Chapter 1

Introduction to the thesis

Bacteria and Archaea represent an abundant component of Earths biomass, with estimates between 9.2-31.7 10^{29} cells [29] to 41.8-64.3 10^{29} cells [74]. Species diversity estimates suggest that there are close to 10^7 different Bacteria and Archaeal species [15], although some authors suggest that these could be an overestimation [57]. Nevertheless, Bacteria and Archaea represent the most diverse group of organisms on Earth, and yet we barely have scratched the surface on characterizing such diversity [78, 54].

Our understanding of the phylogenetic and functional diversity of natural microbial communities is limited by our inability to grow most of the microorganisms that are present in the environment. This phenomenon, known as the Great Plate-Count Anomaly [63], shows that current culture collections may not be representatives of what can be seen in natural samples by culture-independent methods [3]. Most of these organisms are currently uncultured, due to our lack of information on their physiology and nutritional requirements [64], but unique cultivation methods and information derived from genomic surveys [66] could help us to change this in the near future. Over the last decade, the development of newer technologies and experimental methods, have allowed us to study natural microbial communities, via culture-independent techniques. In particular, DNA-based methods, driven by the development of what has been called next-generation sequencing [35], has allowed us to investigate natural microbial communities without relying on the cultivation of its members. DNA-based surveys can be divided into two broad categories; marker-based analysis (such as the 16S rRNA gene) and whole DNA sequencing.

Marker-based analysis provides a characterization of the phylogenetic diversity that is present in the community. It allows to estimate the diversity of the community [13, 52, 12] and it can allow the identification of the members of the community based on sequence similarity against a reference database [37]. This approach only provides a partial picture of the metabolic repertoire present in natural microbial communities, because it requires the extrapolation of results from the genomes of isolated microorganisms to the community under study. 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 16S rRNA gene [26].

Metagenomics (also referred as community genomics), relies on the direct sequencing of genetic material isolated from a microbial community [77], which allows to evaluate the phylogenetic and functional diversity, without culturing the microorganisms that are present. Depending on the complexity of the community under study, not all of the members of the microbial community will be observed on the sequence data, as the most abundant members will dominate the metagenomic reads [8]. Several examples can be found of studies where communities with a low to moderate species diversity have been studied using metagenomic methods (e.g. [67, 51, 18, 65, 11]). Highly diverse communities (such as the ones found in soil), require high amounts of sequence information to provide a full picture of the microorganisms that are present [28]. The decrease in sequences cost, driven by the development of novel sequencing technologies, will increase our ability to deeply sample complex communities using metagenomics, where then the challenge will be the computational resources and algorithms needed to analyze vast amounts of information [49, 28]. Microbial communities in extreme environments (such as hypersaline waters) represent tractable systems that can be studied using metagenomic approaches. Because of their relatively low species diversity [5, 51, 24], it is possible to study the phylogenetic and metabolic diversity that is present in the community. In this first chapter, I will provide an overview on metagenomics, with a particular emphasis in assembly-based approximations. This will be followed by an introduction into the microbiology of hypersaline environments, followed by an overview of the study site, the hypersaline waters of Lake Tyrrell in 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 [25], to access the genomic information of uncultured microorganisms in environmental samples. The first metagenomic studies were performed by isolating DNA from environmental samples, cloning and then sequencing these clones using the Sanger method [65, 70]. Technological limitations, make this process expensive and laborious, and it most of the cases is limited either to study microbial communities from low to medium species diversity [65, 51], or to get an overview of the general functional and phylogenetic composition of a community [70, 56]. The development of novel sequencing technologies (”next-generation sequencing”) [36], has removed some of these limitations, and now it is possible to sequence DNA samples directly (no cloning required), with higher-throughputs and a lower cost per base [36]. One of the remaining limitations in current technologies, is the length of the sequence reads, not yet close to the length of reads generated by Sanger sequencing. Newer technological developments are changing this, including the improvement of current methods and the apparition of newer technologies, such as single molecule sequencing [19] and nanopores [9].

Metagenomic studies are driven, in most cases, by discovery, data mining and comparative research, rather than by a specific hypothesis [77]. For this, not only the sequence information is needed, but also all the metadata that is associated with a sample, which allows to put into context (both biological and physical) the microbial community under study. This includes several pieces of information, such as environmental parameters associated with a sample; procedures used to process a sample (e.g.: filtration, centrifugation). All of these become very important information that can be used in the bioinformatic analysis of a sample [41].

Based on the complexity of the microbial community under study, and the choice of sampling and sequencing methods, two main approaches have been used to study microbial communities using metagenomics; gene-centric and assembly-based approaches (Figure 1.1). Gene-centric studies have focused in complex communities (or with a low-depth of sequencing), where assembly of genome fragments from members of the community is not feasible given the number of reads obtained. In these cases, the focus is in the description of the phylogenetic and functional diversity of a community [67, 11], and the comparison among different environments based on this profiles [18, 76, 53]. Gene-centric studies focus in the phylogenetic and functional classification of the sequence reads, either by analyzing every read present in the dataset [7, 47, 4] or via marker-genes that can be used as a proxy to create either taxonomic and/or functional profiles of the community [16, 59]. One of the main drawbacks of this method, is that community profiles are limited to what is available in the reference databases being used, which can lead to missing unique groups that are present in the microbial community under study. For example, Figure 1.2 shows a comparison on the classification of a set of metagenomic reads using two different reference datasets, one that has the novel group that were recovered by assembling metagenomic reads [41, 51] and the other that do not have those genomes. [Waiting on final results for this].

Assembly based metagenomics, consists in the assembly of the sequence information, with the goal of recovering larger fragments that can improve the phylogenetic and functional classification of the organisms in the community. This approach allows to obtain longer fragments, generating better gene models and the recovery of novel groups from the metagenomic sample [41, 51, 30, 17]. By recovering these genomes from a metagenomic sample, it is possible to characterize previously unknown microbial groups in a natural microbial community. This information can be used to cultivate these organisms [66], population genetic studies [62, 73, 2] and to provide a more complete pictures of the interactions among members of a microbial community [65, 22]. In addition, the recovery of near-complete genomes from metagenomics datasets, allows the discovery of novel functions that could be easy missed by just looking at the unassembled reads. In this thesis, the assembly of the reads from the Lake Tyrrell metagenome, led to the discovery of a previously unknown Class of Archaea (Chapter 2), and also the discovery of a novel type of microbial rhodopsins, named as Xenorhodopsin, with amino acid signature different from previously described microbial rhodopsins (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. Even in this case, some of the rarer members of the community are unapproachable by sequencing, because the metagenomic dataset will be dominated by those most abundant organisms. For example, estimations of the number of sequences needed to address highly diverse samples, like soil, show that billion of sequence are needed, to be able to sample some of these most abundant organisms 1.3. The main challenges, even in simpler systems, is the classification of the assembled genomic fragments into phylogenetic unique bins, each representing a different population. By combining various pieces of information, such as sample characteristics, G+C percentage, amino acid counts and library abundance, it is possible to classify these genomic fragments into unique populations [51, 1]. Larger datasets represent a computational challenge, because of the high memory requirements of some of the assembly software used. In this case, the use of methods for digital normalization, has showed to reduce this computational problems [49, 28].

The work presented in this thesis focuses on the study of the microbial community that inhabits the hypersaline waters of Lake Tyrrell, Australia, using an assembly-based approximation. The combination of a relatively low-species diversity, driven by the extreme conditions found in this habitat [5] and a deep sequencing approach, allowed the reconstruction of some of the most abundant members of the community and the discovery of novel microorganisms.

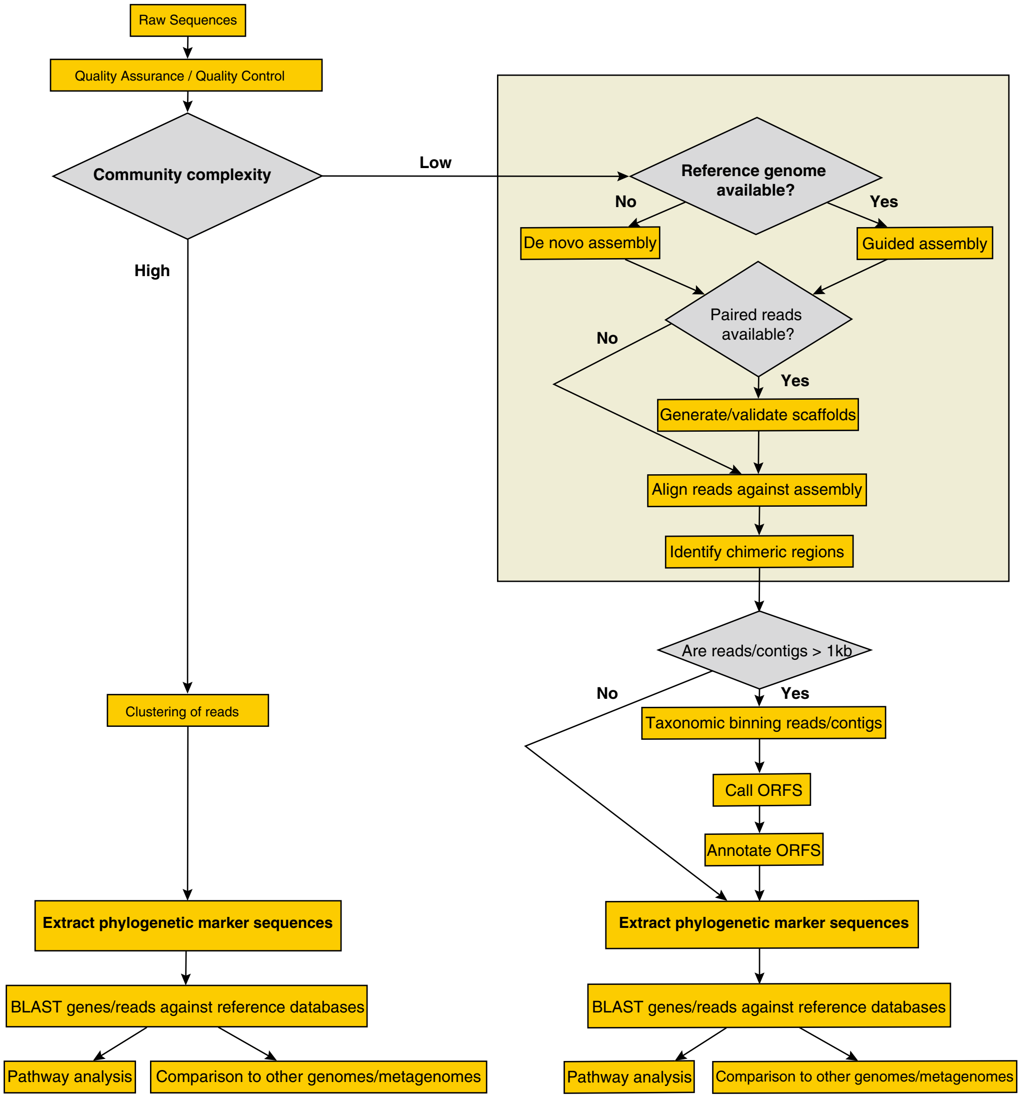
Raw Sequences
Quality Assurance / Quality Control
Community complexity
Low
High
Reference genome available?
No
Yes
De novo assembly
Guided assembly
Paired reads available?
No
Yes
Generate/validate scaffolds
Align reads against assembly
Identify chimeric regions
Are reads/contigs > 1kb
No
Yes
Taxonomic binning reads/contigs
Call ORFS
Annotate ORFS
Clustering of reads
Extract phylogenetic marker sequences
Extract phylogenetic marker sequences
BLAST genes/reads against reference databases
BLAST genes/reads against reference databases
Pathway analysis
Comparison to other genomes/metagenomes
Pathway analysis
Comparison to other genomes/metagenomes

Figure 1.1: Diagram showing two approaches for the analysis of metagenomic data from natural microbial communities. Figure from Bragg and Tyson, 2014 [8].

*

Figure 1.2: Comparison of a set of reads (Illumin XXbp, total of XX reads) classified using Phylosift? with two different reference databases. The first case shows the reads classified in abscence of any reference genomes, the second one where reference genomes obtained through metagenomics [41, 51] are included in the database

*

Figure 1.3: Sequencing required for *de novo* assembly using Illumina short reads. Numbers based on REF

1.2 Microbial communities in hypersaline environments

Hypersaline environments can be found in multiple places around the globe, with different types of saline systems. Some examples include salt lakes, saline soils, salt flats, solar salterns, brine pools, salted foods and fermented foods (REF, Oren Halophilic microorganisms and their environments) . Aquatic systems are the ones more studied, which can be 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 in the range of 150-300 g/L NaCl [5]. Besides the high NaCl concentrations, other salts are also important to consider when taking in account the ionic composition of these systems, including elements such as Mg2+ and Ca+2, which can influence the microbial community that will be present in these habitats [38, 50]. [Show Figure 19.1 from Aron] Figure 1.6.

Both Bacterial and Archaeal genera can be found in moderate and extreme saline systems. Within the Archaea, the phylogenetic diversity appears to be limited to the *Euryarchaeota*, in particular in the classes *Halobacteria* and *Methanomicrobia* (Table 1.1) [71]. The discovery of the *Nanohaloarchaea* (described in Chapter 2, and [41]) increased the phylogenetic diversity in halophilic Archaea, by describing a novel Class, within the *Euryarchaeota*, that is present in hypersaline systems, in several parts of the world [41]. Recently, a phylogenomic analysis that included novel Archeal groups identified via single-cell genomics, suggested that the *Nanohaloarchaea* are a new phylum, sister to the *Euryarchaeota* [54], although more work (and more genomes and isolates) are needed to fully resolve the phylogenetic relationships between these groups [75].

Even within the *Halobacteria*, there is still room for discovery of novel taxonomic groups. Our group recently described the genome of a newly isolated halophilic Archaea, *Candidatus* Halobonum tyrrellensis strain G22 [68], which phylogenetic analysis suggests is a new genus within the *Halobacteria* (Appendix A). These phylogenetic placement was supported by analysis done using the 16S rRNA gene (Figure 1.4) and a phylogenomic approach using the markers genes implemented in the software PhyloPhlan [58] (Figure 1.5)

The breadth of Bacterial species that can be found in saline systems is wider than in the case of the Archaea (Table 1.2). In moderate saline environments, Bacteria can be even more abundant that Archaeal species [42, 24, 23, 14], but as salinity increases, Archaeal groups become more abundant [24, 14, 40]. One of the most abundant bacterial species found in extreme hypersaline systems is *Salinibacter ruber* [6]. This bacterium shares many phenotypical characteristics of halophilic Archaea [46] and multiple gene clusters appear to be acquired via horizontal gene transfer from Archaea [39].

Microorganisms that live in hypersaline systems require strategies to deal with the high salt concentrations and the osmotic pressure that generates. Two different strategies can be found in halophiles; a "salt-in" strategy, that involves the accumulation of inorganic ions in the cytoplasm, and a "salt-out" strategy, that pumps out ions from the cytoplasm [45]. The "salt-in" strategy is found in the Archaeal populations, in particular within the order *Halobacteriales* and in Bacteria in the order *Halanerobiales* and also in *S. ruber* [45]. These organisms accumulate inorganic ions, such as K^+ and Cl^-. This requires special adaptations in the enzymes of these organisms, which is reflected in their amino acid composition, where most of the proteins show an increase in the acidic amino acid, such as glutamate and aspartate and a decrease in basic amino acids, such as lysine and arginine. Based on this amino acid composition profile, it has been suggested that the uncultured members of the Class *Nanohaloarchaea* use this strategy, as well. The "salt-out" approach is found in most of the halophilic Bacteria and halophilic methanogenic Archaea [45]. 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 keep the osmotic potential in the cytoplasm [45].

Halophilic microorganisms show a diverse range of metabolic processes, both in the lab and natural communities. Several dissimilatory metabolic pathways have been described in halophiles (Figure 1.7), in a variety of salinity ranges. The majority of Bacterial and Archeal groups isolated from hypersaline sites are aerobic chemoorganotrophs. In addition, as oxygen, has a low solubility in salt saturated waters [61], several anaerobic energy pathways are available, including substrates such as nitrate, fumarate and dimethyl sulfoxide as electron acceptors [42, 45]. 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* and in high salt environments, the primary producers are planktonic algae of the genus *Dunaliella* [43].

Another important characteristic of halophilic organisms, particularly in the Archaea, is the presence of microbial rhodopsins, photoactive proteins that can be found in all three domains of life. These proteins can serve as light-drive proton pumps (bacteriorhodopsin), chloride pumps (halorhodopsin) or phototactic and photophobic receptors (sensory rhodopsins) [10]. Halorhodopsins and sensory rhodopsins appear to be limited in their distribution to the *Halobacteria*, with just a few examples found in other organisms [60]. Chapter 3, describes the discovery a novel type of microbial rhodopsin, named Xenorhodopsin, which was found in the genome of the *Nanohaloarchaea*, and appear to be the subject of horizontal gene transfer (although it is not possible yet to establish the direction) between Archaea and Bacteria, as it closest homolog is a putative sensory rhodopsin found in the bacterium *Anabaena* sp. PCC 7120 [72, 69].

In this section, I wanted to provide a broad overview of saline systems and their microbiology. Even when they have been highly characterized [42, 44], recent studies using culture independent techniques such as metagenomics and single-cell genomics, have recovered novel, previously uncharacterized microbial groups [41, 24, 33]. The limited microbial diversity, driven by the salinity extremes found in these systems, makes them ideal model systems to study microbial diversity using metagenomic methods, as we can fully characterize the community using deep-sequencing approaches, with the goal of reconstructing the genomes of the members of the community. This allows not only to identify the functional and phylogenetic diversity of the community, but also the association of such functions to members of the community. Also, 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. These studies focused on the population genomics of single species, such as *Haloquadratum walsbyi* [32] and *Salinibacter ruber* [48], or in the dynamic changes of the microbial and viral diversity over salinity gradientes and over time [76, 55]. Only recent studies have started to explore the microbial and viral of these systems by using assembly-based metagenomics and also single-cell genomics [41, 51, 24, 23, 20, 21].

1.3 Lake Tyrrell, Australia, as a model ecosystem

In this work, the chosen study site was Lake Tyrrell, in the Victoria region of Australia. This lake, is located in the center of the Murray Basin Plains (Figure 1.8), in a region characterized by a semi-arid climate with rainfall being 300 mm/year, mainly during the winter season, and where the evaporation rates are 2000 mm/year [34]. This system 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 [**?**]. The lake shows seasonal salinity variations, where during the winter the salt content is close to >250 g/L and in summer, due to water evaporation, the residual brines have concentrations >330 g/L [34].

Lake Tyrrell has been described and studied in detail in terms of its hydrological and geochemical features [**?**, 34, 27], presenting it as a great candidate

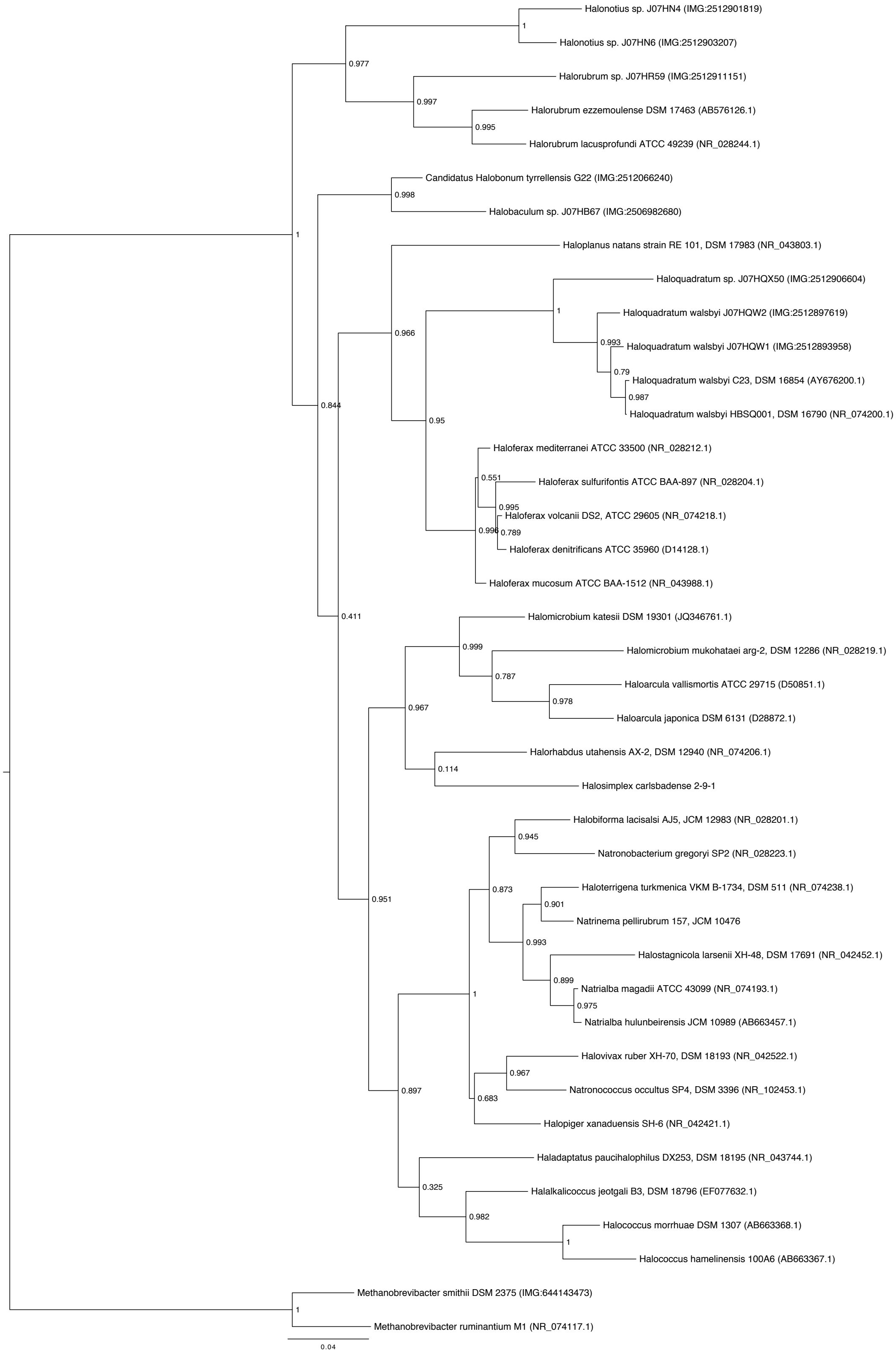
Halonotius sp. J07HN4 (IMG:2512901819)
1
Halonotius sp. J07HN6 (IMG:2512903207)
0.977
Halorubrum sp. J07HR59 (IMG:2512911151)
0.997
Halorubrum ezzemoulense DSM 17463 (AB576126.1)
0.995
Halorubrum lacusprofundi ATCC 49239 (NR_028244.1)
Candidatus Halobonum tyrrellensis G22 (IMG:2512066240)
0.998
Halobaculum sp. J07HB67 (IMG:2506982680)
1
Haloplanus natans strain RE 101, DSM 17983 (NR_043803.1)
Haloquadratum sp. J07HQX50 (IMG:2512906604)
1
Haloquadratum walsbyi J07HQW2 (IMG:2512897619)
0.966
0.993
Haloquadratum walsbyi J07HQW1 (IMG:2512893958)
0.79
Haloquadratum walsbyi C23, DSM 16854 (AY676200.1)
0.987
0.844
Haloquadratum walsbyi HBSQ001, DSM 16790 (NR_074200.1)
0.95
Haloferax mediterranei ATCC 33500 (NR_028212.1)
0.551
Haloferax sulfurifontis ATCC BAA-897 (NR_028204.1)
0.995
Haloferax volcanii DS2, ATCC 29605 (NR_074218.1)
0.996
0.789
Haloferax denitrificans ATCC 35960 (D14128.1)
Haloferax mucosum ATCC BAA-1512 (NR_043988.1)
Halomicrobium katesii DSM 19301 (JQ346761.1)
0.411
0.999
Halomicrobium mukohataei arg-2, DSM 12286 (NR_028219.1)
0.787
Haloarcula vallismortis ATCC 29715 (D50851.1)
0.978
0.967
Haloarcula japonica DSM 6131 (D28872.1)
Halorhabdus utahensis AX-2, DSM 12940 (NR_074206.1)
0.114
Halosimplex carlsbadense 2-9-1
Halobiforma lacisalsi AJ5, JCM 12983 (NR_028201.1)
0.945
Natronobacterium gregoryi SP2 (NR_028223.1)
Haloterrigena turkmenica VKM B-1734, DSM 511 (NR_074238.1)
0.873
0.901
0.951
Natrinema pellirubrum 157, JCM 10476
0.993
Halostagnicola larsenii XH-48, DSM 17691 (NR_042452.1)
0.899
Natrialba magadii ATCC 43099 (NR_074193.1)
1
0.975
Natrialba hulunbeirensis JCM 10989 (AB663457.1)
Halovivax ruber XH-70, DSM 18193 (NR_042522.1)
0.967
0.897
Natronococcus occultus SP4, DSM 3396 (NR_102453.1)
0.683
Halopiger xanaduensis SH-6 (NR_042421.1)
Haladaptatus paucihalophilus DX253, DSM 18195 (NR_043744.1)
0.325
Halalkalicoccus jeotgali B3, DSM 18796 (EF077632.1)
0.982
Halococcus morrhuae DSM 1307 (AB663368.1)
1
Halococcus hamelinensis 100A6 (AB663367.1)
Methanobrevibacter smithii DSM 2375 (IMG:644143473)
1
Methanobrevibacter ruminantium M1 (NR_074117.1)
0.04

Figure 1.4: Phylogenetic tree of G22, using 16S

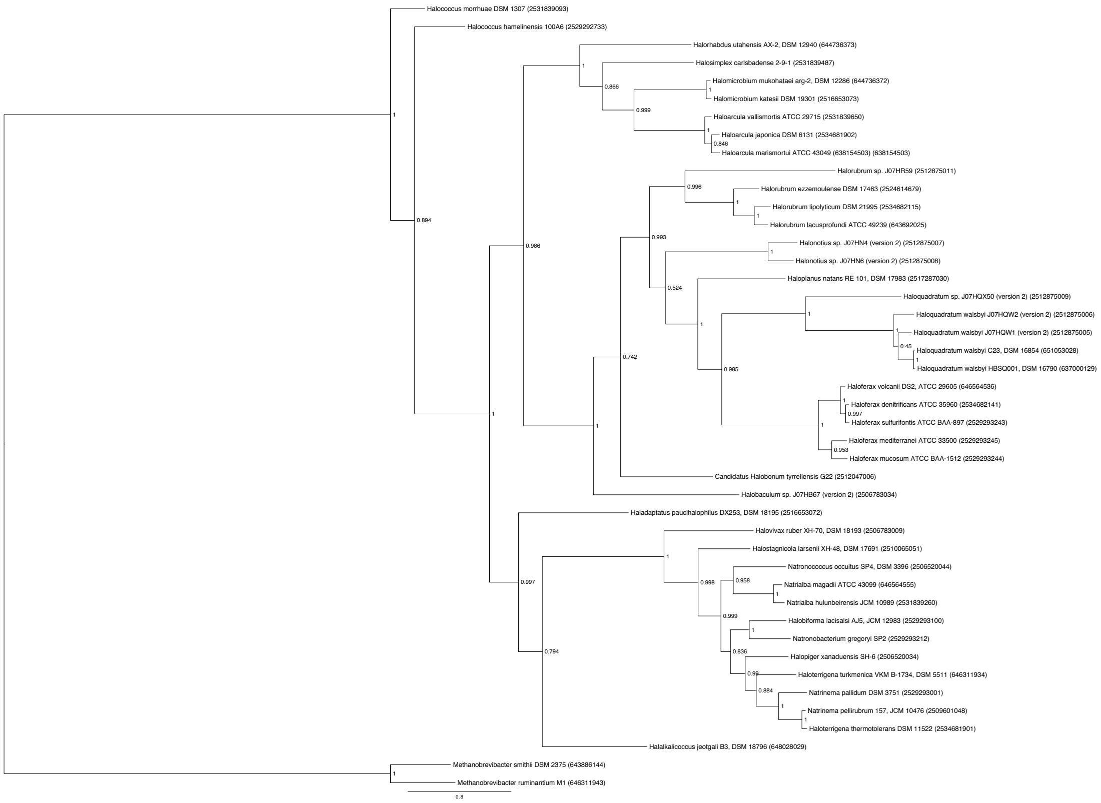
Halococcus morrhuae DSM 1307 (2531839093)
Halococcus hamelinensis 100A6 (2529292733)
Halorhabdus utahensis AX-2, DSM 12940 (644736373)
1
Halosimplex carlsbadense 2-9-1 (2531839487)
0.866
Halomicrobium mukohataei arg-2, DSM 12286 (644736372)
1
Halomicrobium katesii DSM 19301 (2516653073)
1
0.999
Haloarcula vallismortis ATCC 29715 (2531839650)
1
Haloarcula japonica DSM 6131 (2534681902)
0.846
Haloarcula marismortui ATCC 43049 (638154503) (638154503)
Halorubrum sp. J07HR59 (2512875011)
0.996
Halorubrum ezzemoulense DSM 17463 (2524614679)
1
Halorubrum lipolyticum DSM 21995 (2534682115)
1
Halorubrum lacusprofundi ATCC 49239 (643692025)
0.894
0.986
0.993
Halonotius sp. J07HN4 (version 2) (2512875007)
1
Halonotius sp. J07HN6 (version 2) (2512875008)
Haloplanus natans RE 101, DSM 17983 (2517287030)
0.524
Haloquadratum sp. J07HQX50 (version 2) (2512875009)
1
Haloquadratum walsbyi J07HQW2 (version 2) (2512875006)
1
Haloquadratum walsbyi J07HQW1 (version 2) (2512875005)
0.45
Haloquadratum walsbyi C23, DSM 16854 (651053028)
1
Haloquadratum walsbyi HBSQ001, DSM 16790 (637000129)
0.742
0.985
Haloferax volcanii DS2, ATCC 29605 (646564536)
1
Haloferax denitrificans ATCC 35960 (2534682141)
0.997
Haloferax sulfurifontis ATCC BAA-897 (2529293243)
1
1
Haloferax mediterranei ATCC 33500 (2529293245)
0.953
Haloferax mucosum ATCC BAA-1512 (2529293244)
Candidatus Halobonum tyrrellensis G22 (2512047006)
Halobaculum sp. J07HB67 (version 2) (2506783034)
Haladaptatus paucihalophilus DX253, DSM 18195 (2516653072)
Halovivax ruber XH-70, DSM 18193 (2506783009)
1
Halostagnicola larsenii XH-48, DSM 17691 (2510065051)
Natronococcus occultus SP4, DSM 3396 (2506520044)
0.997
0.998
0.958
Natrialba magadii ATCC 43099 (646564555)
1
Natrialba hulunbeirensis JCM 10989 (2531839260)
0.999
Halobiforma lacisalsi AJ5, JCM 12983 (2529293100)
1
Natronobacterium gregoryi SP2 (2529293212)
0.794
0.836
Halopiger xanaduensis SH-6 (2506520034)
0.99
Haloterrigena turkmenica VKM B-1734, DSM 5511 (646311934)
0.884
Natrinema pallidum DSM 3751 (2529293001)
1
Natrinema pellirubrum 157, JCM 10476 (2509601048)
1
Haloterrigena thermotolerans DSM 11522 (2534681901)
Halalkalicoccus jeotgali B3, DSM 18796 (648028029)
Methanobrevibacter smithii DSM 2375 (643886144)
1
Methanobrevibacter ruminantium M1 (646311943)
0.8

Figure 1.5: Phylogenetic tree of G22, using Phylophlan [58]

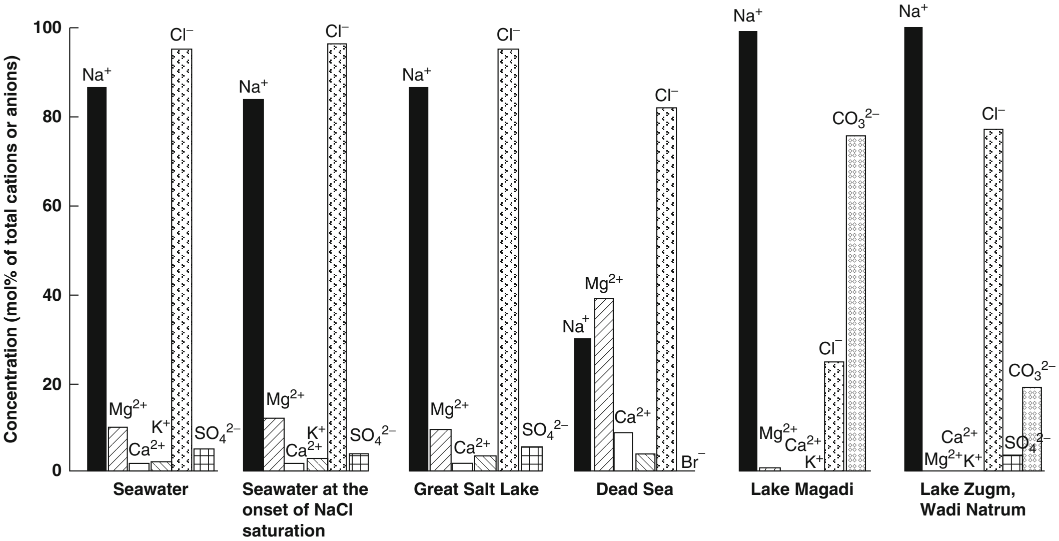

Figure 1.6: Ionic composition of several aquatic systems. Figure from Oren, 2013. [45]

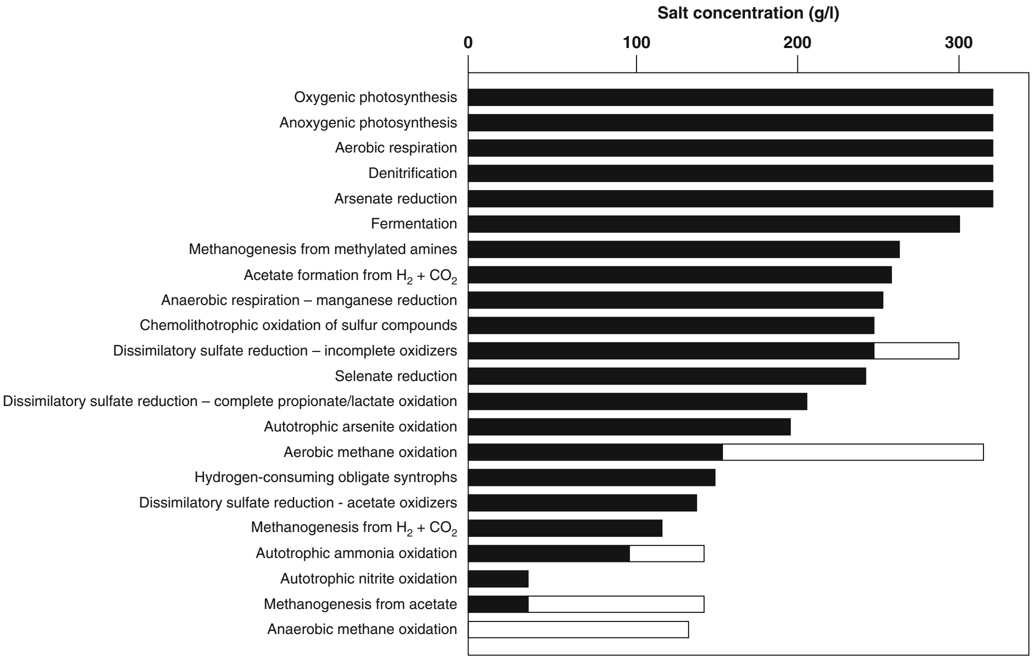

Figure 1.7: Salt concentration limits for some microbial metabolic processes. Black bars indicate information based on laboratory studies, while white bars indicate activities measured in natural microbial communities. Figure from Oren, 2013 [45].

Table 1.1: Halophilic Archaea, modified from [71] to include the recently discovered *Nanohaloarchaea* [41].

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 [71].

Phylum	**Class**	**Genera**
Actinobacteria	*Actinobacteria*	*Actinopolyspora, Amycolatopsis, Georgenia, Corynebacterium, Haloactinobacterium, Haloactinopolyspora, Haloechinothrix, Haloglycomyces, Nesterenjonia, Nocardiopsis, Haloactinospora, Streptomonospora, Isoptericola, Prauserella, Saccharomonospora, Saccharopolyspora*
Bacteoidetes	*Bacteroidia*	*Anaerophaga*
	Flavobacteria	*Gramella, Psychroflexus*
	Sphingobacteria	*Salinibacter, Salisaeta*
Cyanobacteria		*Rubidibacter, Prochlorococcus, Halospirulina*
Firmicutes	*Bacilli*	*Alkalibacillus, Aquisalibacillus, Bacillus, Filobacillus, Gracilibacillus, Halalkalibacillus...*

for microbiological characterisation. Recent projects (this thesis is the outcome of one of such projects) have used a diverse array of microbiological techniques to study its microbial diversity, including Eukaryotes [31], Archaea and Bacteria [51, 41, 68] and Viruses [20, 21]. Future work will combine the metagenomic, proteomic and available geochemical information, providing a more integrative picture of the interactions between microbes, viruses and the environment in this system.

The current dissertation explores the microbial diversity that is present in the Lake Tyrrell ecosystem, based on the data generated through a metagenomic study. In particular, highlights how by assembling metagenomic data, we can obtain a more complete picture of the microbial diversity present in the community, and also how we can exploit this information to look not only a a broad picture of the diversity, but also at a microscale.

Chapter 2 shows how the assembly of metagenomic datasets, combined with additional metadata, such as size fractionation, sequence composition and phylogenetic binning, allows the recovery and identification of novel microbial groups from the sample. In this case, from the Lake Tyrrell metagenome, two near-complete genomes from a novel Class of Archaea, the *Nanohaloarchaea*, were recovered.

Chapter 3 provides the bioinformatic analysis and description of a novel type of microbial rhodopsin that was identified in the genomes of the *Nanohaloarchaea*, named Xenorhodopsins. The results show how this rhodopsin appears to be a new class, based on phylogenetic analysis and also on the presence of unique amino acid signatures, which makes it different from the already previously described groups.

Chapter 4 describes the assembly of several genomes from the Lake Tyrrell metagenome, based on the combination of assembly-based approaches and metadata. The results here provide a framework for future analysis of this ecosystem, providing a set of habitat-specific genomes that can be used for future analysis, including phylogenetic, functional and genetic diversity.

Chapter 5 leverages on the assembled habitat-specific genomes, and shows a bioinformatic approach for the analysis of genetic diversity in a metagenome sample. Using this approach, I identified difference between the genetic heterogeneity in the Lake Tyrrell microbial populations, which results in evidence of positive selection in specific functions, in the members of the community.

1.4 Acknowledgments