Direct*, 5(1):33, 2010.

[68] E. R. Mardis. Next-generation DNA sequencing methods. *Annual Review Of Genomics And Human Genetics*, 9:387–402, Jan. 2008.

[69] E. R. Mardis. Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, 6(1):287–303, June 2013.

[70] F. A. Matsen and S. N. Evans. Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. *PLoS ONE*, 8(3):e56859, 2013.

[71] C. Matz and S. Kjelleberg. Off the hook – how bacteria survive protozoan grazing. *Trends in Microbiology*, 13(7):302–307, July 2005.

[72] D. McDonald, M. N. Price, J. Goodrich, E. P. Nawrocki, T. Z. DeSantis, A. Probst, G. L. Andersen, R. Knight, and P. Hugenholtz. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, 6(3):610–618, Dec. 2011.

[73] J. H. McDonald and M. Kreitman. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351(6328):652–654, June 1991.

[74] T. J. McGenity, R. T. Gemmell, W. D. Grant, and H. Stan-Lotter. Origins of halophilic microorganisms in ancient salt deposits. *Environmental Microbiology*, 2(3):243–250, June 2000.

[75] W. McKinney. Data structures for statistical computing in python. In S. van der Walt and J. Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[76] E. F. Mongodin, K. E. Nelson, S. Daugherty, R. DeBoy, J. Wister, H. Khouri, J. Weidman, D. A. Walsh, R. T. Papke, G. Sanchez Perez, A. K. Sharma, C. Nesbø, D. MacLeod, E. Bapteste, W. F. Doolittle, R. L. Charlebois, B. Legault, and F. Rodriguez-Valera. The genome of Salinibacter ruber: Convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proceedings of the National Academy of Sciences*, 102(50):18147–18152, Dec. 2005.

[77] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, July 2008.

[78] M. B. Mutlu, M. Martínez-García, F. Santos, A. Peña, K. Guven, and J. Antón. Prokaryotic diversity in Tuz Lake, a hypersaline environment in Inland Turkey. *FEMS Microbiology Ecology*, 65(3):474–483, Sept. 2008.

[79] P. Narasingarao, S. Podell, J. A. Ugalde, C. Brochier-Armanet, J. B. Emerson, J. J. Brocks, K. B. Heidelberg, J. F. Banfield, and E. E. Allen. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *The ISME Journal*, 6(1):81–93, Jan. 2012.

[80] T. Nogueira, M. Touchon, and Eduardo P. C. Rocha. Rapid Evolution of the Sequences and Gene Repertoires of Secreted Proteins in Bacteria. *PLoS ONE*, 7(11):e49403, Nov. 2012.

[81] T. E. Oliphant. Python for scientific computing. *Computing in Science & Engineering*, 9(3):10–20, 2007.

[82] A. Oren. *Halophilic Microorganisms and Their Environments*. Kluwer Academic Publishers, Aug. 2002.

[83] A. Oren. Microbial life at high salt concentrations: phylogenetic and metabolic diversity. *Saline Systems*, 4:2, Jan. 2008.

[84] A. Oren. Saltern evaporation ponds as model systems for the study of primary production processes under hypersaline conditions. *Aquatic Microbial Ecology*, 56:193–204, Sept. 2009.

[85] A. Oren. Approaches Toward the Study of Halophilic Microorganisms in Their Natural Environments: Who Are They and What Are They Doing? In *link.springer.com*, pages 1–33. Springer Netherlands, Dordrecht, Dec. 2012.

[86] A. Oren. Life at High Salt Concentrations. In *link.springer.com*, pages 421–440. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[87] A. Oren. Salinibacter: an extremely halophilic bacterium with archaeal properties Aharon Oren. *FEMS Microbiology Letters*, pages n/a–n/a, Feb. 2013.

[88] B. Palenik, Q. Ren, V. Tai, and I. T. Paulsen. Coastal Synechococcus metagenome reveals major roles for horizontal gene transfer and plasmids in population diversity. *Environmental Microbiology*, 11(2):349–359, Feb. 2009.

[89] D. H. Parks, N. J. MacDonald, and R. G. Beiko. Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinformatics*, 12:328, 2011.

[90] L. Pašić, B. Rodriguez-Mueller, A.-B. Martin-Cuadrado, A. Mira, F. Rohwer, and F. Rodriguez-Valera. Metagenomic islands of hyperhalophiles: the case of Salinibacter ruber. *BMC Genomics*, 10(1):570, Dec. 2009.

[91] A. Pati, L. S. Heath, N. C. Kyrpides, and N. Ivanova. ClaMS: A Classifier for Metagenomic Sequences. *Standards in Genomic Sciences*, 5(2):248–253, Nov. 2011.

[92] J. Pell, A. Hintze, R. Canino-Koning, A. Howe, J. M. Tiedje, and C. T. Brown. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proceedings of the National Academy of Sciences*, 109(33):13272–13277, July 2012.

[93] A. Peña, H. Teeling, J. Huerta-Cepas, F. Santos, P. Yarza, J. Brito-Echeverría, M. Lucio, P. Schmitt-Kopplin, I. Meseguer, and C. Schenowitz. Fine-scale evolution: genomic, phenotypic and ecological differentiation in two coexisting Salinibacter ruber strains. *The ISME Journal*, 4(7):882–895, 2010.

[94] F. Perez and B. E. Granger. IPython: a system for interactive scientific computing. *Computing in Science & Engineering*, 9(3):21–29, 2007.

[95] S. Podell, J. B. Emerson, C. M. Jones, J. A. Ugalde, S. Welch, K. B. Heidelberg, J. F. Banfield, and E. E. Allen. Seasonal fluctuations in ionic concentrations drive microbial succession in a hypersaline lake community. *The ISME Journal*, Dec. 2013.

[96] S. Podell, J. A. Ugalde, P. Narasingarao, J. F. Banfield, K. B. Heidelberg, and E. E. Allen. Assembly-Driven Community Genomics of a Hypersaline Microbial Ecosystem. *PLoS ONE*, 8(4):e61692, Apr. 2013.

[97] K. Porter, B. E. Russ, and M. L. Dyall-Smith. Virus–host interactions in salt lakes. *Current Opinion in Microbiology*, 10(4):418–424, Aug. 2007.

[98] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–842, Mar. 2010.

[99] M. S. Rappé and S. J. Giovannoni. The uncultured microbial majority. *Annual Review of Microbiology*, 57:369–394, Jan. 2003.

[100] D. C. Reed, C. K. Algar, J. A. Huber, and G. J. Dick. Gene-centric approach to integrating environmental genomics and biogeochemical models. *Proceedings of the National Academy of Sciences*, Jan. 2014.

[101] M. L. Reno, N. L. Held, C. J. Fields, P. V. Burke, and R. J. Whitaker. Biogeography of the Sulfolobus islandicus pan-genome. *Proceedings of the National Academy of Sciences of the United States of America*, 106(21):8605–8610, May 2009.

[102] C. Rinke, P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J.-F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W.-T. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz, and T. Woyke. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–437, 2013.

[103] B. Rodriguez-Brito, L. Li, L. Wegley, M. Furlan, F. Angly, M. Breitbart, J. Buchanan, C. Desnues, E. Dinsdale, R. Edwards, B. Felts, M. Haynes, H. Liu, D. Lipson, J. Mahaffy, A. B. Martin-Cuadrado, A. Mira, J. Nulton, L. Pasić, S. Rayhawk, J. Rodriguez-Mueller, F. Rodriguez-Valera, P. Salamon, S. Srinagesh, T. F. Thingstad, T. Tran, R. V. Thurber, D. Willner, M. Youle, and F. Rohwer. Viral and microbial community dynamics in four aquatic environments. *The ISME Journal*, 4(6):739–751, June 2010.

[104] F. Rodriguez-Valera, A.-B. Martin-Cuadrado, B. Rodriguez-Brito, L. Pasić, T. F. Thingstad, F. Rohwer, and A. Mira. Explaining microbial population genomics through phage predation. *Nature Reviews Microbiology*, 7(11):828–836, Nov. 2009.

[105] F. Rodriguez-Valera and D. W. Ussery. Is the pan-genome also a pan-selectome? *F1000 Research*, Sept. 2012.

[106] D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y.-H. Rogers, L. I. Falcón, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology*, 5(3):e77, Jan. 2007.

[107] S. Schloissnig, M. Arumugam, S. Sunagawa, M. Mitreva, J. Tap, A. Zhu, A. Waller, D. R. Mende, J. R. Kultima, J. Martin, K. Kota, S. R. Sunyaev,

G. M. Weinstock, and P. Bork. Genomic variation landscape of the human gut microbiome. *Nature*, 493(7430):45–50, Jan. 2013.

[108] P. D. Schloss and J. Handelsman. Status of the microbial census. *Microbiology and molecular biology reviews : MMBR*, 68(4):686–+, Dec. 2004.

[109] N. Segata, D. Börnigen, X. C. Morgan, and C. Huttenhower. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications*, 4:2304, 2013.

[110] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, pages –, June 2012.

[111] B. J. Shapiro and M. F. Polz. Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in Microbiology*, Mar. 2014.

[112] A. Sharma, J. Spudich, and W. Doolittle. Microbial rhodopsins: functional versatility and genetic mobility. *Trends in Microbiology*, 14(11):463–469, Nov. 2006.

[113] J. E. Sherwood, F. Stagnitti, M. J. Kokkinn, and W. D. Williams. Dissolved oxygen concentrations in hypersaline waters. *Limnology And Oceanography*, pages 235–250, 1991.

[114] S. L. Simmons, G. Dibartolo, V. J. Denef, D. S. A. Goltsman, M. P. Thelen, and J. F. Banfield. Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. *PLoS Biology*, 6(7):e177, July 2008.

[115] J. T. Staley and A. Konopka. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Reviews in Microbiology*, 39(1):321–346, 1985.

[116] E. J. Stewart. Growing unculturable bacteria. *Journal Of Bacteriology*, 194(16):4151–4160, Aug. 2012.

[117] V. Tai, A. F. Y. Poon, I. T. Paulsen, and B. Palenik. Selection in Coastal Synechococcus (Cyanobacteria) Populations Evaluated from Environmental Metagenomes. *PLoS ONE*, 6(9):e24249, Sept. 2011.

[118] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41, Sept. 2003.

[119] X. Tian, J. E. Strassmann, and D. C. Queller. A conserved extraordinarily long serine homopolymer in Dictyostelid amoebae. *Heredity*, 112(2):215–218, Oct. 2013.

[120] T. J. Treangen and Eduardo P. C. Rocha. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. *PLoS genetics*, 7(1):e1001284, Jan. 2011.

[121] G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, Mar. 2004.

[122] G. W. Tyson, I. Lo, B. J. Baker, E. E. Allen, P. Hugenholtz, and J. F. Banfield. Genome-directed isolation of the key nitrogen fixer Leptospirillum ferrodiazotrophum sp. nov. from an acidophilic microbial community. *Applied and Environmental Microbiology*, 71(10):6319–6324, Oct. 2005.

[123] J. A. Ugalde, M. J. Gallardo, C. Belmar, P. Muñoz, N. Ruiz-Tagle, S. Ferrada-Fuentes, C. Espinoza, E. E. Allen, and V. A. Gallardo. Microbial Life in a Fjord: Metagenomic Analysis of a Microbial Mat in Chilean Patagonia. *PLoS ONE*, 8(8):e71952, Aug. 2013.

[124] J. A. Ugalde, P. Narasingarao, S. Kuo, S. Podell, and E. E. Allen. Draft Genome Sequence of "Candidatus Halobonum tyrrellensis" Strain G22, Isolated from the Hypersaline Waters of Lake Tyrrell, Australia. *Genome announcements*, 1(6), 2013.

[125] J. A. Ugalde, S. Podell, P. Narasingarao, and E. E. Allen. Xenorhodopsins, an enigmatic new class of microbial rhodopsins horizontally transferred between archaea and bacteria. *Biology Direct*, 6:52, 2011.

[126] J. C. Venter. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, Apr. 2004.

[127] A. Ventosa, M. C. Márquez, C. Sánchez-Porro, and R. Rafael. Taxonomy of Halophilic Archaea and Bacteria. pages 59–80, 2012.

[128] L. Vogeley, O. Sineshchekov, V. Trivedi, J. Sasaki, J. Spudich, and H. Luecke. Anabaena sensory rhodopsin: a photochromic color sensor at 2.0 A. *Science's STKE*, 306(5700):1390, 2004.

[129] M. Vos. A species concept for bacteria based on adaptive divergence. *Trends in Microbiology*, 19(1):1–7, 2011.

[130] D. M. Ward, F. M. Cohan, D. Bhaya, J. F. Heidelberg, M. Kühl, and A. Grossman. Genomics, environmental genomics and the issue of microbial species. *Heredity*, 100(2):207–219, Feb. 2008.

[131] S. Weber. Light-driven enzymatic catalysis of DNA repair: a review of recent biophysical studies on photolyase. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1707(1):1–23, Feb. 2005.

[132] R. J. Whitaker and J. F. Banfield. Population genomics in natural microbial communities. *Trends in ecology & evolution (Personal edition)*, 21(9):508–516, Sept. 2006.

[133] W. B. Whitman, D. C. Coleman, and W. J. Wiebe. Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, 95(12):6578–6583, 1998.

[134] C. A. Whittaker and R. O. Hynes. Distribution and evolution of von Willebrand/integrin A domains: widely dispersed domains with roles in cell adhesion and elsewhere. *Molecular biology of the cell*, 13(10):3369–3387, 2002.

[135] J. Wiedenbeck and F. M. Cohan. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiology Reviews*, 35(5):957–976, July 2011.

[136] T. A. Williams and T. M. Embley. Archaeal "dark matter" and the origin of eukaryotes. *Genome Biology and Evolution*, 6(3):474–481, Mar. 2014.

[137] S. J. Williamson, D. B. Rusch, S. Yooseph, A. L. Halpern, K. B. Heidelberg, J. I. Glass, C. Andrews-Pfannkoch, D. Fadrosh, C. S. Miller, and G. Sutton. The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE*, 3(1):e1456, 2008.

[138] D. Willner, R. Thurber, and F. Rohwer. Metagenomic signatures of 86 microbial and viral metagenomes. *Environmental Microbiology*, Mar. 2009.

[139] P. Wilmes, S. L. Simmons, V. J. Denef, and J. F. Banfield. The dynamic genetic repertoire of microbial communities. *FEMS Microbiology Reviews*, 33(1):109–132, Jan. 2009.

[140] J. C. Wooley, A. Godzik, and I. Friedberg. A Primer on Metagenomics. *PLoS Computational Biology*, 6(2):e1000667, Feb. 2010.

[141] D. Wu, P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, and B. J. Tindall. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, 462(7276):1056–1060, 2009.

[142] Z. Xu, H. Chen, and R. Zhou. Genome-wide evidence for positive selection and recombination in Actinobacillus pleuropneumoniae. *BMC Evolutionary Biology*, 11(1):203, 2011.

[143] Z. Yang and J. P. Bielawski. Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*, 15(12):496–503, Dec. 2000.

Appendix A

Draft Genome Sequence of *Candidatus* Halobonum tyrrellensis Strain G22, Isolated from the Hypersaline Waters of Lake Tyrrell, Australia

genomeA
Journals.ASM.org

Draft Genome Sequence of "*Candidatus* Halobonum tyrrellensis" Strain G22, Isolated from the Hypersaline Waters of Lake Tyrrell, Australia

Juan A. Ugalde,[a] Priya Narasingarao,[a] Sidney Kuo,[b] Sheila Podell,[a] Eric E. Allen[a,b]

Marine Biology Research Division, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California, USA[a]; Division of Biological Sciences, University of California, San Diego, La Jolla, California, USA[b]

We report the draft 3.675-Mbp genome sequence of "*Candidatus* Halobonum tyrrellensis" strain G22, a novel halophilic archaeon isolated from the surface hypersaline waters of Lake Tyrrell, Australia. The availability of the first genome from the "*Candidatus* Halobonum" genus provides a new genomic resource for the comparative genomic analysis of halophilic *Archaea*.

Received 25 October 2013 **Accepted** 11 November 2013 **Published** 12 December 2013

Citation Ugalde JA, Narasingarao P, Kuo S, Podell S, Allen EE. 2013. Draft genome sequence of "*Candidatus* Halobonum tyrrellensis" strain G22, isolated from the hypersaline waters of Lake Tyrrell, Australia. Genome Announc. 1(6):e01001-13. doi:10.1128/genomeA.01001-13.

Address correspondence to Eric E. Allen, eallen@ucsd.edu.

Halophilic *Archaea* of the class *Halobacteria* (phylum *Euryarchaeota*) are dominant members of extreme hypersaline environments worldwide (1). Numerous genera have been isolated from diverse hypersaline habitats, and many representative genome sequences are available (1, 2). However, recent metagenomic analyses of hypersaline ecosystems have revealed that these reference halophiles are not adequately representative of the dominant microbial populations present in many natural hypersaline habitats (3–5). Here, we report the genome sequence of a novel member of the class *Halobacteria* isolated from the hypersaline surface waters of Lake Tyrrell, Victoria, Australia.

Surface water samples were plated onto minimal medium containing a 23% salt solution amended with various carbon substrates, including glycerol, acetate, or glucose, and incubated at room temperature under aerobic conditions. After 3 to 4 weeks of incubation, the colonies were restreaked for purity and characterized via 16S rRNA gene amplification and sequencing to screen for novel species/strains. "*Candidatus* Halobonum tyrrellensis" strain G22 was isolated from minimal medium containing glycerol as the sole carbon source, incubated aerobically at room temperature. Genomic DNA was sequenced using 454 Titanium chemistry at the J. Craig Venter Institute (Rockville, MD). The total number of reads generated was 568,949, with an average length of 428 bp. The sequences were assembled using Newbler (version 2.7), resulting in a total of 72 contigs (N_{50}, 119,067 bp; mean contig length, 45,962 bp; maximum contig length, 303,316 bp), with an estimated genome size of 3,675,087 bp and a G+C content of 70.1%. Functional annotation of predicted gene sequences was performed using the IMG-ER platform (6). A total of 3,525 predicted coding sequences were identified, including 47 tRNAs and a single copy of the rRNA operon.

A phylogenetic tree based on 16S rRNA genes (http://dx.doi.org/10.6084/m9.figshare.830514) suggests that "*Ca*. Halobonum tyrrellensis" is a member of the *Halobacteriaceae* family and a sister group of the *Halobaculum* genus, sharing 92% 16S rRNA gene sequence identity with *Halobaculum gomorrense* (7). A comparison of the "*Ca*. Halobonum tyrrellensis" genome with the partial genome sequence available for *H. gomorrense* (632,433 bp) (8) revealed an average nucleotide identity (ANI) of 81.56 ± 1.18%. A phylogenomic approach using multiple amino acid markers (9) supports the placement of "*Ca*. Halobonum tyrrellensis" as a new genus (http://dx.doi.org/10.6084/m9.figshare.830514). A detailed characterization of the physiology and metabolism of "*Ca*. Halobonum tyrrellensis" and a formal description of this strain are currently in progress.

The features found in the genome include the presence of a putative sensory rhodopsin, a high number of ABC transporters and carbon metabolism genes, including trehalose utilization genes, and the absence of conserved haloarchaeal genes encoding a flagellar system or gas vesicle synthesis proteins.

The "*Ca*. Halobonum tyrrellensis" genome represents the first high-quality draft sequence for a member of the new candidate genus *Halobonum*. These data expand the breadth of the reference genome sequence information for halophilic *Archaea*, providing a new resource for comparative genomic analyses and the phylogenetic binning of metagenomic sequence data recovered from hypersaline environments.

Nucleotide sequence accession number. The draft genome sequence of "*Ca*. Halobonum tyrrellensis" strain G22 is deposited at DDBJ/EMBL/Genbank databases under the accession no. ASGZ00000000.

ACKNOWLEDGMENTS

Funding for this work was provided by NSF award no. 0626526 (to J. F. Banfield, K. B. Heidelberg, and E.E.A.) and NIH award R21HG005107-02 (E.E.A.). J.A.U. was supported by a Fulbright-Conicyt fellowship.

REFERENCES

1. **Andrei AŞ, Banciu HL, Oren A.** 2012. Living with salt: metabolic and phylogenetic diversity of archaea inhabiting saline ecosystems. FEMS Microbiol. Lett. **330:**1–9.
2. **Lynch EA, Langille MG, Darling A, Wilbanks EG, Haltiner C, Shao KS, Starr MO, Teiling C, Harkins TT, Edwards RA, Eisen JA, Facciotti MT.**

2012. Sequencing of seven Haloarchaeal genomes reveals patterns of genomic flux. PLoS One **7:**e41389. doi:10.1371/journal.pone.0041389.

3. **Ghai R, Pasić L, Fernández AB, Martin-Cuadrado A-B, Mizuno CM, Mcmahon KD, Papke RT, Stepanauskas R, Rodriguez-Brito B, Rohwer F, Sánchez-Porro C, Ventosa A, Rodriguez-Valera F.** 2011. New abundant microbial groups in aquatic hypersaline environments. Sci. Rep. **1:**135.
4. **Podell S, Ugalde JA, Narasingarao P, Banfield JF, Heidelberg KB, Allen EE.** 2013. Assembly-driven community genomics of a hypersaline microbial ecosystem. PLoS One **8:**e61692. doi:10.1371/journal.pone.0061692.
5. **Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, Heidelberg KB, Banfield JF, Allen EE.** 2012. *De novo* metagenomic assembly reveals abundant novel major lineage of *Archaea* in hypersaline microbial communities. ISME J. **6:**81–93.
6. **Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC.** 2011. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res. **40:**D115–D122. doi:10.1093/nar/gkr1044.
7. **Oren A, Gurevich P, Gemmell RT, Teske A.** 1995. *Halobaculum gomorrense* gen. nov., sp. nov., a novel extremely halophilic archaeon from the Dead Sea. Int. J. Syst. Bacteriol. **45:**747–754.
8. **Goo YA, Roach J, Glusman G, Baliga NS, Deutsch K, Pan M, Kennedy S, DasSarma S, Ng WV, Hood L.** 2004. Low-pass sequencing for microbial comparative genomics. BMC Genomics **5:**3. doi:10.1186/1471-2164-5-3.
9. **Segata N, Börnigen D, Morgan XC, Huttenhower C.** 2013. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat. Commun. **4:**2304. doi:10.1038/ncomms3304.

Appendix A is a full reprint of: Draft Genome Sequence of "*Candidatus* Halobonum tyrrellensis" Strain G22, Isolated from the Hypersaline Waters of Lake Tyrrell, Australia. J.A. Ugalde, P. Narasingarao, S. Kuo, S. Podell and E.E. Allen. *Genome Announcements*, **1**(6):e01001-13. 2013. With permission from all coauthors.

Appendix B

Genome coverage plots

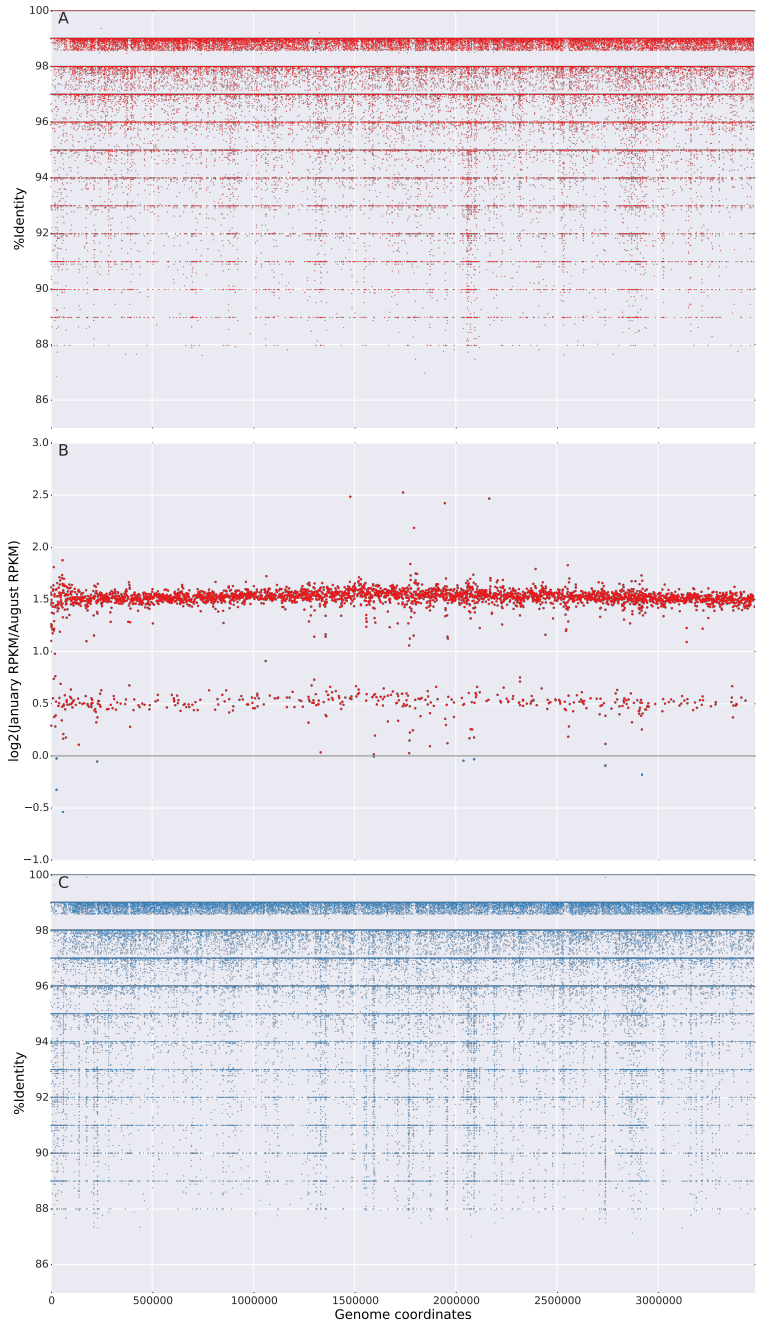


Figure B.1: Coverage and gene abundance for J07HQW1. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

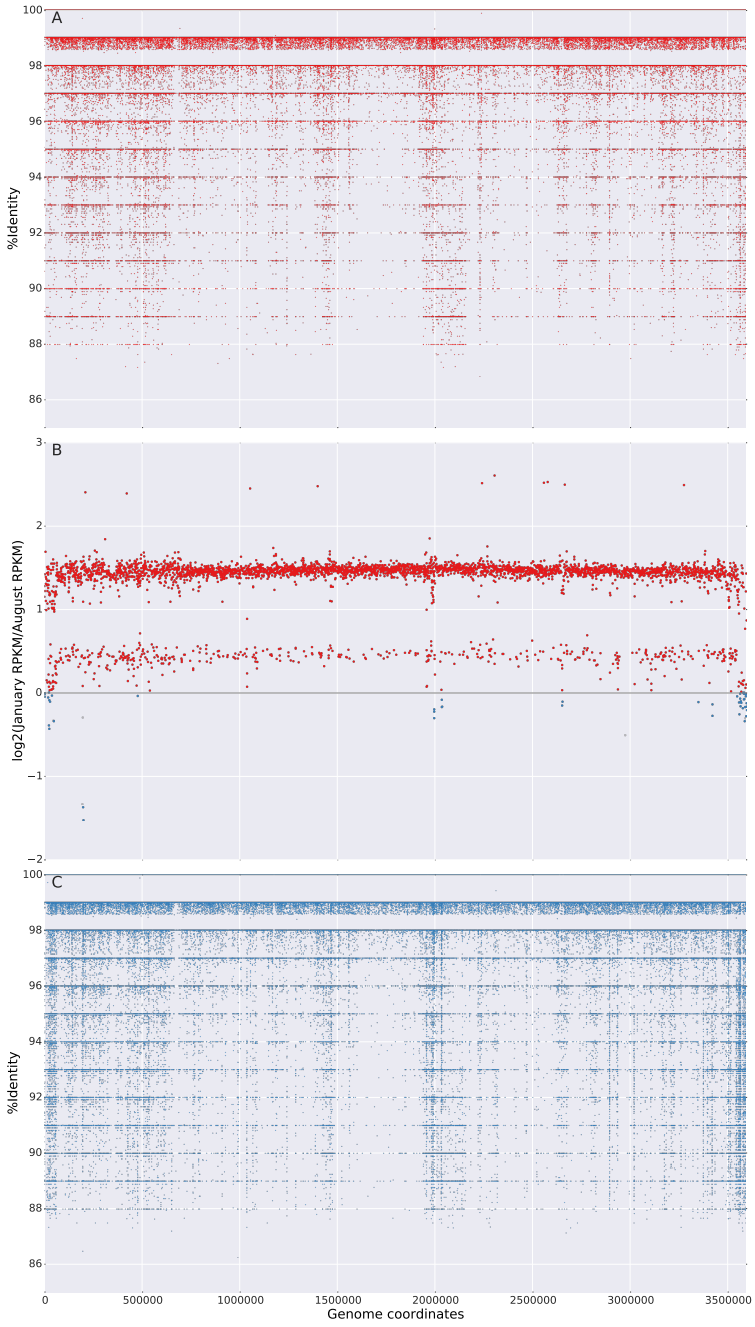


Figure B.2: Coverage and gene abundance for J07HQW2. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

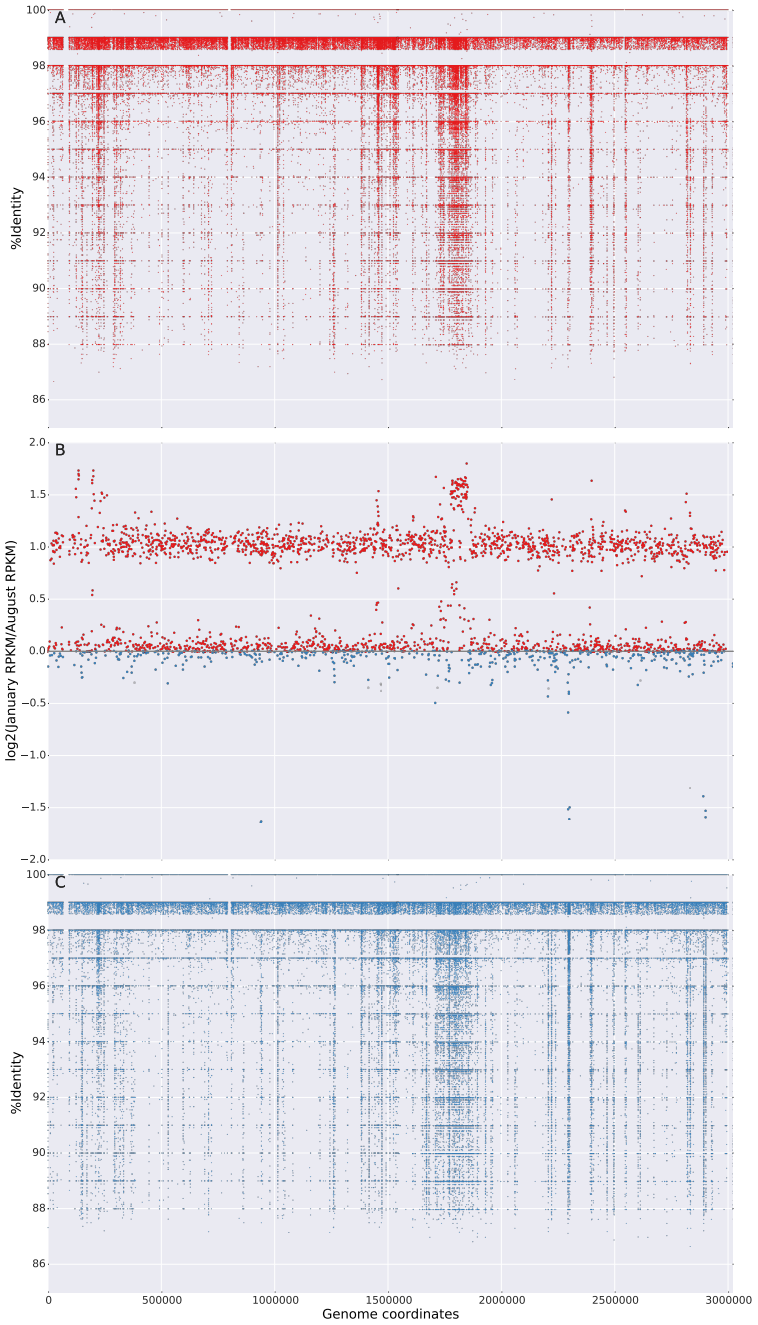


Figure B.3: Coverage and gene abundance for J07HQX50. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

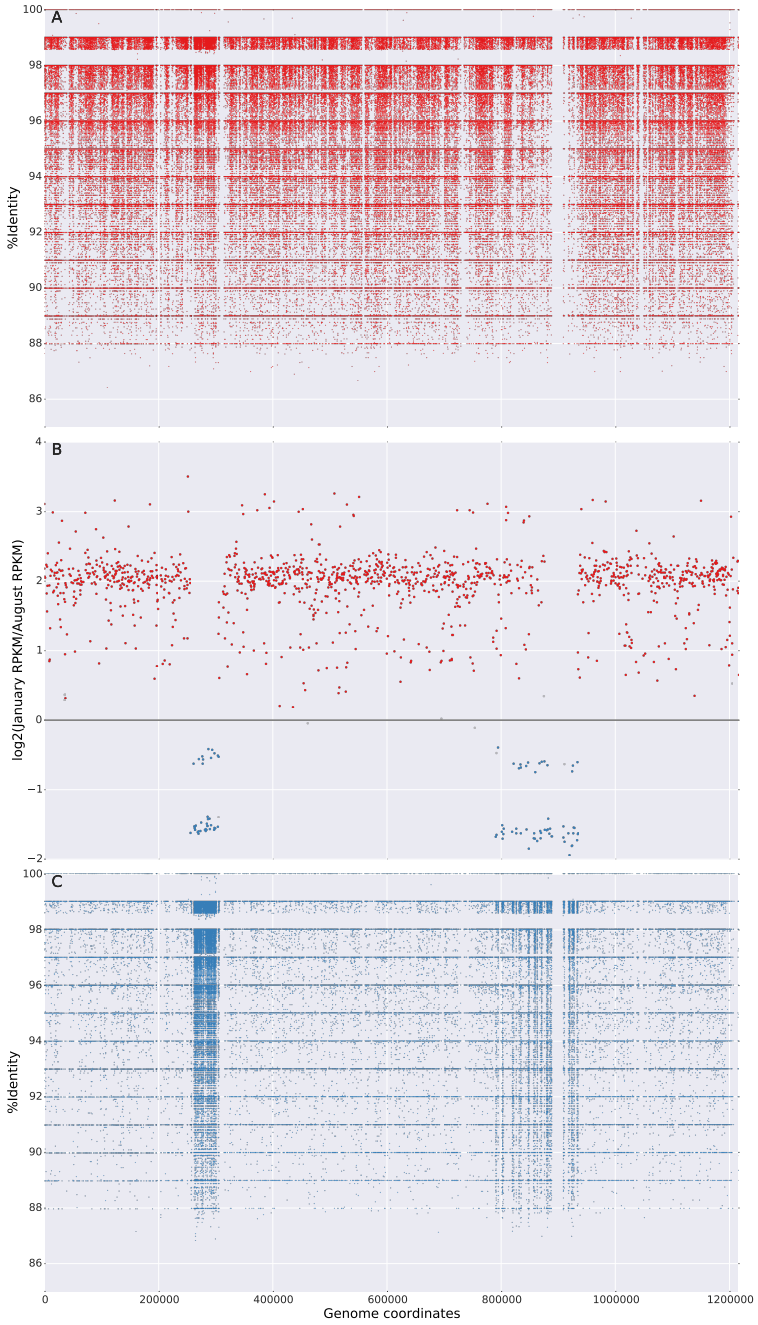


Figure B.4: Coverage and gene abundance for J07AB56. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

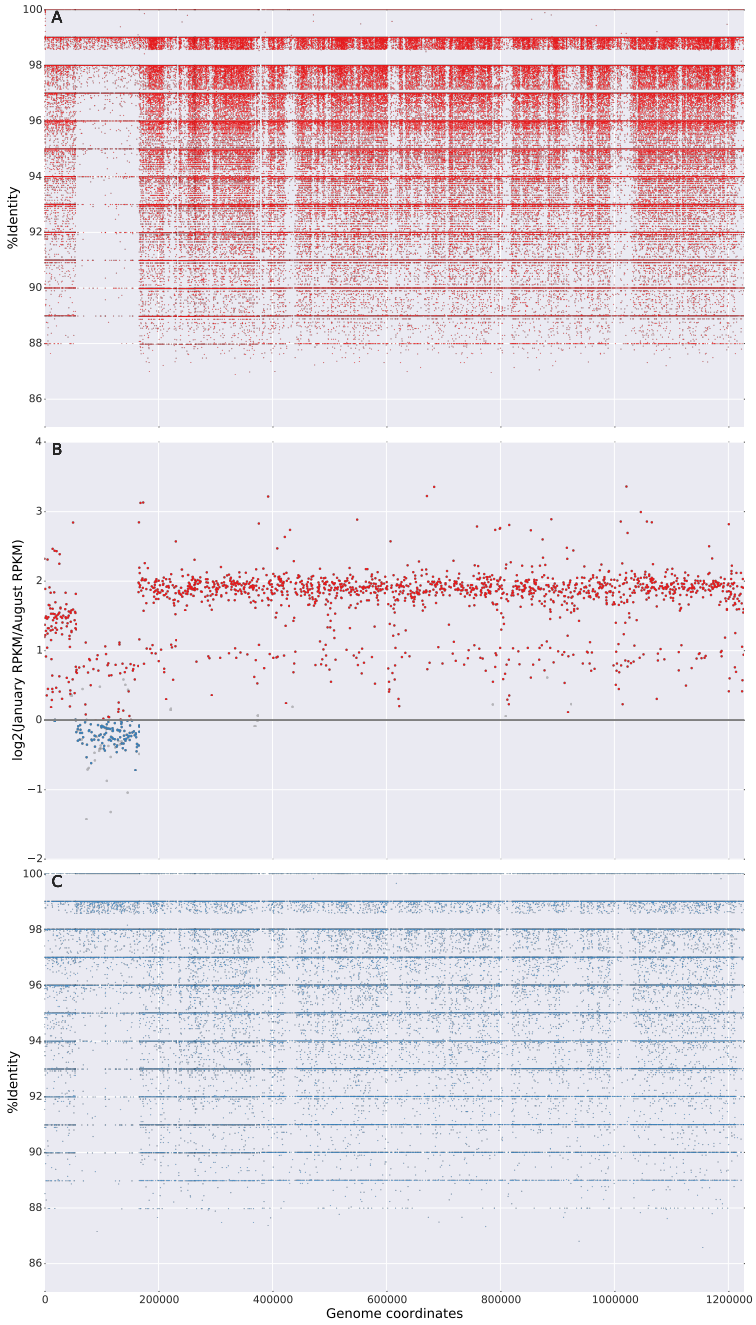


Figure B.5: Coverage and gene abundance for J07AB43. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

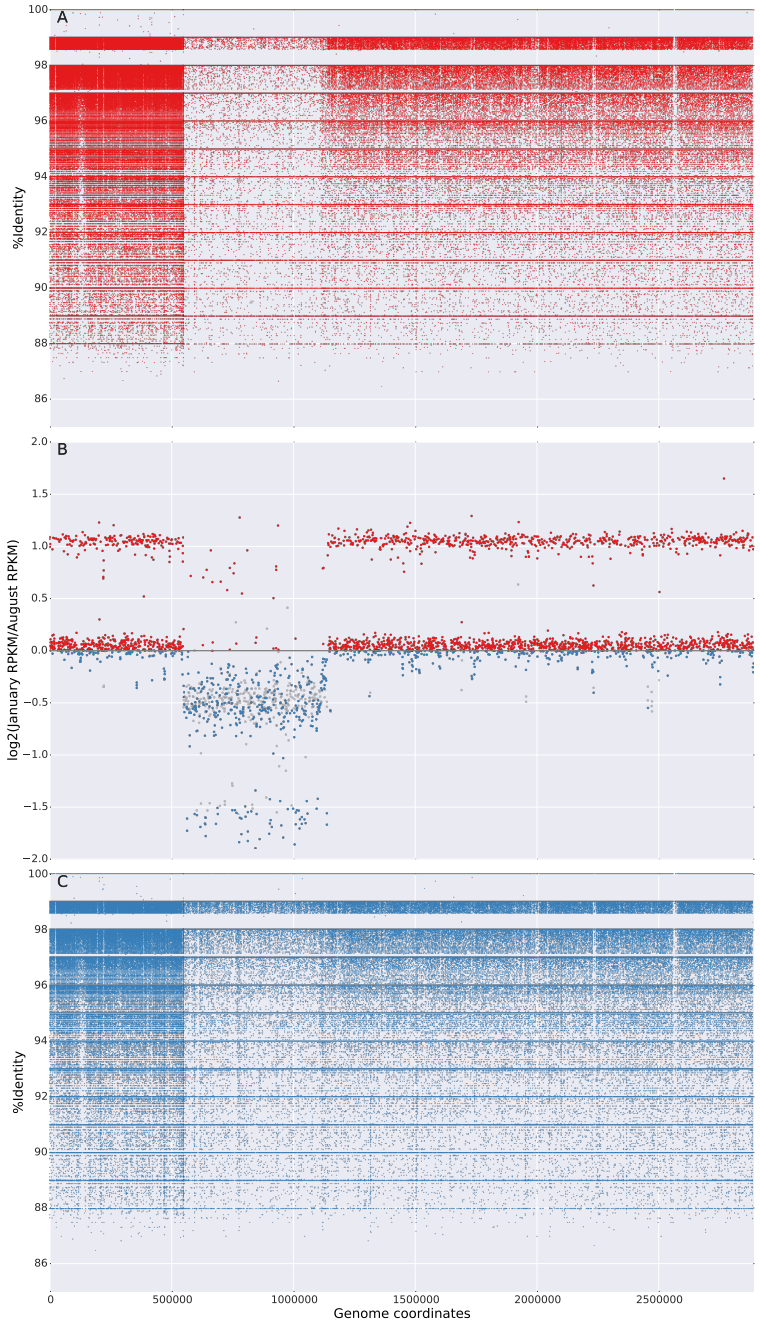


Figure B.6: Coverage and gene abundance for J07HN4. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

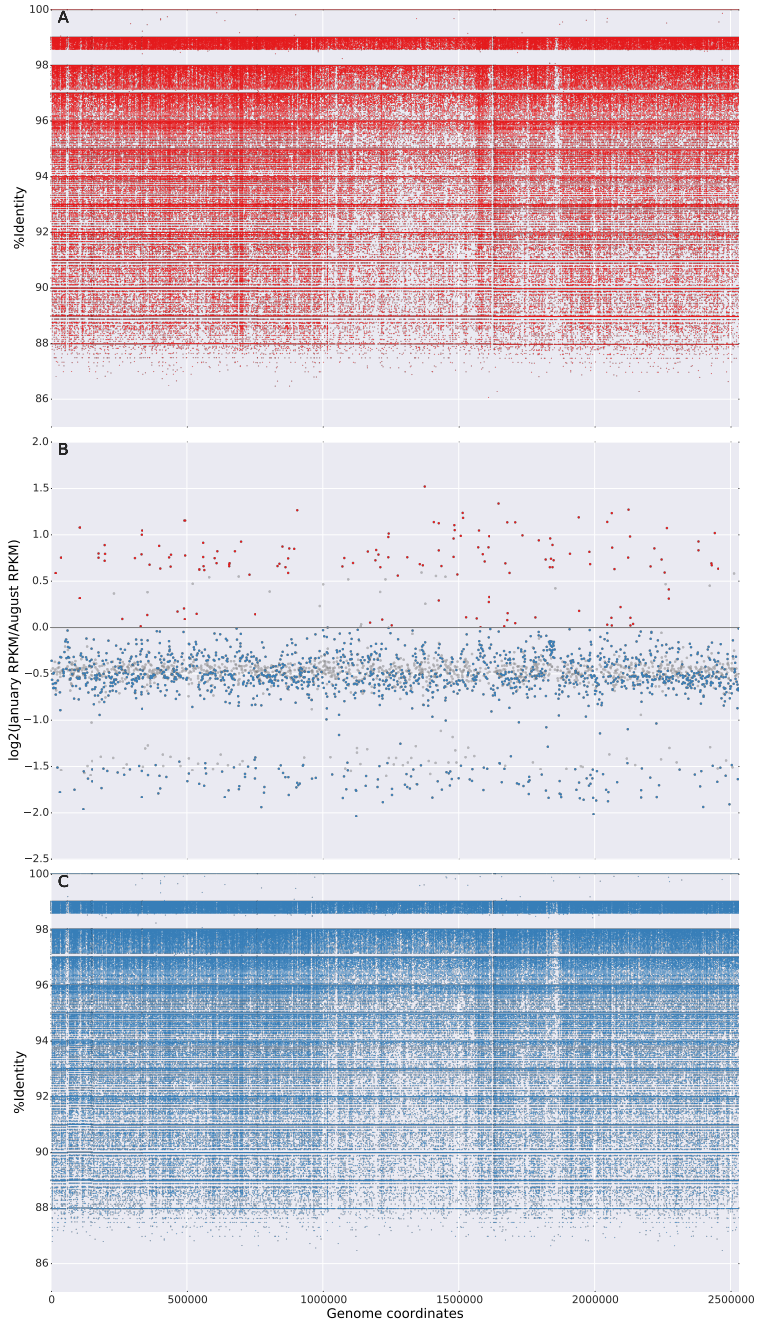


Figure B.7: Coverage and gene abundance for J07HN6. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

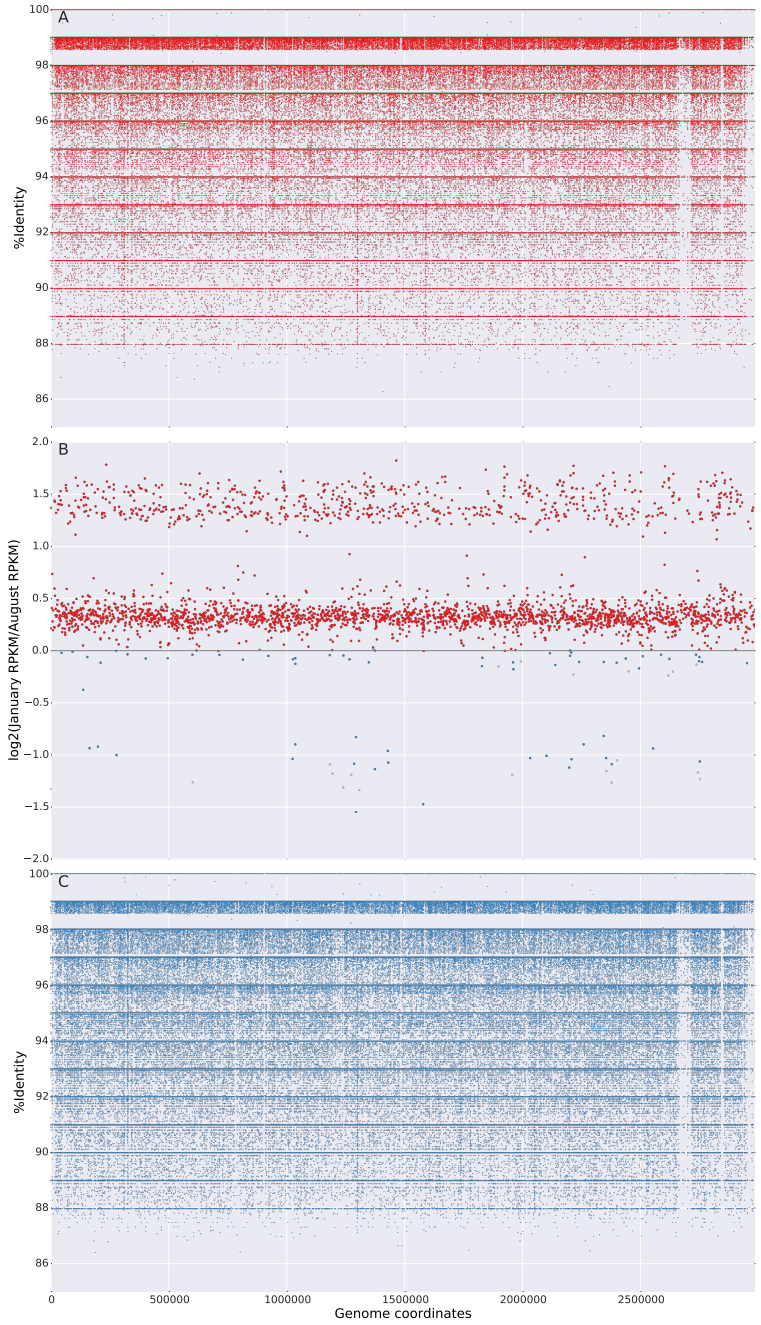


Figure B.8: Coverage and gene abundance for J07HN64. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

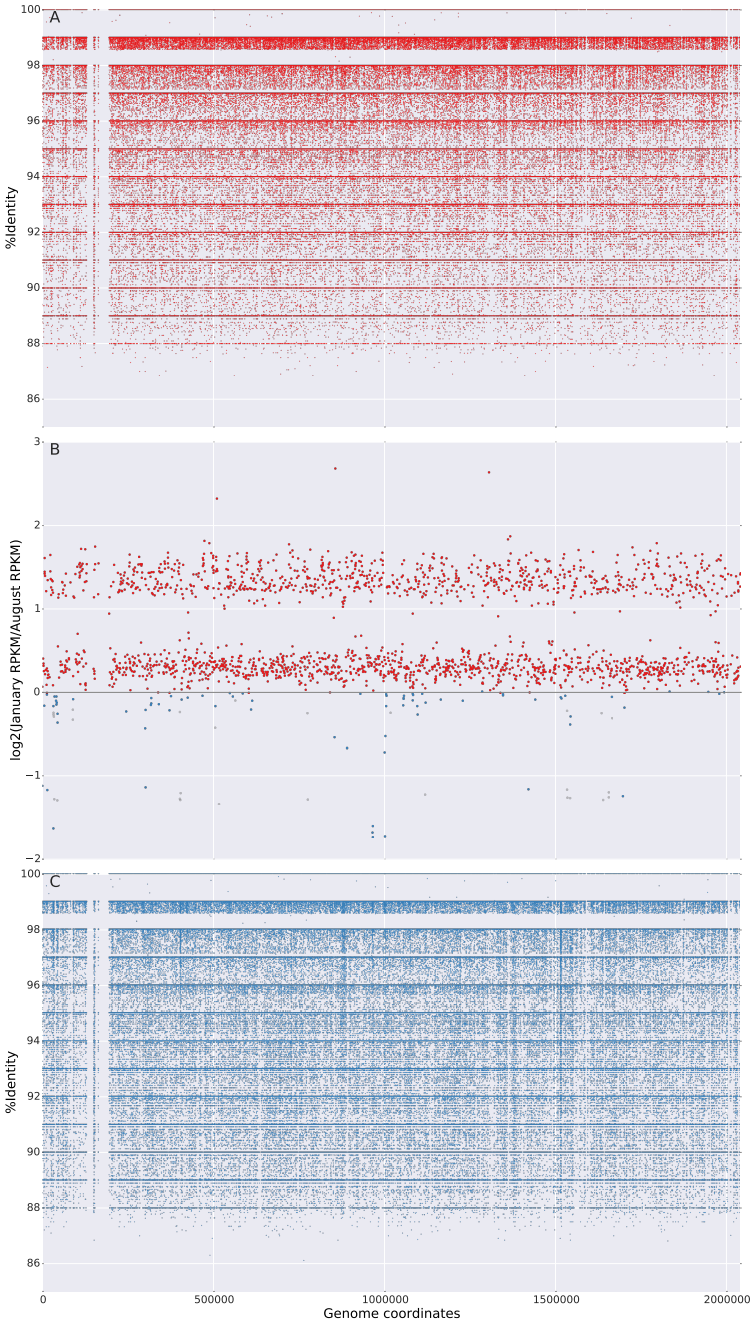


Figure B.9: Coverage and gene abundance for J07HX5. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

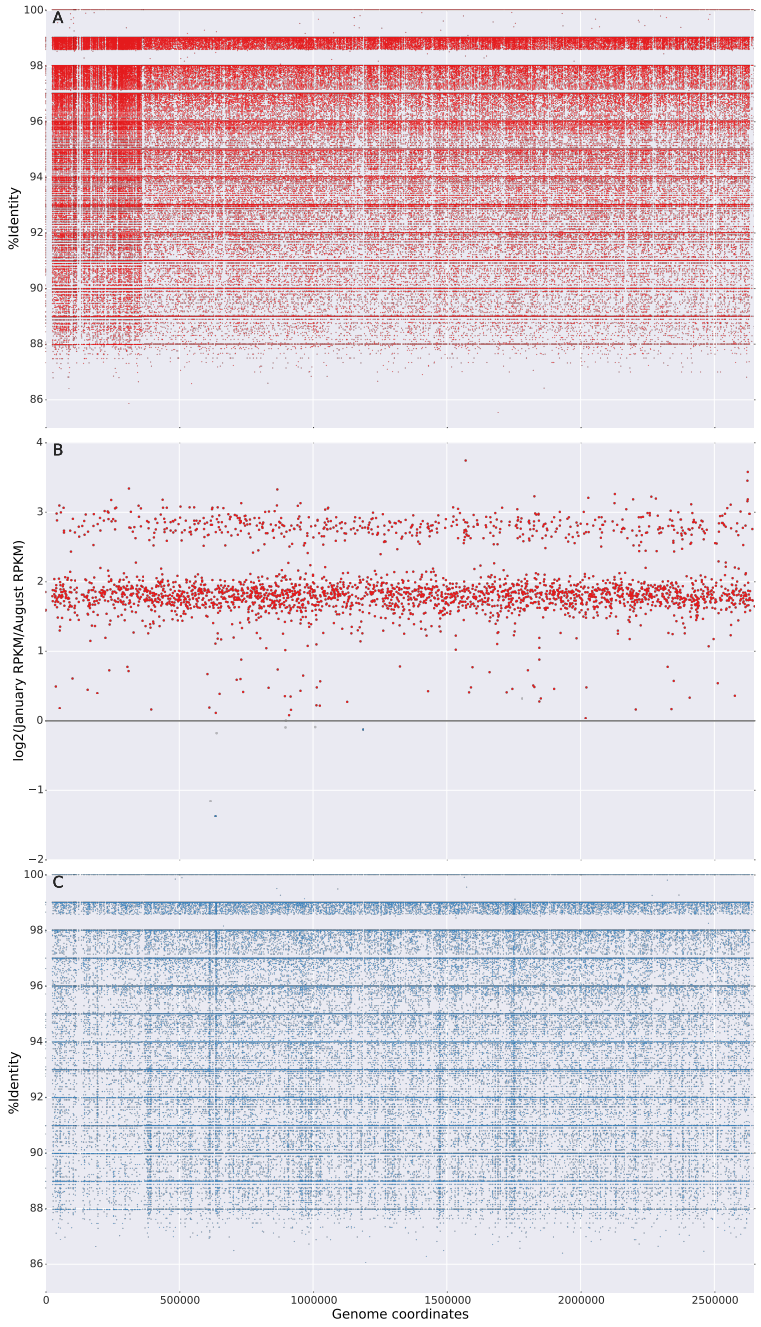


Figure B.10: Coverage and gene abundance for J07HB67. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

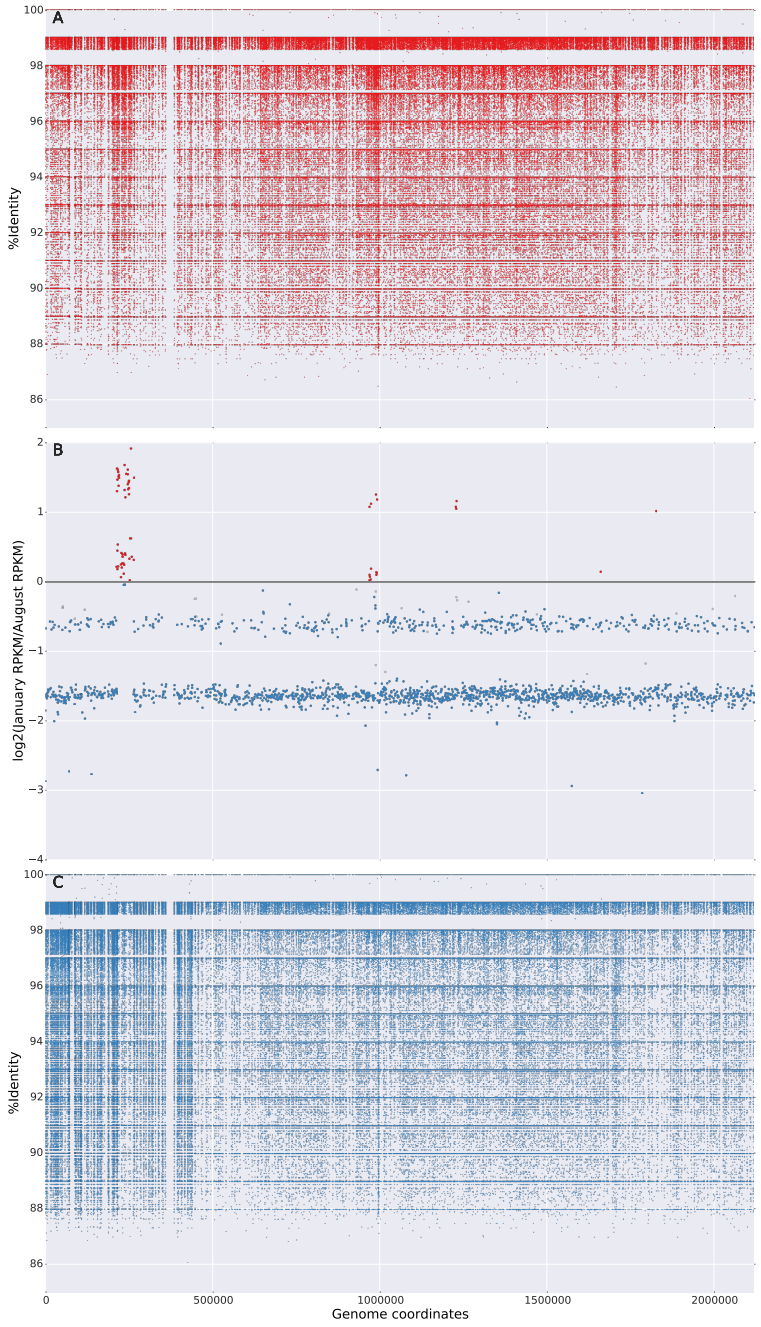


Figure B.11: Coverage and gene abundance for J07HR59. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

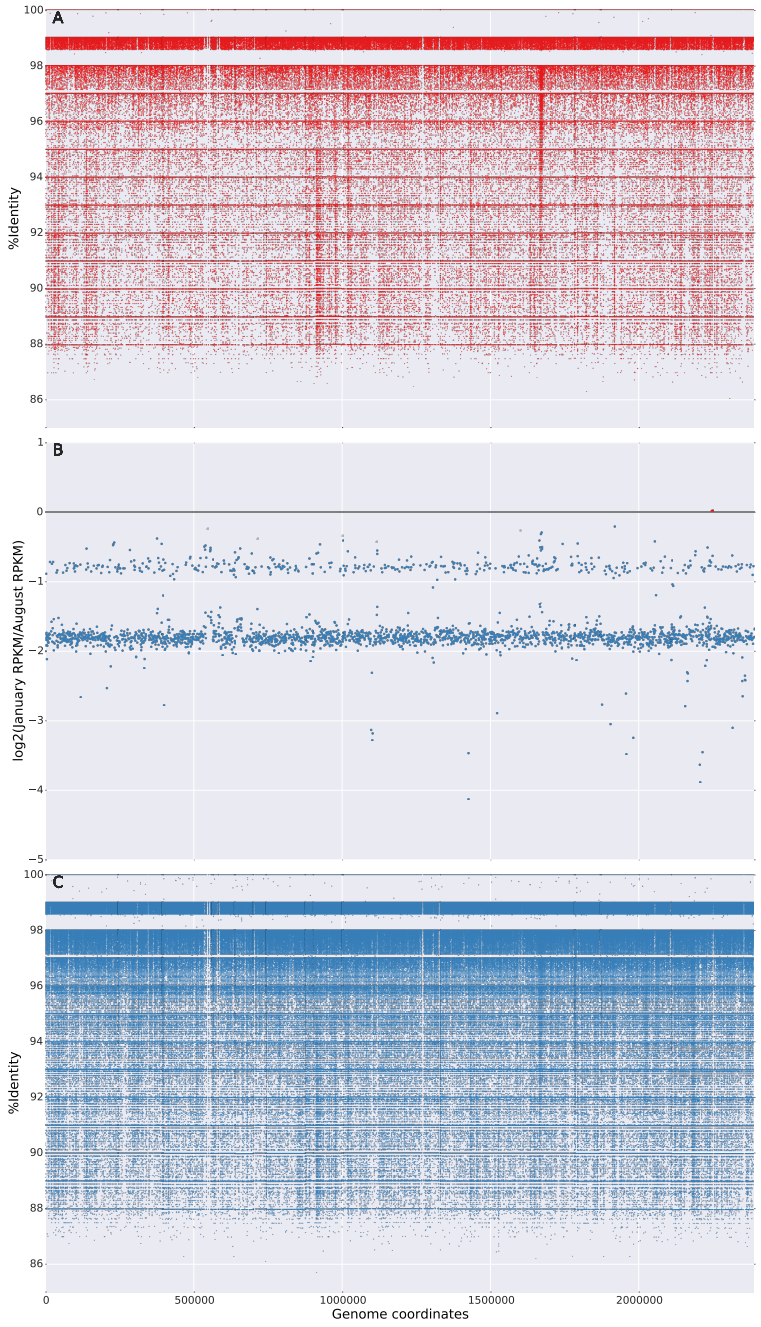


Figure B.12: Coverage and gene abundance for A07HB70. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

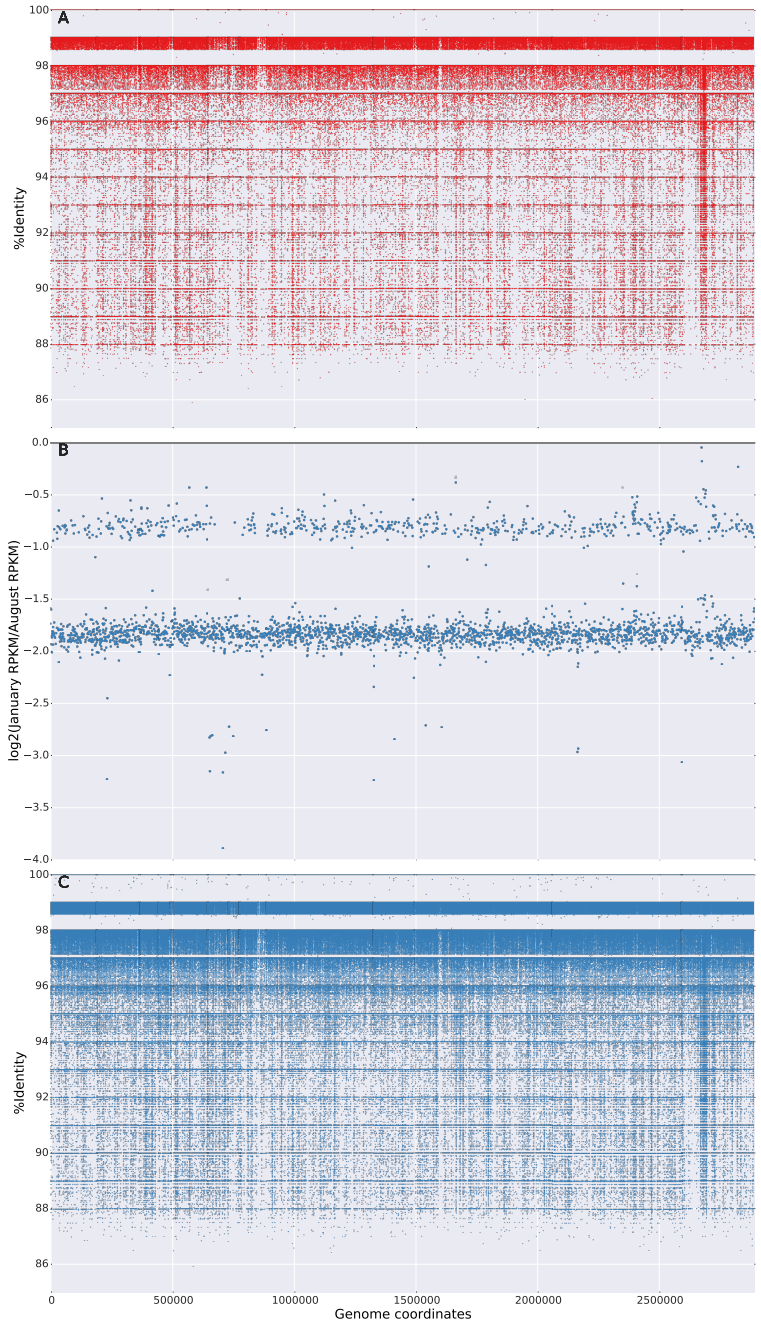


Figure B.13: Coverage and gene abundance for A07HR67. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

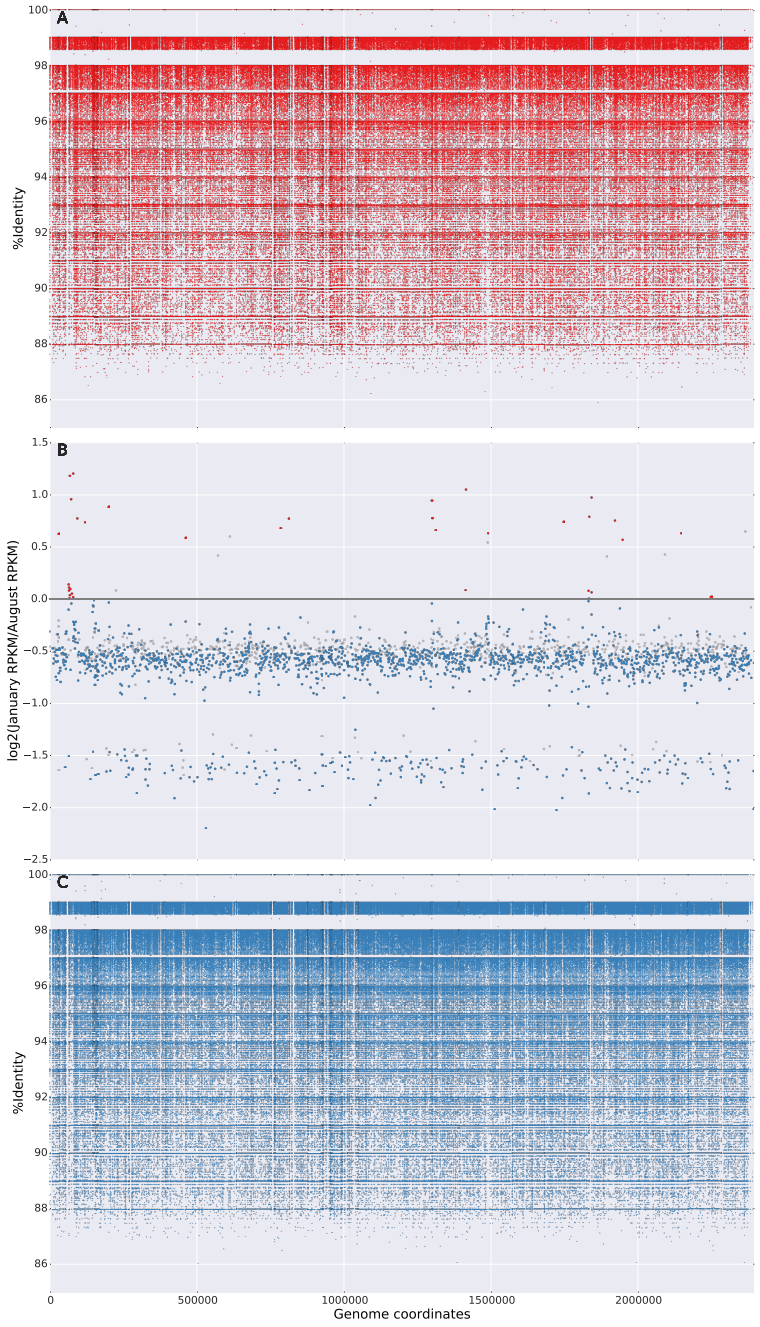


Figure B.14: Coverage and gene abundance for A07HN63. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

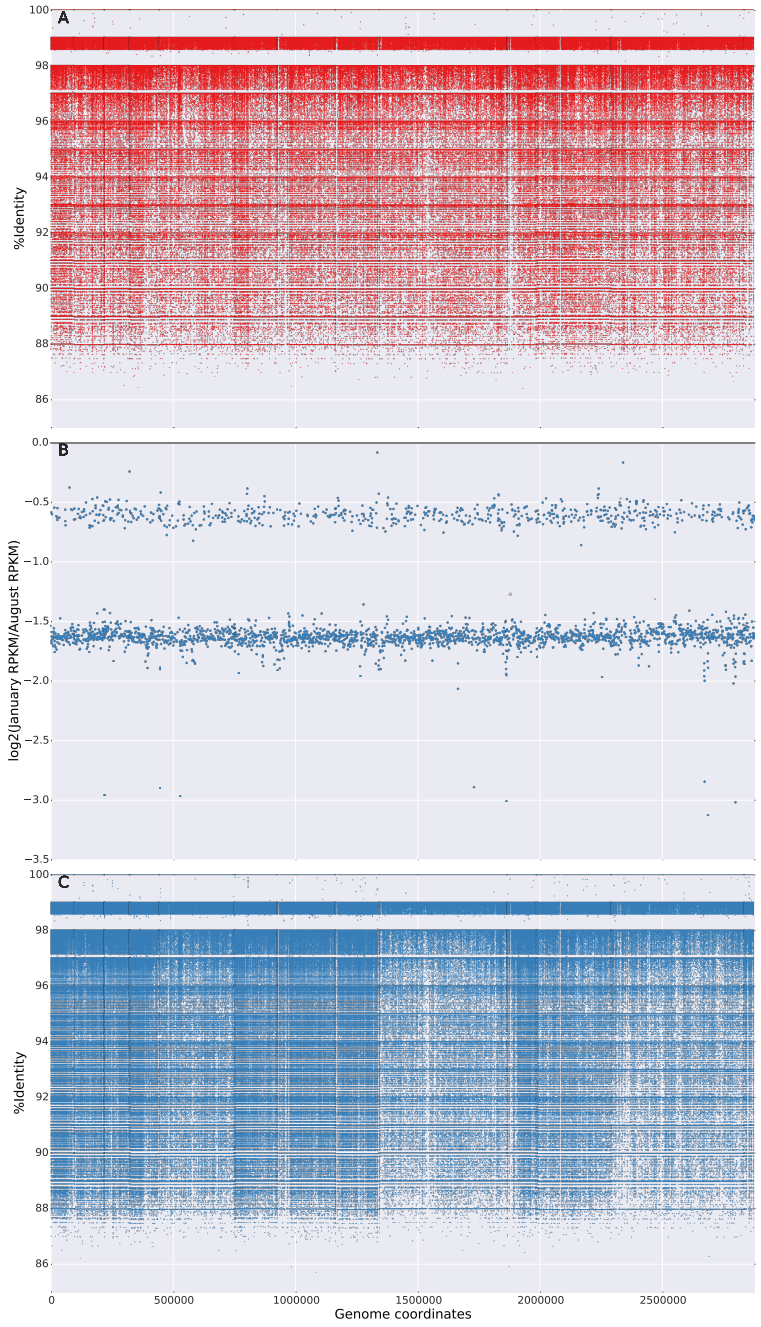


Figure B.15: Coverage and gene abundance for A07HR60. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

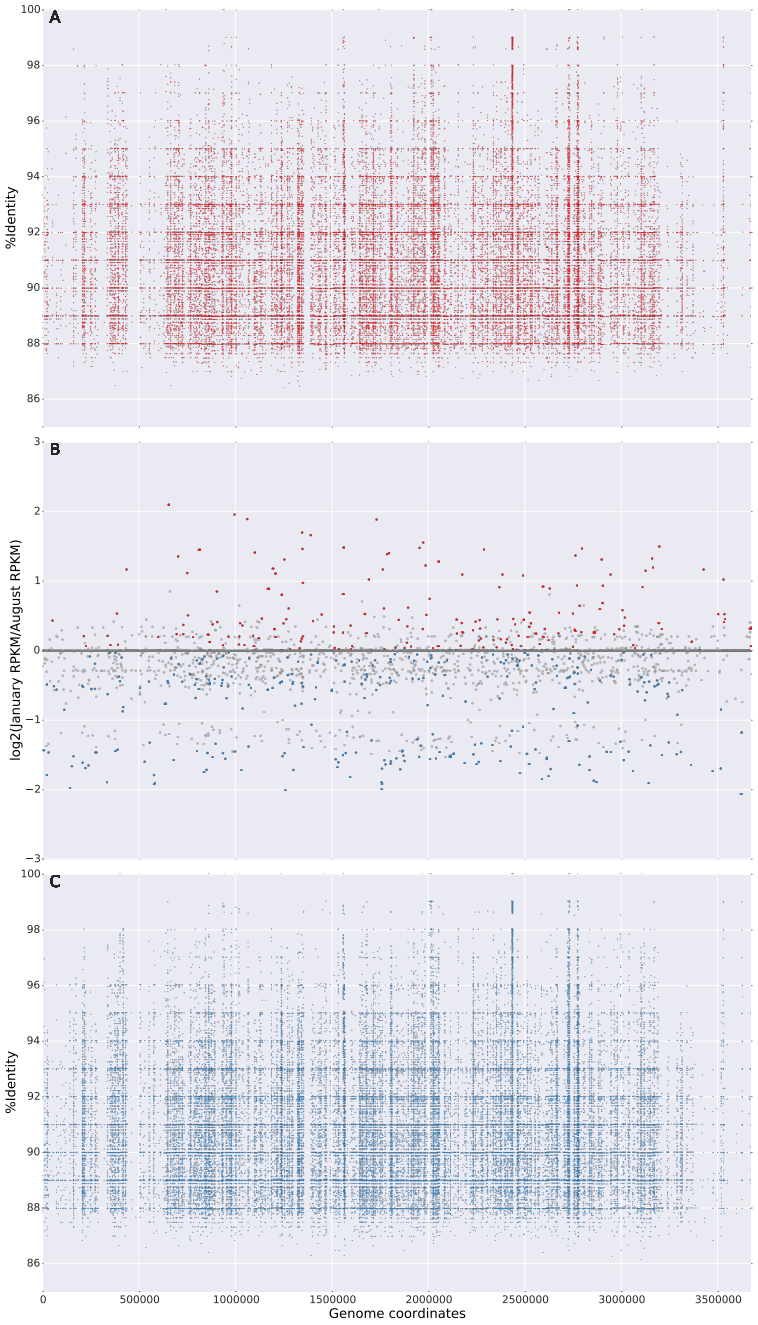


Figure B.16: Coverage and gene abundance for G22. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

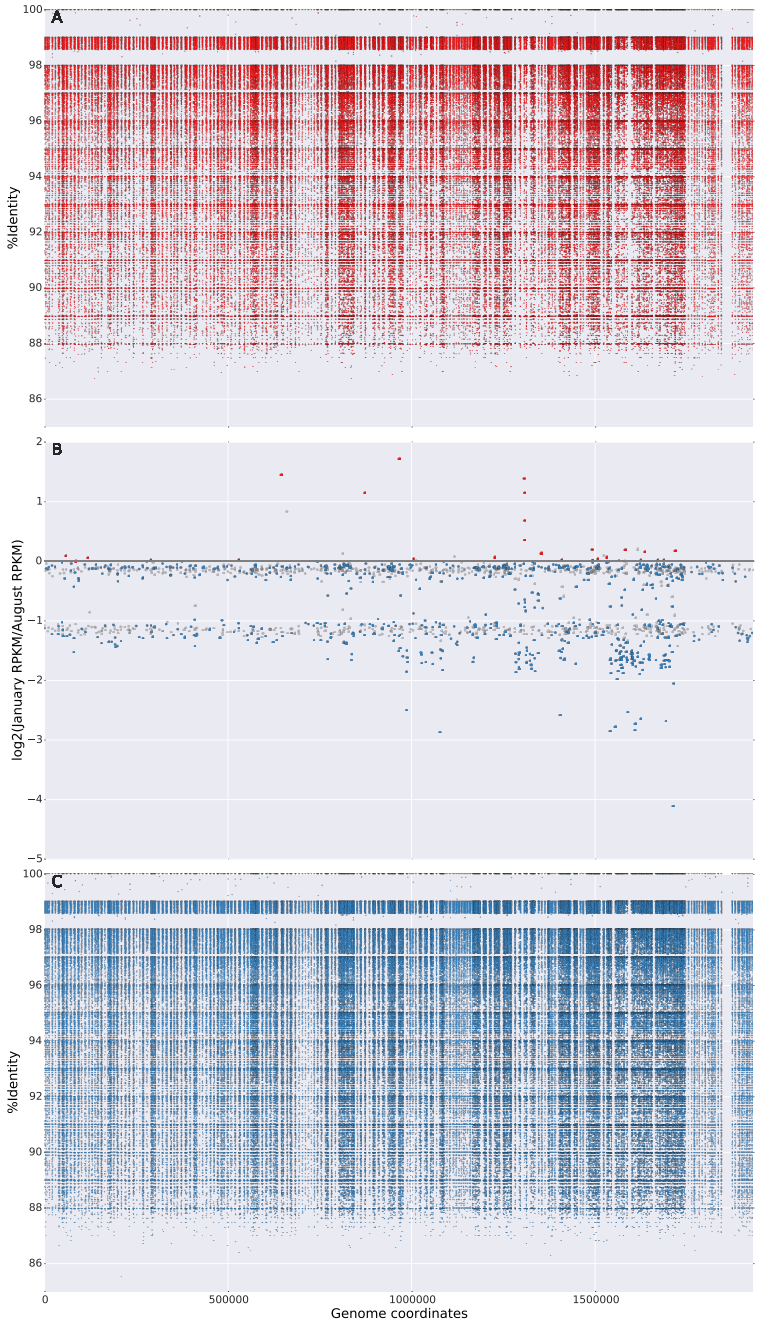


Figure B.17: Coverage and gene abundance for J07SB. **A** and **C** shows reads recruited to the January and August genomes, respectively. **B** indicates the number of reads recruited to each individual gene, expressed as RPKM values, where the color indicates the sample from where the read originated (January vs. August.)

Appendix C

COG categories with genes under positive selection

Table C.1: COG categories with genes under positive selection in the January sample for J07HWQ1. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	4	1.654	0.318
Amino acid transport and metabolism	2	0.426	0.311
Carbohydrate transport and metabolism	1	0.510	1.000
Coenzyme transport and metabolism	5	2.102	0.180
Transcription	3	1.031	1.000
Translation, ribosomal structure and biogenesis	1	0.320	0.362
Cell wall/membrane/envelope biogenesis	1	0.725	1.000
Replication, recombination and repair	4	1.215	0.574
Post-translational modification, protein turnover, and chaperones	2	0.969	1.000
Inorganic ion transport and metabolism	3	1.054	0.761
Function unknown	7	1.727	0.199
General function prediction only	5	0.739	0.669
Intracellular trafficking, secretion, and vesicular transport	2	5.088	0.070
Signal transduction mechanisms	3	2.031	0.203
Defense mechanisms	1	1.906	0.422

Table C.2: COG categories with genes under positive selection in the August sample for J07HWQ1. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	4	1.613	0.326
Amino acid transport and metabolism	2	0.416	0.312
Carbohydrate transport and metabolism	1	0.499	0.720
Coenzyme transport and metabolism	6	2.522	0.046
Transcription	3	1.006	1.000
Translation, ribosomal structure and biogenesis	1	0.313	0.363
Cell wall/membrane/envelope biogenesis	1	0.709	1.000
Replication, recombination and repair	4	1.186	0.773
Post-translational modification, protein turnover, and chaperones	2	0.947	1.000
Inorganic ion transport and metabolism	3	1.029	1.000
Function unknown	7	1.681	0.207
General function prediction only	5	0.720	0.670
Intracellular trafficking, secretion, and vesicular transport	2	4.970	0.073
Signal transduction mechanisms	3	1.983	0.212
Defense mechanisms	1	1.863	0.429

Table C.3: COG categories with genes under positive selection in the January sample for J07HWQ2. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	1	0.625	1.000
Amino acid transport and metabolism	4	1.094	0.782
Cell cycle control, cell division, chromosome partitioning	1	3.986	0.237
Coenzyme transport and metabolism	1	0.698	1.000
Transcription	1	0.463	0.719
Translation, ribosomal structure and biogenesis	2	0.917	1.000
Cell wall/membrane/envelope biogenesis	2	2.384	0.222
Replication, recombination and repair	4	1.368	0.539
Inorganic ion transport and metabolism	2	0.943	1.000
Function unknown	1	0.298	0.360
General function prediction only	10	2.652	**0.020**
Signal transduction mechanisms	1	1.215	0.571
Defense mechanisms	2	5.271	0.065

Table C.4: COG categories with genes under positive selection in the August sample for J07HWQ2. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	1	0.625	1.000
Amino acid transport and metabolism	5	1.418	0.410
Cell cycle control, cell division, chromosome partitioning	1	3.986	0.237
Coenzyme transport and metabolism	1	0.698	1.000
Transcription	1	0.463	0.719
Translation, ribosomal structure and biogenesis	2	0.917	1.000
Cell wall/membrane/envelope biogenesis	2	2.384	0.222
Replication, recombination and repair	4	1.368	0.539
Inorganic ion transport and metabolism	2	0.943	1.000
Function unknown	1	0.298	0.360
General function prediction only	9	2.283	**0.043**
Signal transduction mechanisms	1	1.215	0.571
Defense mechanisms	2	5.271	0.065

Table C.5: COG categories with genes under positive selection in the January sample for J07HQX50. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	Count	Odds ratio	pvalue
Transcription	1	1.529	0.501
Cell wall/membrane/envelope biogenesis	3	9.513	**0.007**
Replication, recombination and repair	2	2.106	0.279
Post-translational modification, protein turnover, and chaperones	2	3.759	0.121
Function unknown	1	0.825	1.000
General function prediction only	2	1.043	1.000
Intracellular trafficking, secretion, and vesicular transport	1	6.163	0.166
Signal transduction mechanisms	1	2.745	0.326

Table C.6: COG categories with genes under positive selection in the August sample for J07HQX50. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	Count	Odds ratio	pvalue
Transcription	1	1.529	0.501
Cell wall/membrane/envelope biogenesis	3	9.513	**0.007**
Replication, recombination and repair	2	2.106	0.279
Post-translational modification, protein turnover, and chaperones	2	3.759	0.121
Function unknown	1	0.825	1.000
General function prediction only	2	1.043	1.000
Intracellular trafficking, secretion, and vesicular transport	1	6.163	0.166
Signal transduction mechanisms	1	2.745	0.326

Table C.7: COG categories with genes under positive selection in the January sample for J07AB56. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	7	2.029	0.101
Chromatin structure and dynamics	1	2.118	0.421
Amino acid transport and metabolism	3	0.824	1.000
Cell cycle control, cell division, chromosome partitioning	3	1.687	0.429
Carbohydrate transport and metabolism	1	0.260	0.246
Nucleotide transport and metabolism	3	1.019	1.000
Coenzyme transport and metabolism	4	2.163	0.148
Transcription	5	0.935	1.000
Translation, ribosomal structure and biogenesis	8	0.484	0.065
Cell wall/membrane/envelope biogenesis	1	0.301	0.355
Replication, recombination and repair	5	0.723	0.670
Post-translational modification, protein turnover, and chaperones	10	2.127	0.058
Cell motility	4	6.278	**0.011**
Inorganic ion transport and metabolism	3	0.925	1.000
Function unknown	6	1.002	1.000
General function prediction only	6	0.533	0.162
Intracellular trafficking, secretion, and vesicular transport	6	3.712	0.013
Signal transduction mechanisms	1	1.509	0.518
Defense mechanisms	2	1.110	0.703

Table C.8: COG categories with genes under positive selection in the August sample for J07AB56. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	4	1.518	0.353
Chromatin structure and dynamics	1	2.848	0.338
Amino acid transport and metabolism	3	1.118	0.749
Cell cycle control, cell division, chromosome partitioning	2	1.500	0.644
Carbohydrate transport and metabolism	2	0.713	1.000
Nucleotide transport and metabolism	1	0.445	0.717
Coenzyme transport and metabolism	2	1.423	0.652
Transcription	5	1.281	0.591
Translation, ribosomal structure and biogenesis	3	0.230	0.005
Cell wall/membrane/envelope biogenesis	1	0.404	0.724
Replication, recombination and repair	6	1.212	0.632
Post-translational modification, protein turnover, and chaperones	4	1.068	0.786
Cell motility	2	4.130	0.115
Inorganic ion transport and metabolism	2	0.822	1.000
Function unknown	5	1.129	0.798
General function prediction only	9	1.168	0.693
Intracellular trafficking, secretion, and vesicular transport	5	4.182	**0.014**
Defense mechanisms	2	1.500	0.644

Table C.9: COG categories with genes under positive selection in the January sample for J07AB43. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	Count	Odds ratio	pvalue
Energy production and conversion	5	1.776	0.223
Amino acid transport and metabolism	2	0.526	0.572
Cell cycle control, cell division, chromosome partitioning	2	1.933	0.310
Carbohydrate transport and metabolism	2	0.419	0.307
Nucleotide transport and metabolism	3	0.951	1.000
Lipid transport and metabolism	1	1.557	0.504
Coenzyme transport and metabolism	3	1.169	0.741
Transcription	6	1.470	0.434
Translation, ribosomal structure and biogenesis	9	0.712	0.411
Cell wall/membrane/envelope biogenesis	4	1.222	0.768
Replication, recombination and repair	6	0.898	1.000
Post-translational modification, protein turnover, and chaperones	7	1.707	0.202
Secondary metabolites biosynthesis, transport, and catabolism	1	2.500	0.373
Function unknown	7	1.094	0.830
General function prediction only	11	1.310	0.454
Intracellular trafficking, secretion, and vesicular transport	2	1.471	0.647

Table C.10: COG categories with genes under positive selection in the August sample for J07AB43. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	4	1.537	0.348
Amino acid transport and metabolism	1	0.283	0.246
Cell cycle control, cell division, chromosome partitioning	1	1.042	1.000
Carbohydrate transport and metabolism	2	0.458	0.423
Nucleotide transport and metabolism	3	1.043	0.762
Coenzyme transport and metabolism	2	0.841	1.000
Transcription	6	1.619	0.280
Translation, ribosomal structure and biogenesis	11	0.999	1.000
Cell wall/membrane/envelope biogenesis	4	1.342	0.543
Replication, recombination and repair	6	0.990	1.000
Post-translational modification, protein turnover, and chaperones	7	1.883	0.179
Secondary metabolites biosynthesis, transport, and catabolism	1	2.734	0.349
Function unknown	6	1.017	1.000
General function prediction only	9	1.149	0.696
Intracellular trafficking, secretion, and vesicular transport	1	0.793	1.000
Signal transduction mechanisms	1	1.130	0.606

Table C.11: COG categories with genes under positive selection in the January sample for J07HN4. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	Count	Odds ratio	pvalue
Energy production and conversion	6	0.463	0.063
Amino acid transport and metabolism	19	1.202	0.500
Cell cycle control, cell division, chromosome partitioning	4	1.853	0.286
Carbohydrate transport and metabolism	8	1.233	0.538
Nucleotide transport and metabolism	6	0.985	1.000
Lipid transport and metabolism	3	0.865	1.000
Coenzyme transport and metabolism	13	1.156	0.634
Transcription	11	0.915	0.877
Translation, ribosomal structure and biogenesis	18	1.342	0.250
Cell wall/membrane/envelope biogenesis	2	0.645	0.766
Replication, recombination and repair	12	1.341	0.380
Post-translational modification, protein turnover, and chaperones	9	0.861	0.867
Cell motility	8	2.483	0.026
Secondary metabolites biosynthesis, transport, and catabolism	2	0.985	1.000
Inorganic ion transport and metabolism	10	0.874	0.873
Function unknown	16	0.860	0.700
General function prediction only	22	0.874	0.653
Intracellular trafficking, secretion, and vesicular transport	1	0.440	0.721
Signal transduction mechanisms	10	0.790	0.645
Defense mechanisms	3	1.747	0.422

Table C.12: COG categories with genes under positive selection in the August sample for J07HN4. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	Count	Odds ratio	pvalue
Energy production and conversion	6	0.466	0.063
Amino acid transport and metabolism	18	1.139	0.589
Cell cycle control, cell division, chromosome partitioning	4	1.864	0.285
Carbohydrate transport and metabolism	8	1.240	0.536
Nucleotide transport and metabolism	6	0.990	1.000
Lipid transport and metabolism	3	0.870	1.000
Coenzyme transport and metabolism	13	1.163	0.633
Transcription	12	1.010	1.000
Translation, ribosomal structure and biogenesis	16	1.186	0.561
Cell wall/membrane/envelope biogenesis	2	0.649	0.766
Replication, recombination and repair	13	1.470	0.219
Post-translational modification, protein turnover, and chaperones	9	0.866	0.867
Cell motility	8	2.497	0.025
Secondary metabolites biosynthesis, transport, and catabolism	2	0.991	1.000
Inorganic ion transport and metabolism	10	0.879	0.873
Function unknown	16	0.865	0.700
General function prediction only	22	0.880	0.653
Intracellular trafficking, secretion, and vesicular transport	1	0.443	0.720
Signal transduction mechanisms	10	0.794	0.644
Defense mechanisms	3	1.757	0.421

Table C.13: COG categories with genes under positive selection in the January sample for J07HN6. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Amino acid transport and metabolism	3	0.731	0.794
Carbohydrate transport and metabolism	1	0.695	1.000
Nucleotide transport and metabolism	1	0.574	1.000
Lipid transport and metabolism	4	4.077	0.023
Coenzyme transport and metabolism	2	0.718	1.000
Transcription	5	1.600	0.376
Translation, ribosomal structure and biogenesis	1	0.238	0.176
Cell wall/membrane/envelope biogenesis	3	3.354	0.073
Replication, recombination and repair	2	0.832	1.000
Post-translational modification, protein turnover, and chaperones	3	1.264	0.732
Cell motility	1	1.045	1.000
Inorganic ion transport and metabolism	3	1.056	0.761
Function unknown	4	0.754	0.813
General function prediction only	8	1.266	0.525
Intracellular trafficking, secretion, and vesicular transport	3	5.064	**0.028**
Signal transduction mechanisms	4	1.251	0.565

Table C.14: COG categories with genes under positive selection in the August sample for J07HN6. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	2	0.565	0.578
Amino acid transport and metabolism	2	0.414	0.317
Carbohydrate transport and metabolism	2	1.233	0.679
Nucleotide transport and metabolism	1	0.500	1.000
Lipid transport and metabolism	4	3.518	0.036
Coenzyme transport and metabolism	2	0.623	0.767
Transcription	6	1.685	0.271
Translation, ribosomal structure and biogenesis	1	0.207	0.125
Cell wall/membrane/envelope biogenesis	3	2.902	0.099
Replication, recombination and repair	2	0.722	1.000
Post-translational modification, protein turnover, and chaperones	5	1.895	0.201
Cell motility	1	0.910	1.000
Inorganic ion transport and metabolism	4	1.242	0.567
Function unknown	5	0.829	0.828
General function prediction only	7	0.923	1.000
Intracellular trafficking, secretion, and vesicular transport	4	5.958	**0.007**
Signal transduction mechanisms	4	1.079	0.786

Table C.15: COG categories with genes under positive selection in the January sample for J07HX64. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	Count	Odds ratio	pvalue
Energy production and conversion	6	0.812	0.843
Amino acid transport and metabolism	23	1.470	0.108
Cell cycle control, cell division, chromosome partitioning	2	1.629	0.373
Carbohydrate transport and metabolism	9	1.487	0.289
Nucleotide transport and metabolism	6	1.186	0.641
Lipid transport and metabolism	2	0.347	0.168
Coenzyme transport and metabolism	8	1.062	0.846
Transcription	7	0.975	1.000
Translation, ribosomal structure and biogenesis	9	0.743	0.526
Cell wall/membrane/envelope biogenesis	4	0.924	1.000
Replication, recombination and repair	10	1.203	0.582
Post-translational modification, protein turnover, and chaperones	5	0.637	0.437
Cell motility	2	1.447	0.650
Secondary metabolites biosynthesis, transport, and catabolism	5	1.423	0.409
Inorganic ion transport and metabolism	6	0.935	1.000
Function unknown	15	0.933	0.891
General function prediction only	27	1.000	1.000
Intracellular trafficking, secretion, and vesicular transport	2	0.863	1.000
Signal transduction mechanisms	6	0.859	1.000
Defense mechanisms	1	0.444	0.720

Table C.16: COG categories with genes under positive selection in the August sample for J07HX64. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	Count	Odds ratio	pvalue
Energy production and conversion	6	0.817	0.842
Amino acid transport and metabolism	24	1.558	0.061
Cell cycle control, cell division, chromosome partitioning	2	1.640	0.370
Carbohydrate transport and metabolism	9	1.497	0.287
Nucleotide transport and metabolism	6	1.194	0.640
Lipid transport and metabolism	2	0.349	0.168
Coenzyme transport and metabolism	8	1.069	0.846
Transcription	7	0.982	1.000
Translation, ribosomal structure and biogenesis	9	0.748	0.526
Cell wall/membrane/envelope biogenesis	4	0.930	1.000
Replication, recombination and repair	10	1.211	0.580
Post-translational modification, protein turnover, and chaperones	5	0.641	0.437
Cell motility	2	1.456	0.649
Secondary metabolites biosynthesis, transport, and catabolism	5	1.433	0.406
Inorganic ion transport and metabolism	6	0.941	1.000
Function unknown	14	0.871	0.782
General function prediction only	27	1.008	1.000
Intracellular trafficking, secretion, and vesicular transport	2	0.868	1.000
Signal transduction mechanisms	5	0.716	0.681
Defense mechanisms	1	0.446	0.719

Table C.17: COG categories with genes under positive selection in the January sample for J07HX5. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	7	1.164	0.666
Amino acid transport and metabolism	11	0.818	0.647
Cell cycle control, cell division, chromosome partitioning	1	0.623	1.000
Carbohydrate transport and metabolism	6	1.386	0.450
Nucleotide transport and metabolism	4	0.775	0.814
Lipid transport and metabolism	3	0.730	0.793
Coenzyme transport and metabolism	7	1.197	0.661
Transcription	6	1.165	0.644
Translation, ribosomal structure and biogenesis	10	1.349	0.437
Cell wall/membrane/envelope biogenesis	4	1.367	0.539
Replication, recombination and repair	7	0.713	0.482
Post-translational modification, protein turnover, and chaperones	2	0.357	0.164
Cell motility	1	1.395	0.537
Secondary metabolites biosynthesis, transport, and catabolism	1	0.566	1.000
Inorganic ion transport and metabolism	12	2.120	**0.031**
Function unknown	9	0.771	0.624
General function prediction only	19	1.134	0.590
Intracellular trafficking, secretion, and vesicular transport	2	1.399	0.655
Signal transduction mechanisms	1	0.217	0.120
Defense mechanisms	3	1.651	0.434

Table C.18: COG categories with genes under positive selection in the August sample for J07HX5. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	6	0.997	1.000
Amino acid transport and metabolism	11	0.826	0.647
Cell cycle control, cell division, chromosome partitioning	1	0.629	1.000
Carbohydrate transport and metabolism	6	1.399	0.447
Nucleotide transport and metabolism	4	0.782	0.814
Lipid transport and metabolism	3	0.736	0.793
Coenzyme transport and metabolism	7	1.208	0.659
Transcription	6	1.175	0.642
Translation, ribosomal structure and biogenesis	10	1.361	0.336
Cell wall/membrane/envelope biogenesis	4	1.379	0.537
Replication, recombination and repair	7	0.720	0.482
Post-translational modification, protein turnover, and chaperones	2	0.360	0.234
Cell motility	1	1.407	0.534
Secondary metabolites biosynthesis, transport, and catabolism	1	0.571	1.000
Inorganic ion transport and metabolism	12	2.141	**0.031**
Function unknown	9	0.778	0.623
General function prediction only	19	1.146	0.587
Intracellular trafficking, secretion, and vesicular transport	2	1.411	0.654
Signal transduction mechanisms	1	0.219	0.120
Defense mechanisms	3	1.665	0.432

Table C.19: COG categories with genes under positive selection in the January sample for J07HB67. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	7	0.785	0.716
Amino acid transport and metabolism	18	0.930	0.899
Cell cycle control, cell division, chromosome partitioning	2	1.057	0.715
Carbohydrate transport and metabolism	7	1.262	0.507
Nucleotide transport and metabolism	7	1.262	0.507
Lipid transport and metabolism	4	0.741	0.816
Coenzyme transport and metabolism	6	0.831	0.842
Transcription	10	1.216	0.579
Translation, ribosomal structure and biogenesis	17	1.485	0.162
Cell wall/membrane/envelope biogenesis	4	1.189	0.773
Replication, recombination and repair	4	0.463	0.184
Post-translational modification, protein turnover, and chaperones	7	1.025	0.841
Cell motility	2	0.733	1.000
Secondary metabolites biosynthesis, transport, and catabolism	4	1.220	0.573
Inorganic ion transport and metabolism	5	0.642	0.436
Function unknown	20	1.326	0.270
General function prediction only	21	0.827	0.493
Intracellular trafficking, secretion, and vesicular transport	4	1.820	0.291
Signal transduction mechanisms	9	1.147	0.704
Defense mechanisms	2	0.690	1.000

Table C.20: COG categories with genes under positive selection in the August sample for J07HB67. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	Count	Odds ratio	pvalue
Energy production and conversion	7	0.890	1.000
Chromatin structure and dynamics	1	2.290	0.390
Amino acid transport and metabolism	21	1.274	0.351
Cell cycle control, cell division, chromosome partitioning	3	1.802	0.417
Carbohydrate transport and metabolism	4	0.799	0.815
Nucleotide transport and metabolism	6	1.216	0.636
Lipid transport and metabolism	4	0.837	1.000
Coenzyme transport and metabolism	5	0.779	0.831
Transcription	8	1.089	0.844
Translation, ribosomal structure and biogenesis	15	1.476	0.188
Cell wall/membrane/envelope biogenesis	1	0.329	0.361
Replication, recombination and repair	5	0.659	0.552
Post-translational modification, protein turnover, and chaperones	7	1.162	0.668
Cell motility	1	0.410	0.726
Secondary metabolites biosynthesis, transport, and catabolism	3	1.027	1.000
Inorganic ion transport and metabolism	6	0.878	1.000
Function unknown	13	0.935	1.000
General function prediction only	17	0.745	0.332
Intracellular trafficking, secretion, and vesicular transport	4	2.057	0.156
Signal transduction mechanisms	8	1.149	0.689
Defense mechanisms	3	1.177	0.741

Table C.21: COG categories with genes under positive selection in the January sample for J07HR59. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	2	1.828	0.328
Amino acid transport and metabolism	2	0.999	1.000
Cell cycle control, cell division, chromosome partitioning	1	11.232	0.102
Nucleotide transport and metabolism	1	1.656	0.472
Lipid transport and metabolism	1	1.656	0.472
Coenzyme transport and metabolism	4	4.326	**0.025**
Transcription	1	1.375	0.534
Replication, recombination and repair	1	0.902	1.000
Post-translational modification, protein turnover, and chaperones	1	1.402	0.528
Inorganic ion transport and metabolism	1	1.349	0.541
Intracellular trafficking, secretion, and vesicular transport	1	5.585	0.182
Defense mechanisms	1	5.585	0.182

Table C.22: COG categories with genes under positive selection in the August sample for J07HR59. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	3	2.165	0.191
Amino acid transport and metabolism	3	1.183	0.738
Cell cycle control, cell division, chromosome partitioning	1	8.558	0.129
Nucleotide transport and metabolism	1	1.262	0.562
Lipid transport and metabolism	1	1.262	0.562
Coenzyme transport and metabolism	4	3.124	0.058
Transcription	2	2.200	0.254
Translation, ribosomal structure and biogenesis	1	0.459	0.714
Replication, recombination and repair	1	0.687	1.000
Post-translational modification, protein turnover, and chaperones	1	1.068	0.621
Inorganic ion transport and metabolism	1	1.028	1.000
Function unknown	1	0.416	0.718
Intracellular trafficking, secretion, and vesicular transport	1	4.255	0.229
Defense mechanisms	1	4.255	0.229

Table C.23: COG categories with genes under positive selection in the January sample for A07HB70. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	Count	Odds ratio	pvalue
Energy production and conversion	8	0.943	1.000
Chromatin structure and dynamics	1	5.003	0.229
Amino acid transport and metabolism	20	1.500	0.113
Carbohydrate transport and metabolism	8	1.317	0.403
Nucleotide transport and metabolism	4	0.915	1.000
Lipid transport and metabolism	4	1.152	0.777
Coenzyme transport and metabolism	3	0.396	0.156
Transcription	2	0.296	0.086
Translation, ribosomal structure and biogenesis	8	0.787	0.610
Cell wall/membrane/envelope biogenesis	1	0.293	0.371
Replication, recombination and repair	4	0.630	0.519
Post-translational modification, protein turnover, and chaperones	6	0.994	1.000
Cell motility	1	0.994	1.000
Secondary metabolites biosynthesis, transport, and catabolism	4	2.534	0.094
Inorganic ion transport and metabolism	8	1.123	0.693
Function unknown	12	0.994	1.000
General function prediction only	23	1.224	0.377
Intracellular trafficking, secretion, and vesicular transport	1	0.618	1.000
Signal transduction mechanisms	7	1.498	0.335
Defense mechanisms	1	0.827	1.000

Table C.24: COG categories with genes under positive selection in the August sample for A07HB70. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	Count	Odds ratio	pvalue
Energy production and conversion	9	1.043	0.856
Chromatin structure and dynamics	1	4.885	0.233
Amino acid transport and metabolism	21	1.545	0.086
Carbohydrate transport and metabolism	7	1.114	0.678
Nucleotide transport and metabolism	4	0.893	1.000
Lipid transport and metabolism	4	1.124	0.780
Coenzyme transport and metabolism	3	0.387	0.112
Transcription	2	0.289	0.088
Translation, ribosomal structure and biogenesis	8	0.768	0.611
Cell wall/membrane/envelope biogenesis	1	0.286	0.253
Replication, recombination and repair	4	0.615	0.521
Post-translational modification, protein turnover, and chaperones	7	1.141	0.672
Cell motility	1	0.971	1.000
Secondary metabolites biosynthesis, transport, and catabolism	4	2.473	0.101
Inorganic ion transport and metabolism	8	1.095	0.844
Function unknown	12	0.968	1.000
General function prediction only	24	1.252	0.318
Intracellular trafficking, secretion, and vesicular transport	1	0.604	1.000
Signal transduction mechanisms	7	1.461	0.342
Defense mechanisms	1	0.808	1.000

Table C.25: COG categories with genes under positive selection in the January sample for A07HR67. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	6	0.657	0.469
Amino acid transport and metabolism	22	1.234	0.370
Cell cycle control, cell division, chromosome partitioning	2	1.813	0.326
Carbohydrate transport and metabolism	11	1.376	0.354
Nucleotide transport and metabolism	7	1.431	0.349
Lipid transport and metabolism	5	1.380	0.420
Coenzyme transport and metabolism	4	0.529	0.319
Transcription	1	0.132	**0.013**
Translation, ribosomal structure and biogenesis	6	0.549	0.182
Cell wall/membrane/envelope biogenesis	5	1.131	0.802
Replication, recombination and repair	8	1.225	0.540
Post-translational modification, protein turnover, and chaperones	5	0.774	0.832
Cell motility	4	2.305	0.118
Secondary metabolites biosynthesis, transport, and catabolism	3	1.319	0.504
Inorganic ion transport and metabolism	8	0.996	1.000
Function unknown	8	0.540	0.107
General function prediction only	32	1.501	0.062
Intracellular trafficking, secretion, and vesicular transport	5	2.442	0.072
Signal transduction mechanisms	3	0.430	0.211
Defense mechanisms	3	1.783	0.419

Table C.26: COG categories with genes under positive selection in the August sample for A07HR67. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	6	0.648	0.374
Amino acid transport and metabolism	23	1.280	0.308
Cell cycle control, cell division, chromosome partitioning	2	1.789	0.332
Carbohydrate transport and metabolism	11	1.356	0.357
Nucleotide transport and metabolism	7	1.411	0.354
Lipid transport and metabolism	5	1.361	0.425
Coenzyme transport and metabolism	4	0.522	0.241
Transcription	1	0.130	**0.014**
Translation, ribosomal structure and biogenesis	6	0.541	0.183
Cell wall/membrane/envelope biogenesis	5	1.115	0.803
Replication, recombination and repair	8	1.208	0.544
Post-translational modification, protein turnover, and chaperones	5	0.763	0.680
Cell motility	4	2.273	0.122
Secondary metabolites biosynthesis, transport, and catabolism	3	1.301	0.509
Inorganic ion transport and metabolism	8	0.982	1.000
Function unknown	9	0.603	0.191
General function prediction only	32	1.475	0.065
Intracellular trafficking, secretion, and vesicular transport	5	2.408	0.075
Signal transduction mechanisms	3	0.424	0.154
Defense mechanisms	3	1.759	0.421

Table C.27: COG categories with genes under positive selection in the January sample for A07HN63. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	2	0.824	1.000
Amino acid transport and metabolism	6	2.020	0.135
Carbohydrate transport and metabolism	2	1.507	0.399
Lipid transport and metabolism	1	1.349	0.535
Coenzyme transport and metabolism	4	1.844	0.287
Transcription	4	1.614	0.327
Cell wall/membrane/envelope biogenesis	2	3.433	0.130
Replication, recombination and repair	3	1.670	0.432
Post-translational modification, protein turnover, and chaperones	1	0.539	1.000
Secondary metabolites biosynthesis, transport, and catabolism	1	2.519	0.344
Inorganic ion transport and metabolism	2	0.887	1.000
Function unknown	1	0.226	0.174
General function prediction only	8	1.658	0.233
Signal transduction mechanisms	1	0.347	0.517

Table C.28: COG categories with genes under positive selection in the August sample for A07HN63. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	Count	Odds ratio	pvalue
Energy production and conversion	2	0.706	1.000
Amino acid transport and metabolism	7	2.039	0.098
Carbohydrate transport and metabolism	2	1.291	0.670
Nucleotide transport and metabolism	1	0.621	1.000
Lipid transport and metabolism	1	1.161	0.588
Coenzyme transport and metabolism	5	2.009	0.188
Transcription	3	1.004	1.000
Cell wall/membrane/envelope biogenesis	2	2.943	0.164
Replication, recombination and repair	4	1.948	0.276
Post-translational modification, protein turnover, and chaperones	2	0.949	1.000
Secondary metabolites biosynthesis, transport, and catabolism	1	2.167	0.386
Inorganic ion transport and metabolism	3	1.169	0.742
Function unknown	1	0.195	0.080
General function prediction only	8	1.382	0.381
Intracellular trafficking, secretion, and vesicular transport	1	2.283	0.371
Signal transduction mechanisms	1	0.299	0.366

Table C.29: COG categories with genes under positive selection in the January sample for A07HR60. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	5	1.064	0.809
Amino acid transport and metabolism	8	1.051	0.846
Carbohydrate transport and metabolism	5	1.803	0.214
Nucleotide transport and metabolism	2	0.900	1.000
Lipid transport and metabolism	5	1.851	0.207
Transcription	6	1.487	0.310
Translation, ribosomal structure and biogenesis	3	0.552	0.483
Cell wall/membrane/envelope biogenesis	4	2.094	0.144
Replication, recombination and repair	6	0.863	1.000
Post-translational modification, protein turnover, and chaperones	4	1.287	0.556
Secondary metabolites biosynthesis, transport, and catabolism	3	1.780	0.257
Inorganic ion transport and metabolism	8	2.183	0.060
Function unknown	6	0.878	1.000
General function prediction only	4	0.323	**0.018**
Intracellular trafficking, secretion, and vesicular transport	1	1.275	0.556
Signal transduction mechanisms	1	0.732	1.000
Defense mechanisms	2	2.184	0.251

Table C.30: COG categories with genes under positive selection in the August sample for A07HR60. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	5	1.019	1.000
Amino acid transport and metabolism	8	1.005	1.000
Carbohydrate transport and metabolism	6	2.102	0.125
Nucleotide transport and metabolism	2	0.863	1.000
Lipid transport and metabolism	5	1.773	0.220
Transcription	6	1.423	0.445
Translation, ribosomal structure and biogenesis	4	0.716	0.654
Cell wall/membrane/envelope biogenesis	3	1.485	0.462
Replication, recombination and repair	6	0.826	0.841
Post-translational modification, protein turnover, and chaperones	6	1.903	0.147
Secondary metabolites biosynthesis, transport, and catabolism	3	1.707	0.427
Inorganic ion transport and metabolism	8	2.086	0.067
Function unknown	6	0.840	0.841
General function prediction only	4	0.309	**0.013**
Intracellular trafficking, secretion, and vesicular transport	1	1.224	0.570
Signal transduction mechanisms	1	0.703	1.000
Defense mechanisms	2	2.096	0.266

Table C.31: COG categories with genes under positive selection in the January sample for J07SB. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	**Count**	**Odds ratio**	**pvalue**
Energy production and conversion	4	0.801	1.000
Amino acid transport and metabolism	6	0.966	1.000
Carbohydrate transport and metabolism	4	1.248	0.565
Nucleotide transport and metabolism	2	0.702	1.000
Lipid transport and metabolism	2	0.806	1.000
Coenzyme transport and metabolism	1	0.243	0.172
Transcription	7	1.559	0.322
Translation, ribosomal structure and biogenesis	8	1.648	0.236
Cell wall/membrane/envelope biogenesis	5	1.288	0.592
Replication, recombination and repair	6	1.539	0.298
Post-translational modification, protein turnover, and chaperones	4	1.194	0.771
Cell motility	1	0.451	0.717
Secondary metabolites biosynthesis, transport, and catabolism	3	1.489	0.463
Inorganic ion transport and metabolism	3	1.015	1.000
Function unknown	3	0.583	0.473
General function prediction only	9	1.302	0.423
Intracellular trafficking, secretion, and vesicular transport	3	1.544	0.452
Signal transduction mechanisms	3	0.527	0.357

Table C.32: COG categories with genes under positive selection in the August sample for J07SB. The pvalue for each category was calculated using the Odds Ratio and a one-tailed Fisher exact test

COG category	Count	Odds ratio	pvalue
Energy production and conversion	4	0.790	0.812
Amino acid transport and metabolism	6	0.952	1.000
Carbohydrate transport and metabolism	4	1.231	0.570
Nucleotide transport and metabolism	2	0.692	1.000
Lipid transport and metabolism	2	0.795	1.000
Coenzyme transport and metabolism	2	0.487	0.424
Transcription	7	1.536	0.326
Translation, ribosomal structure and biogenesis	7	1.400	0.357
Cell wall/membrane/envelope biogenesis	5	1.269	0.595
Replication, recombination and repair	6	1.516	0.303
Post-translational modification, protein turnover, and chaperones	4	1.177	0.773
Cell motility	1	0.445	0.718
Secondary metabolites biosynthesis, transport, and catabolism	3	1.468	0.467
Inorganic ion transport and metabolism	3	1.001	1.000
Function unknown	4	0.778	0.812
General function prediction only	9	1.283	0.544
Intracellular trafficking, secretion, and vesicular transport	3	1.522	0.456
Signal transduction mechanisms	3	0.520	0.357

Appendix D

Genome pN/pS coverage plots

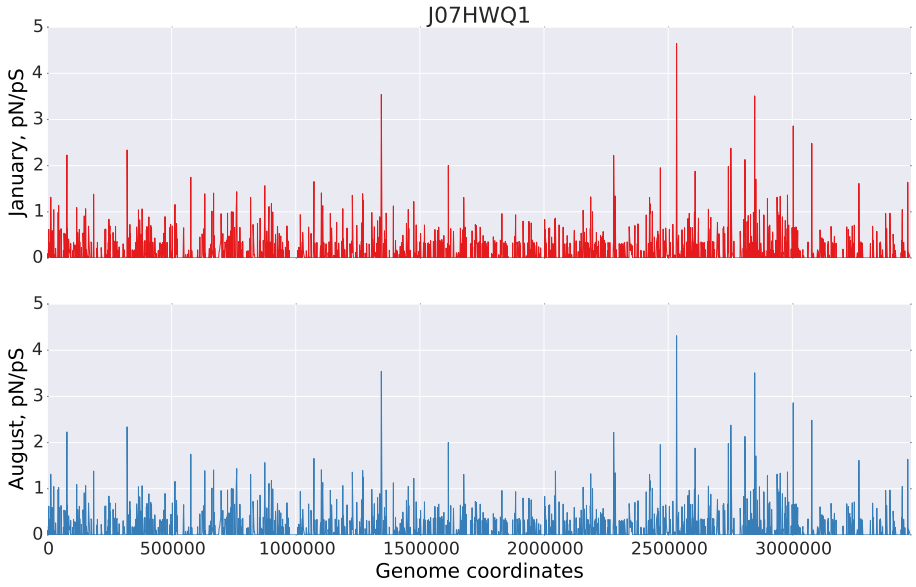


Figure D.1: pN/pS values for each gene in the J07HWQ1 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample.

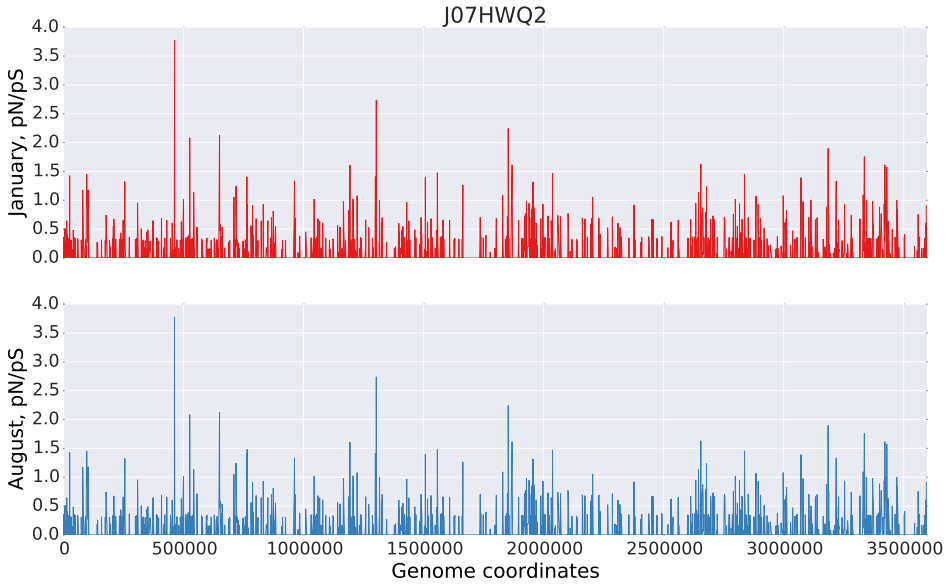


Figure D.2: pN/pS values for each gene in the J07HWQ2 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample

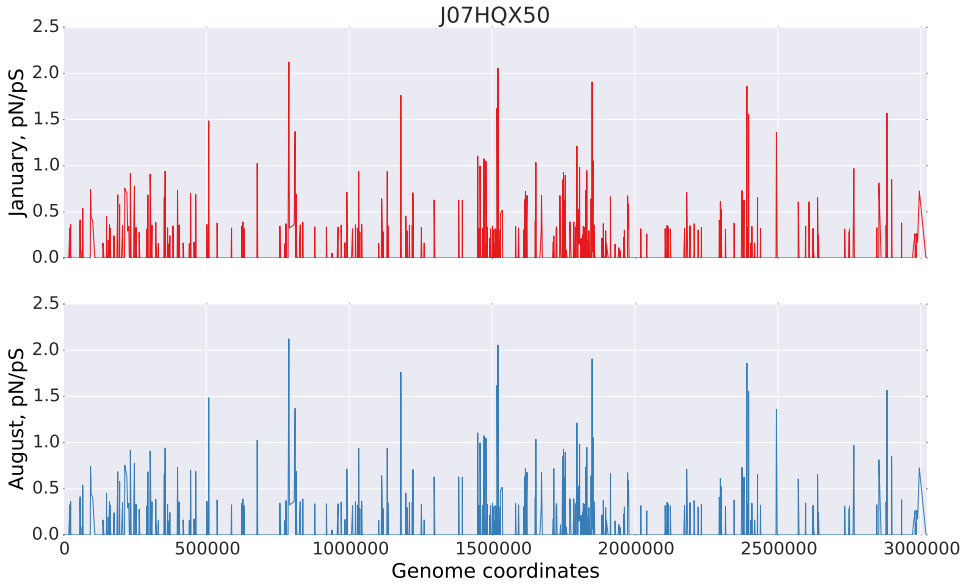


Figure D.3: pN/pS values for each gene in the J07HQX50 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample

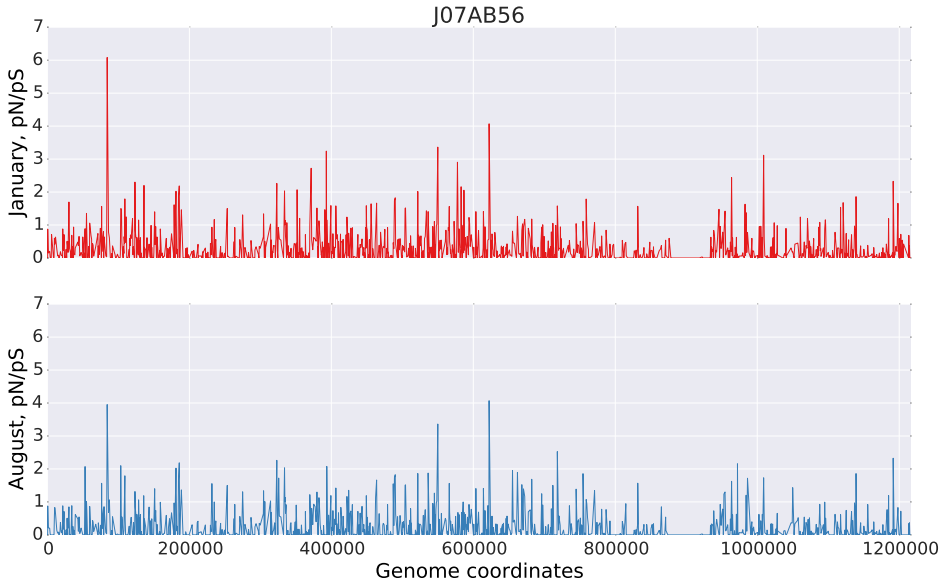


Figure D.4: pN/pS values for each gene in the J07AB56 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample

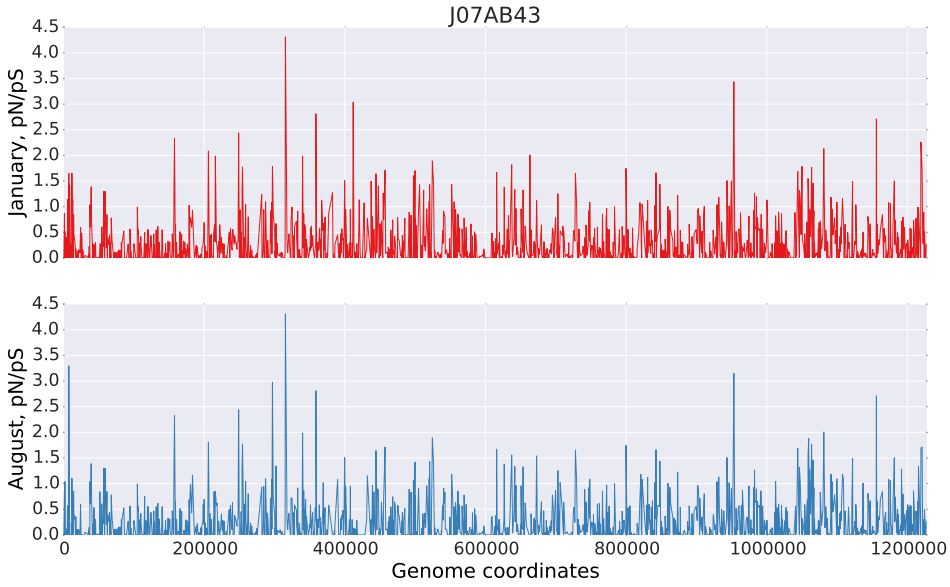


Figure D.5: pN/pS values for each gene in the J07AB56 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample

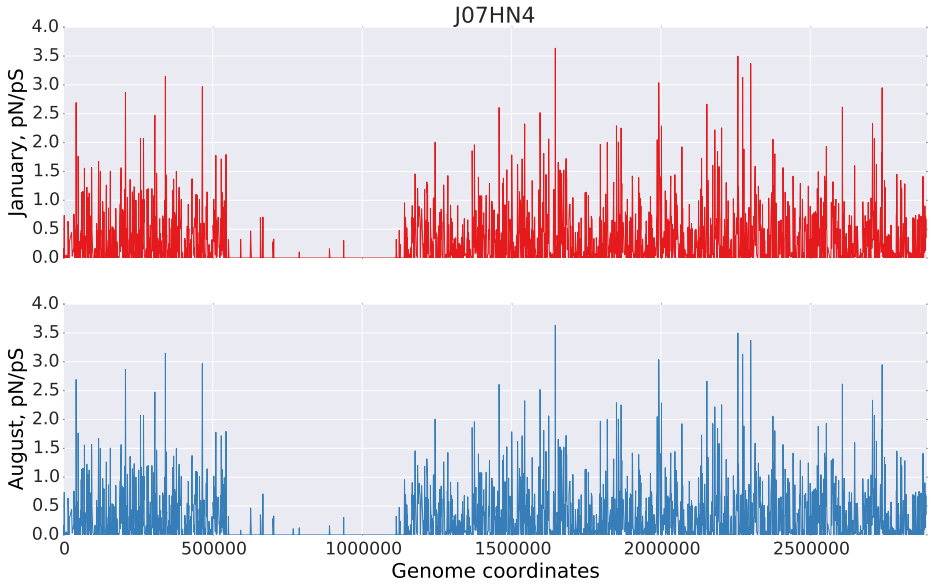


Figure D.6: pN/pS values for each gene in the J07HN4 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample

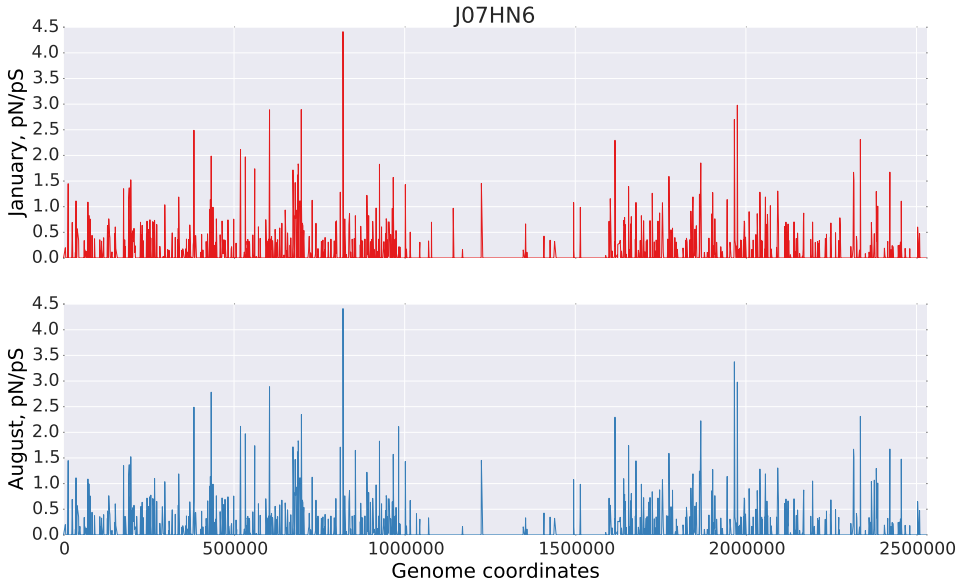


Figure D.7: pN/pS values for each gene in the J07HN6 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample

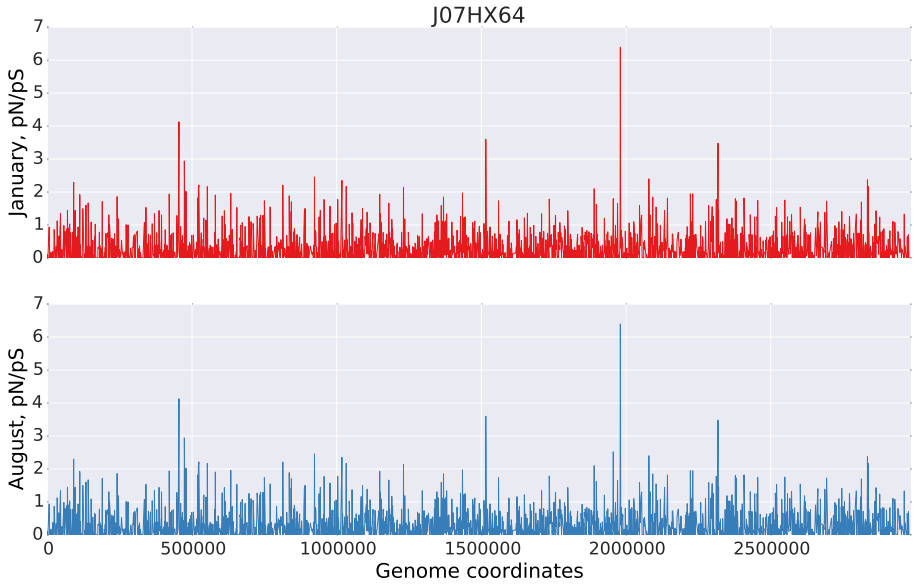


Figure D.8: pN/pS values for each gene in the J07HX64 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample

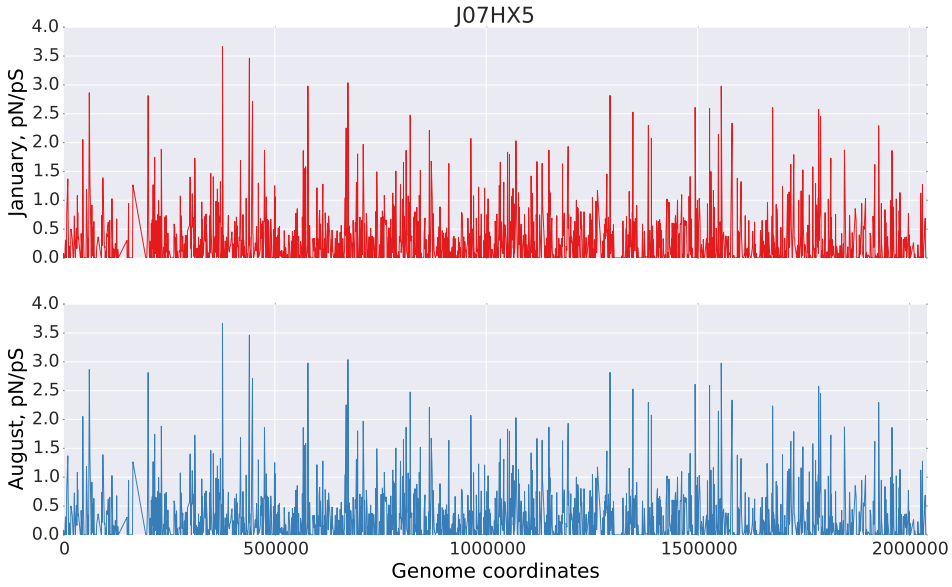


Figure D.9: pN/pS values for each gene in the J07HX5 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample

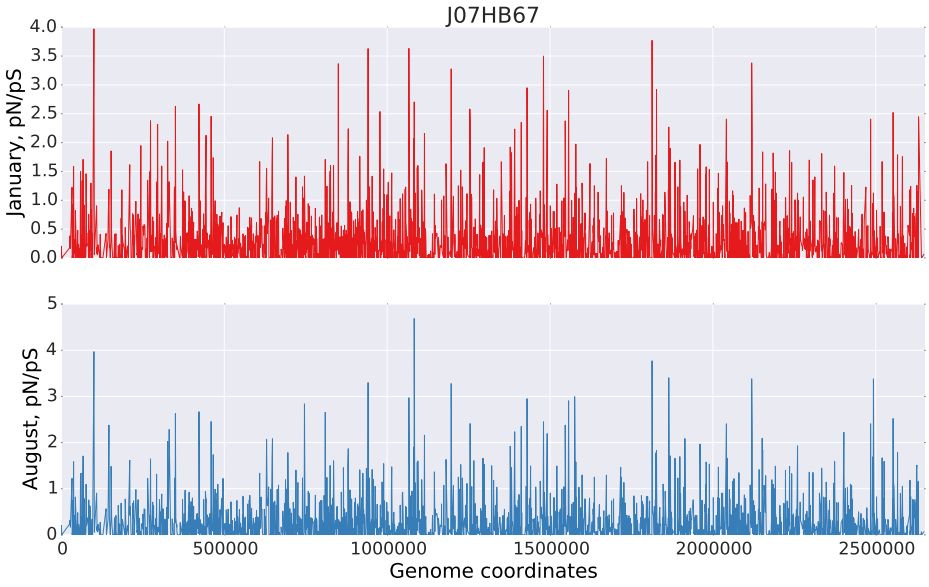


Figure D.10: pN/pS values for each gene in the J07HB67 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample

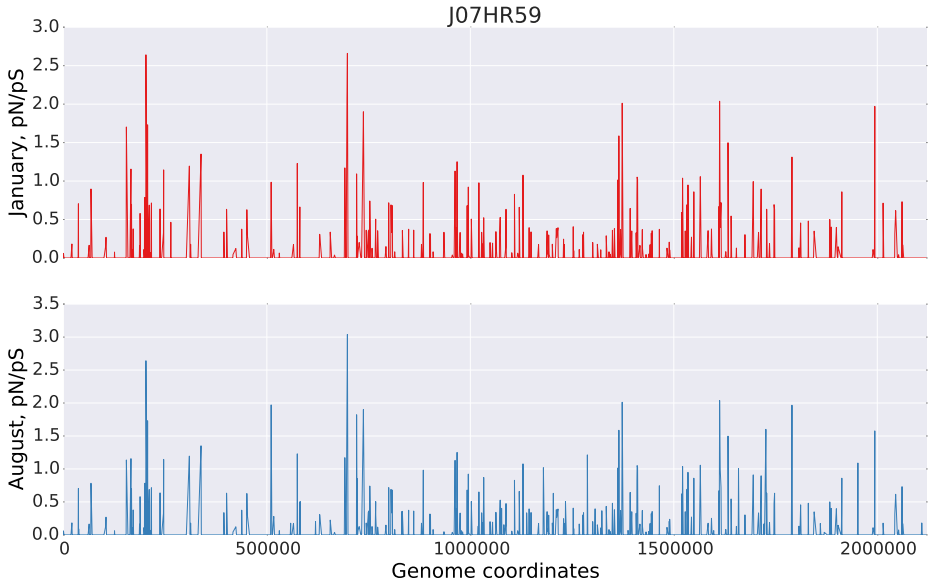


Figure D.11: pN/pS values for each gene in the J07HR59 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample

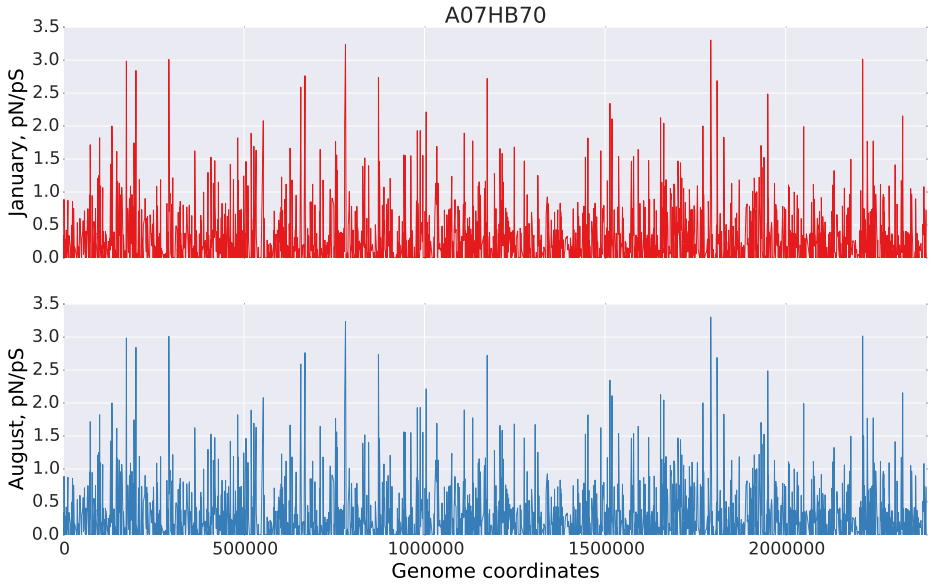


Figure D.12: pN/pS values for each gene in the A07HB70 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample

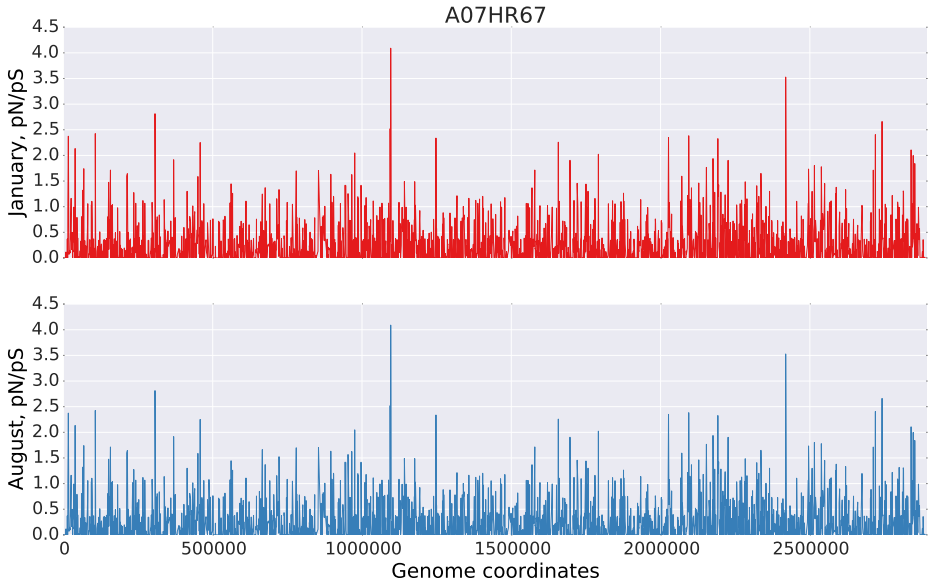


Figure D.13: pN/pS values for each gene in the A07HR67 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample

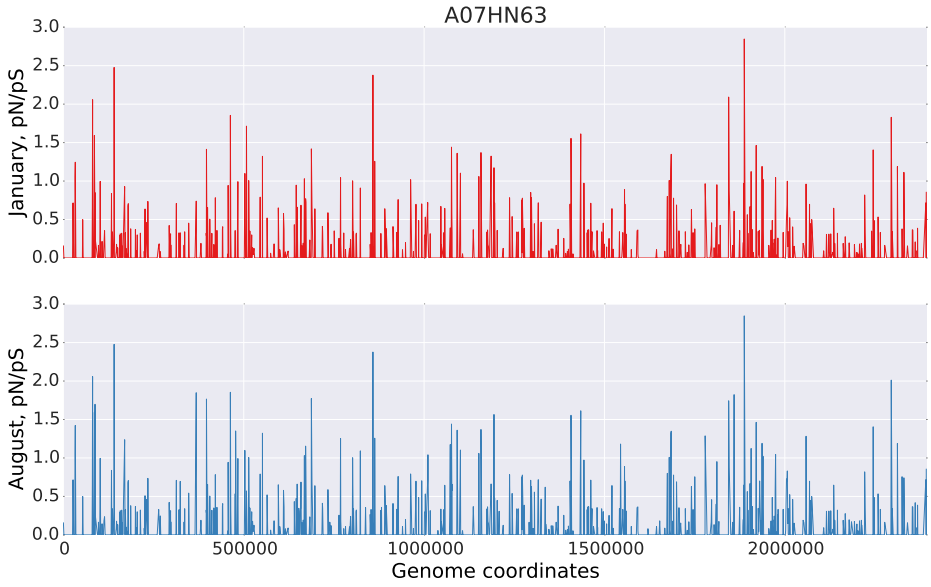


Figure D.14: pN/pS values for each gene in the A07HN63 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample

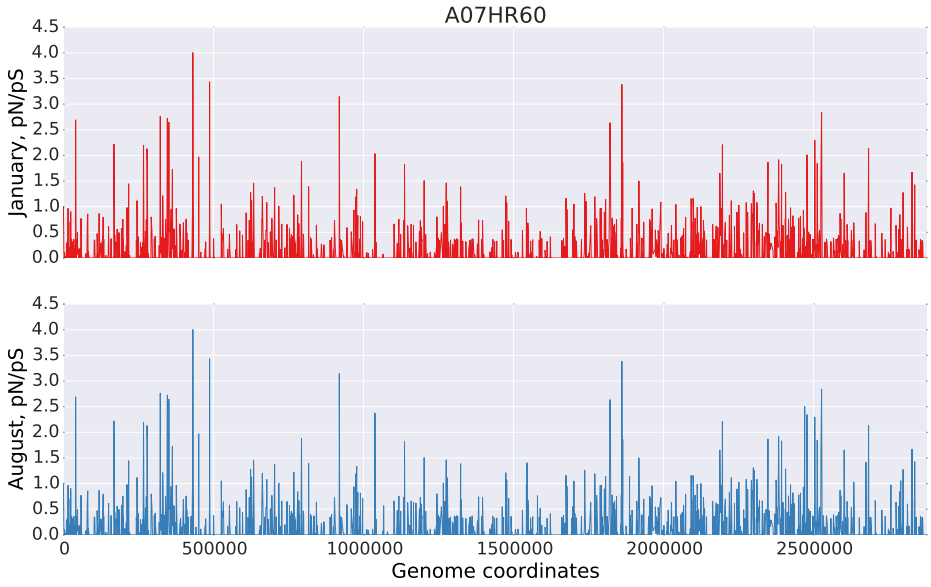


Figure D.15: pN/pS values for each gene in the A07HR60 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample

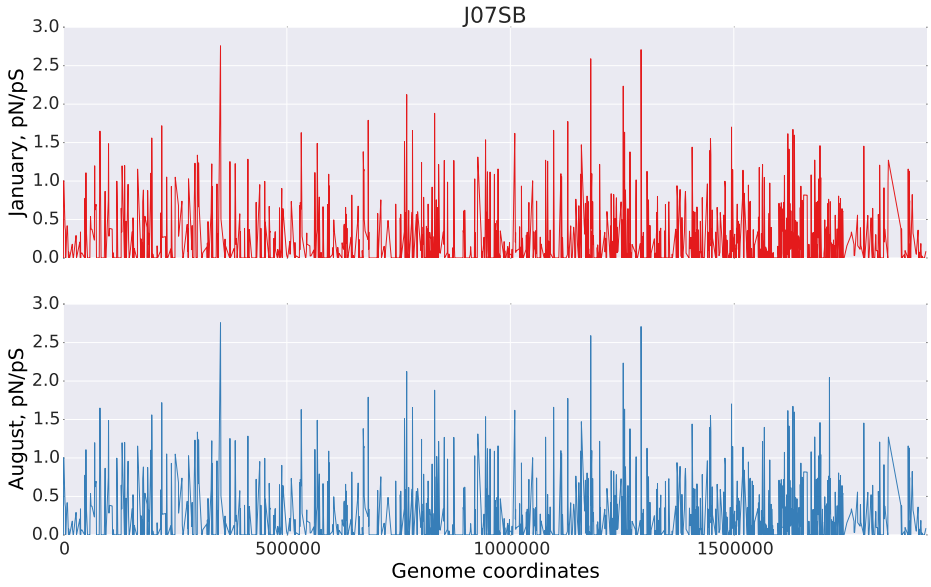


Figure D.16: pN/pS values for each gene in the J07SB genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample

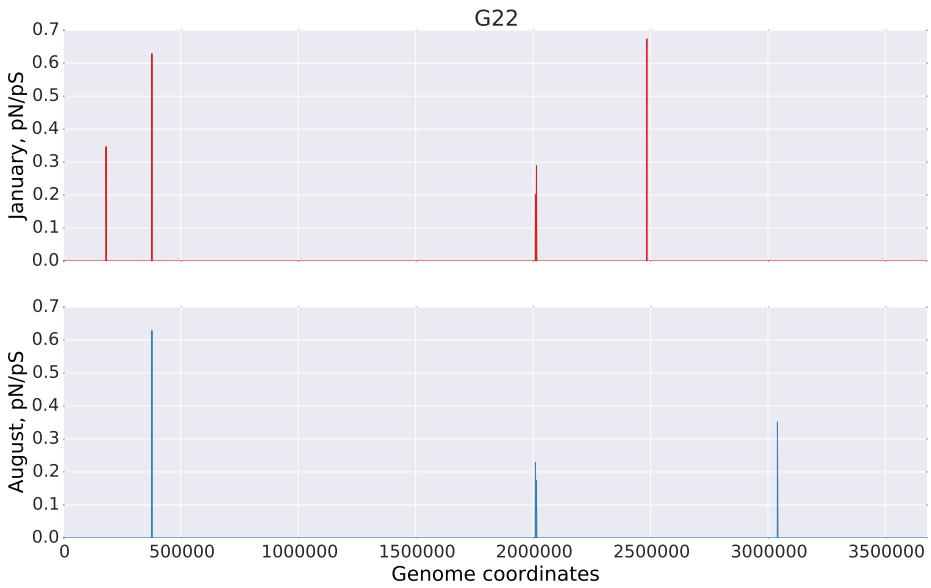


Figure D.17: pN/pS values for each gene in the G22 genome. Top panel shows the values using the reads from the January samples. Bottom panel shows the values using the reads from the August sample