Tanzania Journal of Science 50(5): 1147-1167, 2024
ISSN 0856-1761, e-ISSN 2507-7961

Metagenomic Analysis of Bacterial Communities in Waters of Lake Natron, Arusha, Tanzania

Sadikiel E. Kaale[1,2], Robert S. Machang'u[3] and Thomas J. Lyimo[1*]
[1]*Department of Molecular Biology and Biotechnology, University of Dar es Salaam, P.O. Box 35179, Dar es Salaam, Tanzania*
[2]*School of Life Sciences and Biomedical Engineering, Nelson Mandela Institute of Science and Technology, P.O. Box 447, Tengeru-Arusha, Tanzania*
[3]*Department of Microbiology, Saint Francis University College of Health and Allied Sciences, P. O. Box 175, Ifakara-Morogoro, Tanzania*
sadikiels@gmail.com; machangu@sua.ac.tz
[*]*Corresponding author*: tjlyimo@udsm.ac.tz or tlyimo2000@yahoo.com.
Received 17th Sept 2024, Reviewed 30th Oct., Accepted 27th Dec., Published 31st Dec. 2024
https://dx.doi.org/10.4314/tjs.v50i5.18

Abstract

Metagenomics involves genetic material extraction and sequencing directly from environmental samples in order to gain insights and relationships that exist between microbes and their surroundings. Few studies have been done in extreme environments including soda lakes such as Lake Natron which is found in Arusha region, Tanzania. This study recorded high pH and salinity values confirming the lake's extremity. Full-length 16S rRNA reads sequenced through PacBio was used to reveal first metagenomic snapshots of bacterial communities from 10 random points at the lake shoreline waters. The results showed the dominance of Proteobacteria and Firmicutes with relative abundances of 98.46% and 70.46, respectively. Alphaproteobacteria (93.59%), Bacteroidia (23.80%) and Bacilli (23.19%) were the most dominant classes. *Oceanibaculaceae* (52.43%), *Rhizobiaceae* (66.62%) and *Izemoplasmataceae* (12.50%) were the most dominant families. The dominant genera were *Oceanibaculum* (52.44%), *Allorhizobium* (65.59%) and *Izimaplasma* (12.50%), respectively. The diversity indices showed high level of community diversity, a large number of species, the presence of rare species and an even distribution of bacteria across the sampling points. While this study provides the first report on the occurrence of various taxa in Lake Natron, functional metagenomic analysis is recommended for further investigation of the ecological and biotechnological significance of the identified species.

Keywords: Metagenomics; PacBio sequencing; bacteria diversity; Soda Lakes; Lake Natron

Introduction

Metagenomics is the study involving the extraction of DNA or RNA directly from environmental organisms. It is frequently used for studying particular communities of microorganisms in various habitats. To directly access the genetic information of vast populations of species in their entire communities, metagenomics employs an array of next generation sequencing (NGS) technologies and analysis through bioinformatics tools (Thomas et al. 2012). One of the NGS technologies is PacBio system which has been used recently in the sequencing of full-length 16S rRNA reads. The advantage of this technique is refined classification of bacterial communities up to the species level (Hur and Park 2019). These technologies have enhanced the understanding of microorganisms, particularly bacteria, in their natural habitats including extreme environments. A habitat is considered to be an

extreme environment if it has unusual and difficult-to-survive conditions, such as extremely high or low salinity, pH, and temperatures (Ando et al. 2021). Extreme natural environments in Tanzania include the soda lakes along the Gregory Rift Valley which possess alkaline characteristics, difficult for different life forms to survive. Their natural settings maintain a pH of 9, or above, consistently, frequently, or for extended periods of time (Hamisi et al. 2017).

Several earlier studies done in Lake Natron focused mainly on flamingos and other non-microbial diversity in the lake (Mgimwa et al. 2021; Yona et al. 2023) with little attention on bacterial species (Yakimov et al. 2001; Nonga et al. 2017). Very few previous studies done on Tanzania soda lakes evidently shows an abundant and functional bacteria composition. For example, the studies done by Kaale et al. (2022) and Lema et al. (2022) in five Momela soda lakes isolated 15 different genera of phylum Actinomycetota with pharmacological and antimicrobial potential. A metagenomic study done by Maghembe et al. (2021) in Big Momela detailed the species belonging to newly described genus of cyanobacteria, *Limnospira* spp. Additionally, our recent study done in Lake Natron sediments isolated 13 different genera of phylum Actinomycetota (Kaale et al. 2024) with bioprospecting of producing variety of compounds with ecological and biotechnological significance. Despite all these efforts, the bacteria composition of Lake Natron waters remained unknown until this study.

There are many hindrances to bacterial studies in the soda lakes of Tanzania, such as limited genomic information of the microorganisms and their characterization. This is because most of the lakes are found in hard to reach areas that limit sampling (Zorz et al. 2019). Also, the elevated pH and high salinity levels characteristic of soda lakes can interfere with DNA extraction, PCR amplification, and sequencing processes, as salts and alkaline conditions inhibit many enzymes, including DNA polymerases (Sorokin et al. 2014). In our previous study (Kaale et al. 2024) we showed that Lake Natron hosts a variety of bacteria with potential biotechnological importance. The primary objective of this study was to investigate the bacterial communities inhabiting the waters of Lake Natron, with a particular focus on the phylum Actinomycetota. This phylum is of significant interest because its species are known to produce bioactive compounds with great potential in biotechnology and various industrial applications. Henceforth, we explored, for the first time, the diversity of bacteria species inhabiting the waters of Lake Natron by using full-length 16S rRNA reads sequenced by a PacBio technique. The study intended to address the question of whether Lake Natron harbors a diverse array of bacterial species with significant biotechnological potential, specifically focusing of the phylum Actinomycetota. These bacteria could serve as a valuable resource pool for novel metabolites with applications in biotechnology and medicine, thus opening up new opportunities in drug discovery and other biotechnological applications.

Materials and Methods

Sampling site

Samplings were done in Lake Natron, a soda lake found in the eastern branch of the East African Gregory Rift Valley in Tanzania. It is among the soda lakes located in Arusha region, northern Ngorongoro and Loliondo districts, Tanzania. The lake is under supervision of governmental organizations namely Tanzania National Parks (TANAPA), Tanzania Wildlife Research Institute (TAWIRI) and Tanzania Wildlife Management Authority (TAWA). The lake has high levels of evaporation which have left behind natron sodium carbonate decahydrate (natron) and sodium sesquicarbonate dihydrate (trona); the pH of over 10 (greater than 12 especially in dry conditions (Mgimwa et al. 2021)) which can burn the eyes and skin of animals not adapted to the conditions. The alkalinity of the water is mainly caused by leaching through the nearby Mount *Ol-Doinyo Lengai* volcanic materials (Yona et al. 2023). It contains extreme conditions as soda ashes which can be seen at the shoreline of the lake.

Typically, the lake unique alkaline ecosystem supports only flamingos and wetland birds; in fact, it is the crucial breeding site in the East Africa region for Africa's lesser flamingoes (Nonga et al. 2017). These characteristics stipulate Lake Natron as one of the most extreme soda lakes.

Water sampling and physicochemical parameter measurements

Samplings were done in October 2023 whereby water samples were aseptically taken from 10 random points in the shoreline of west coast of Lake Natron as shown in Figure 1. Although this study was not quantitative, the distance between each sampling point ranged from 25 to 30 meters. The 10 sampling points were labeled as NWA, NWB, NWC, NWD, NWE, NWF, NWG, NWH, NWI and NWJ (where N = Natron; W = Water and A-J = different sampling points). Triplicate samples of water (200 mL) were collected at the surface in sterile 250 mL bottles. The bottles were tightly capped and transported in a cool box (packed with ice cubes) to the Department of Molecular Biology and Biotechnology Laboratory, University of Dar es Salaam for analysis. Immediately upon arrival in the laboratory (approximately 1 day after sampling), the samples were processed for the DNA extraction. Multi-Parameter meter (Hach 40HQD) was used on-site (*in situ*) to measure the physicochemical parameters in triplicates. The parameters measured were temperature, pH, salinity, electrical conductivity (EC) and total dissolved salts (TDS).

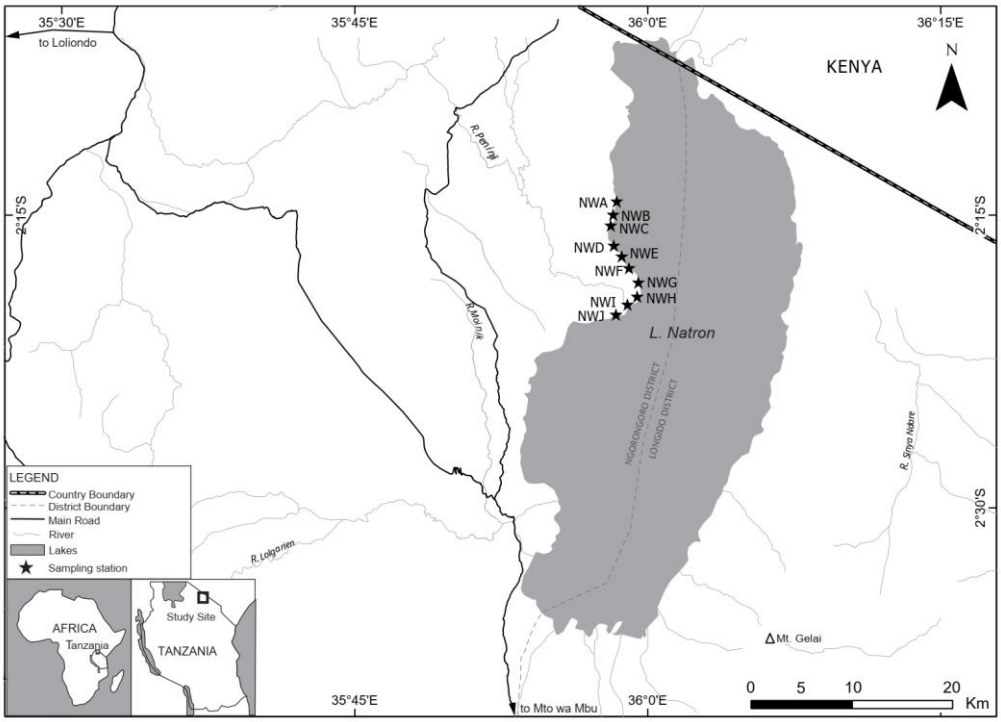

Figure 1: Map of Lake Natron (shaded) in Arusha, Tanzania, showing sampling points (area with asterisks) at the shoreline of the lake

Extraction of genomic DNA

DNA was extracted by filtration of about 100 mL of water samples, and quantity/quality check of eluted DNAs was done following the protocol depicted by Msaki et al. (2023). We also used ZymoBIOMICS™ DNA Miniprep kit (Cat No. D4304, Zymoresearch Corp. USA) following the manufacture's protocol for DNA purification. The isolated genomic DNAs (gDNAs) were kept in storage at -80 °C freezer (Forma 900 series, Thermo Scientific).

PCR Amplification of bacterial 16S rRNA gene

Prior to PCR amplification, the gDNAs from different sampling points were pooled together in equal amounts for the purpose of minimizing the samples (Jeilu et al. 2022) as follows; from sample NWA, NWB and NWC were pooled and named NW1; sample from NWD, NWE and NWF were pooled and named NW2; sample from NWG, NWH, NWI and NWJ were pooled and named NW3. The now pooled samples of DNAs (NW1, NW2 and NW3) were sent to INQABA®, genomics facility, Pretoria, South Africa, for PacBio sequencing. The three samples of gDNAs (NW1, NW2 and NW3) were PCR amplified using the bacterial universal primer pair 27F (5´-AGAGTTTGATCMTGGCTCAG-3´) as a forward primer and 1492R (5´-GGTTACCTTGTTACGACTT-3´) as a reverse primer, targeting the full-length of 16S rRNA gene with an amplicon size of approximately 1500 kb (Palkova et al. 2021). These primers were marked with PacBio M13 adapter sequences on both 5' and 3' ends to enable barcoding of each amplicon so as to multiplex the resultant amplicons using limited cycle PCR.

16S rRNA gene sequencing and bioinformatic analysis

The resulting barcoded amplicons were quantified and pooled equimolar, followed by ampure PB bead-based purification step. The PacBio SMRTbell library was made using the pooled amplicons in accordance with the Manufacturer's instructions (https://www.pacb.com/wp-content/uploads/Procedure-checklist-Preparing-SMRTbell-libraries-using-PacBio-barcoded-M13-primers-for-multiplex-SMRT-sequencing.pdf). Sequencing, primer annealing and polymerase binding were done following SMRT link software protocol to prepare the library for sequencing on PacBio Sequel IIe system. The SMRTlink (v11.0) was used to process the raw sub-reads obtained after full-length 16S rRNA gene amplicon sequencing. Circular consensus sequences (CCS) technique was utilized to get accurate readings (>QV40), which were processed through Vsearch version 2.23.0 (https://github.com/torognes/vsearch). Taxonomic information was ascertained based on the Quantitative Insights into Microbial Ecology (QIMME2) platform, a NGS microbiome bioinformatics tool for microorganism diversity and statistics (Caporaso et al. 2010).

The sequences of pooled samples were then submitted to the NCBI BioProject database, sequence and three accession numbers (NW1 (SAMN38511730), NW2 (SAMN38511731) and NW3 (SAMN38511732)) were retrieved. Thereafter, the statistical analyses of the diversity indices were computed using R-studio software. The Shannon-Wiener diversity, Simpson, Chao1, Goods coverage and dominance were calculated to demonstrate the alpha bacterial community diversity, richness and evenness (Kim et al. 2017).

Results and Discussion

Physicochemical parameters of Lake Natron's water

The results for physicochemical parameters viz. temperature, pH, salinity, EC and TDS were as shown in Table 1. The values recorded had minor variations in all sampling points along the lake shore. The temperature ranged from 35.30 ± 0.39 ^{o}C to 37.19 ± 0.71 ^{o}C, which is within the range previously reported by Nonga et al. (2017) in Lake Natron. This temperature range is favorable for the proliferation of mesophilic bacteria, it being a vital requirement for their metabolic activities (Yona et al. 2023). The pH and salinity ranged from 9.51 ± 0.23 to 9.87 ± 0.31 and 15.02 ± 0.29 psu to 16.01 ± 0.36 psu, respectively. The values of pH and salinity were within similar range with those reported by Yona et al. (2023), they are attributed to the chemical composition of the lake's water which contains different salts (Clarisse et al. 2019). The two parameters are crucial for the distribution and diversity of bacteria species in soda lakes and the main reason why the waters of the lake are not suitable for use by humans and most of animals (Chavan et al. 2013). The EC and TDS recorded moderate values, ranging from 4.92 ± 0.02 mS/cm to 6.85 ± 0.32 mS/cm and 2.95 ± 0.24 g/L to 3.64 ± 0.22 g/L,

respectively. In Lake Natron water, EC and TDS levels are attributed by the presence of carbonate, bicarbonate, chloride, sulfate, phosphate and nitrate salts (Philip and Mosha 2012).

Table 1: Lake Natron waters' physico-chemical parameters measured from the coast of the lake (mean ± standard deviation).

Sampling Point	Parameters				
	Temperature (°C)	**pH**	**Salinity (psu)**	**EC (mS/cm)**	**TDS (g/L)**
NWA	36.40 ± 0.63	9.58 ± 0.10	15.62 ± 0.15	5.45 ± 0.73	3.41 ± 0.14
NWB	36.02 ± 0.34	9.63 ± 0.41	15.47 ± 0.22	5.09 ± 0.32	3.25 ± 0.12
NWC	36.50 ± 0.45	9.71 ± 0.21	15.42 ± 0.23	5.53 ± 0.16	3.51 ± 0.04
NWD	37.19 ± 0.71	9.51 ± 0.23	15.63 ± 0.12	5.32 ± 0.43	3.32 ± 0.23
NEW	36.20 ± 0.15	9.84 ± 0.21	15.47 ± 1.15	5.13 ± 1.51	3.17 ± 0.84
NWF	36.30 ± 0.19	9.61 ± 0.38	15.43 ± 1.01	5.21 ± 0.22	3.12 ± 0.25
NWG	37.15 ± 0.24	9.56 ± 0.15	16.01 ± 0.36	6.85 ± 0.32	3.64 ± 0.22
NWH	36.35 ± 0.38	9.87 ± 0.31	15.37 ± 1.62	4.92 ± 0.02	3.06 ± 0.22
NWI	35.30 ± 0.39	9.65 ± 0.12	15.86 ± 0.32	5.63 ± 0.03	3.55 ± 0.26
NWJ	36.20 ± 0.29	9.61 ± 0.05	15.02 ± 0.29	5.14 ± 0.04	2.95 ± 0.24

Metagenomic analysis of full length 16S gene amplicons

This study revealed the first metagenomic snapshot of bacteria species inhabiting Lake Natron from water samples. All three samples namely NW1 (SAMN38511730), NW2 (SAMN38511731) and NW3 (SAMN38511732) produced about 13979.0, 3906.0 and 1328.0, respectively of quality-filtered non chimeric sequences used in community diversity analysis. Furthermore, using the PacBio system to sequence full-length 16S rRNA reads with NGS technology at a 97% similarity cut off via BLASTn analysis, all sequences belonged to Domain Bacteria by 100%. The taxonomical classification at phylum, class, family, genus and species level is detailed in sections below.

Bacterial diversity at the phylum level

The results revealed a total of 29 different Phyla in the waters of Lake Natron of which 14 phyla in NW1, 20 phyla in NW2 and 25 phyla in NW3. As shown in Figure 2, Phylum Proteobacteria was dominant in NW1 and NW2 with relative abundances of 98.46% and 70.46%, respectively. In NW3, Firmicutes was the dominant phylum with a relative abundance of 35.32% followed by Bacteroidota (25.68%) and Proteobacteria (12.80%). The dominance of Proteobacteria was also observed by Vavourakis et al. (2016) who explored the uncultured bacteria of four hypersaline soda lake brines by metagenomic analysis. The main reason for the dominance of Proteobacteria in direct water samples microbiome is attributed to their adaptability in extreme habitats. Apart from the fact that they are versatile, they also can use a variety of energy and carbon sources, including organic and inorganic compounds, are capable of anaerobic respiration, and perform a variety of highly metabolic activities. The dominance of Firmicutes has also been found in five Siberian soda lakes by Vavourakis et al. (2018). Phylum Bacteroidota was the third dominant species in all sampling points, results similar to Rojas et al. (2018) shows the dominance of these particular species in very alkaliphic soda lakes. In spite of the extremely high salinity and alkaline pH of soda lakes, Firmicutes and Bacteroidota contribute to their ecosystem stability through biogeochemical cycling processes (Sorokin et al. 2014). Various metagenomic soda lake studies have reported the concurrently dominance of Proteobacteria, Firmicutes and Bacteroidota phyla over others as they compete for sources of energy and produce antimicrobial compounds inhibiting the growth of other microorganisms (Omeroglu et al. 2021; Wang et al. 2022).

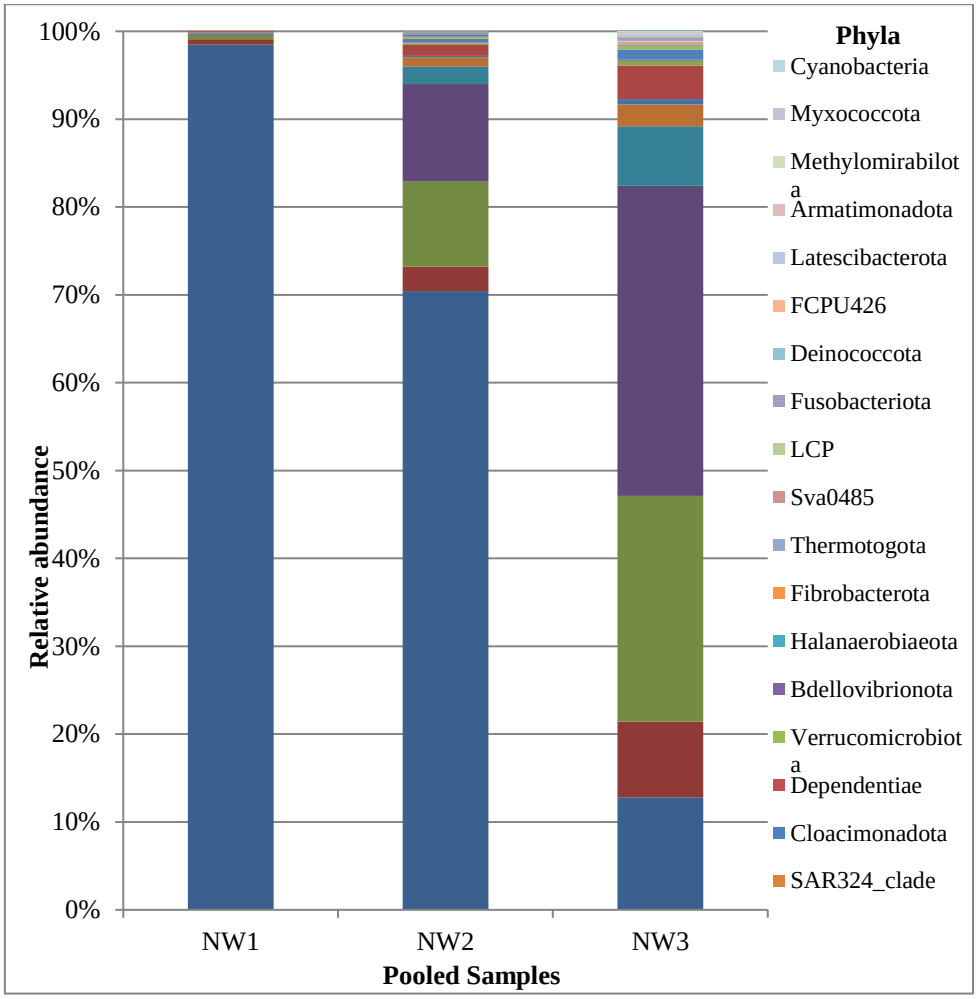

Figure 2: Taxonomic classification at Phylum level of different bacteria from the waters of Lake Natron in Arusha, Tanzania.

Other important phyla found in Lake Natron waters consisted of Actinomycetota, Desulfobacterota, Chloroflexi and Cloacimonadota. While all species of bacteria recovered from the lake's waters were examined in this study, the Phylum Actinomycetota in particular was of great interest due to their antibacterial and biotechnological potential. Actinomycetota, previously known as Actinobacteria (Oren and Garrity 2021) have been studied in various soda lakes including five Momela soda lakes (Kaale et al. 2022) and Lake Natron waters (Kaale et al. 2024) in Tanzania. The Actinomycetota species isolated in these lakes showed prodigious biotechnological and pharmacological potential particularly antibacterial and antifungal activities (Lema et al. 2022). Hence the present study further proves that these particular species are ubiquitous in different soda lakes habitats of Tanzania.

The presence of Desulfobacterota, Desulfobacterota, Chloroflexi and Cloacimonadota phyla in high abundance in this study were similar to other studies that have shown these phyla to be the dominant communities in soda lake habitats, especially from direct water samples. For example, these bacteria have been reported by Chen et al. (2021) from Siberian soda lakes, Luo et al. (2017) from Oloidien Soda Lake in Kenya and Cabello-Yeves et al. (2023) from Lake El Tobar in Spain. They are integral part of sulfur

cycle, partake in nitrogen removal, biofilm aggregation, carbon cycling and organic matter degradation (Bovio-Winkler et al. 2023).

The rest of the other phyla were represented by a relative abundance of less than 1%, although they have been identified in such small quantities, their presence in Lake Natron water groups narrates them as halophiles and potentially (Appendix 1). On the other hand, this first metagenomic study to be conducted in Lake Natron, revealed that few bacteria match with previously culture based studies. Only Phylum Actinomycetota species isolated by Kaale et al. (2024) and Phylum Cyanobacteria species isolated by Nonga et al. (2017) in Lake Natron have also been found in this study, indicating that they are abundantly distributed and culturable. Moreover, Yakimov et al. (2001) isolated *Alcalilimnicola halodurans* belonging to phylum Pseudomonadota in Lake Natron but was not retrieved in this study. This may connote that Pseudomonadota are found in such low abundances (but easily culturable) that metagenomic sequencing could not detect them (Hilton et al. 2016).

Bacterial diversity at the class level

This study has shown a total of 58 different classes of which 22 from NW1, 39 from NW2 and 47 from NW3. The most dominant class was Alphaproteobacteria with a relative abundance of 93.59% in NW1 and 67.72% in NW2 while Class Bacteroidia (23.80%) and Bacilli (23.19%), were the most dominant in NW3 as shown in Figure 3. The dominance of Alphaproteobacteria, Bacteroidia and Bacilli from the soda lakes water samples has been reported previously by Sorokin et al. (2014) and these classes have been deployed in different biogeochemical cycling of nitrogen, carbon and sulfur. The results also reveal the four classes of Actinomycetota species namely Acidimicrobiia, Coriobacteriia, Actinobacteria and Thermoleophilia to be abundant in Lake Natron waters (Figure 3, Appendix 2), this further insinuates the ubiquitous diversity of the species due to their adaptive mechanisms (Oren and Garrity 2021).

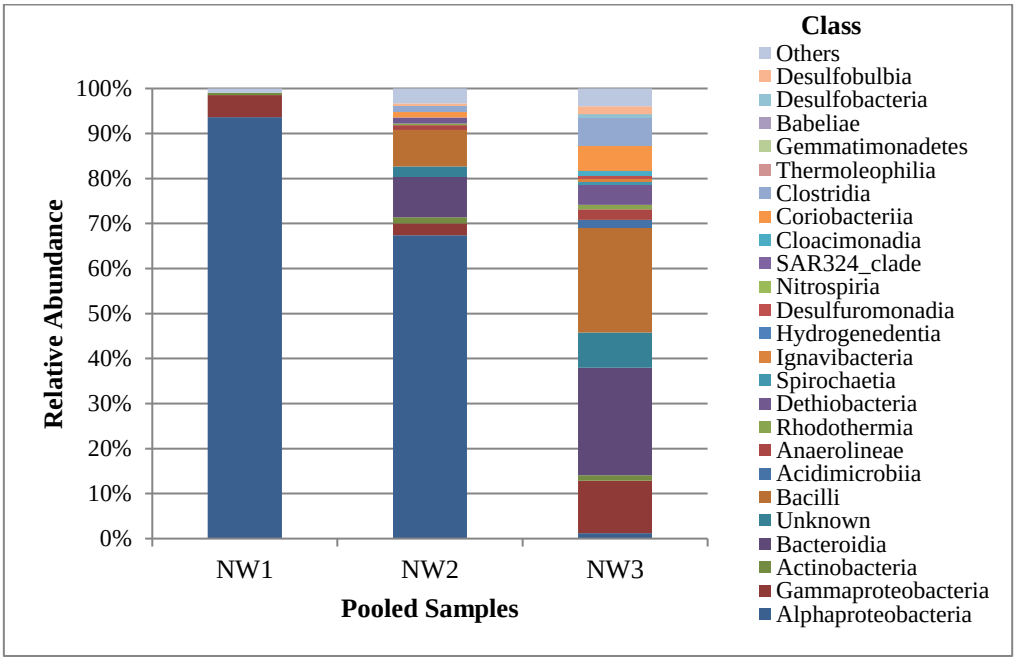

Figure 3: Taxonomic classification at class level of different bacteria from the waters of Lake Natron, Arusha, Tanzania. The Figure shows the classes with the relative abundances of 0.5%≥.

Other classes namely Gammaproteobacteria, Gammaproteobacteria, Dethiobacteria, Clostridia, Anaerolineae, Desulfobulbia, Cloacimonadia, Rhodothermia, Ignavibacteria, Desulfobacteria, Desulfuromonadia, Spirochaetia and Desulfovibrionia were also highly represented in all samples with the relative abundance of more than 0.5% as shown in Figure 3 while others were presented in relative abundances of lower than 0.5% (Appendix 2). Basically, the role of these bacteria are diverse in soda lakes, however, in general, they are involved in nutrient cycling, sulfur metabolism and organic compound degradation (Sorokin et al. 2014).

Bacterial Diversity at the family level

The revealed number of bacterial families inhabiting Lake Natron water were 35 in NW1, 99 in NW2 and 109 in NW3. In total, 139 different bacterial families were found in direct water samples of Lake Natron. The most dominant families were *Oceanibaculaceae*, *Rhizobiaceae* and *Izemoplasmataceae* by relative abundances of 52.43%, 66.62% and 12.50% in NW1, NW2 and NW3, respectively (Figure 4). Based on literature pertaining to metagenomics studies, this is the first study showing the dominance of family *Oceanibaculaceae*, *Rhizobiaceae* and *Izemoplasmataceae* in a soda lake environment. Basically, family *Oceanibaculaceae* has been proposed recently in soda lakes (Koziaeva et al. 2023) and are able to use carbon dioxide as a carbon source and sulfur as an energy source due to their capacity for sulfur-dependent lithoautotrophy. Being a recently proposed family, there is little known of their distribution and characteristics. Similarly, the data on the ecological roles of *Rhizobiaceae* is limited as this family has not been found in soda lakes but mainly in plants. It's possible to infer that just like in other habitats, *Rhizobiaceae* is involved in sulfur-dependent lithoautotrophy and nitrogen fixation (Gopalakrishnan et al. 2015).

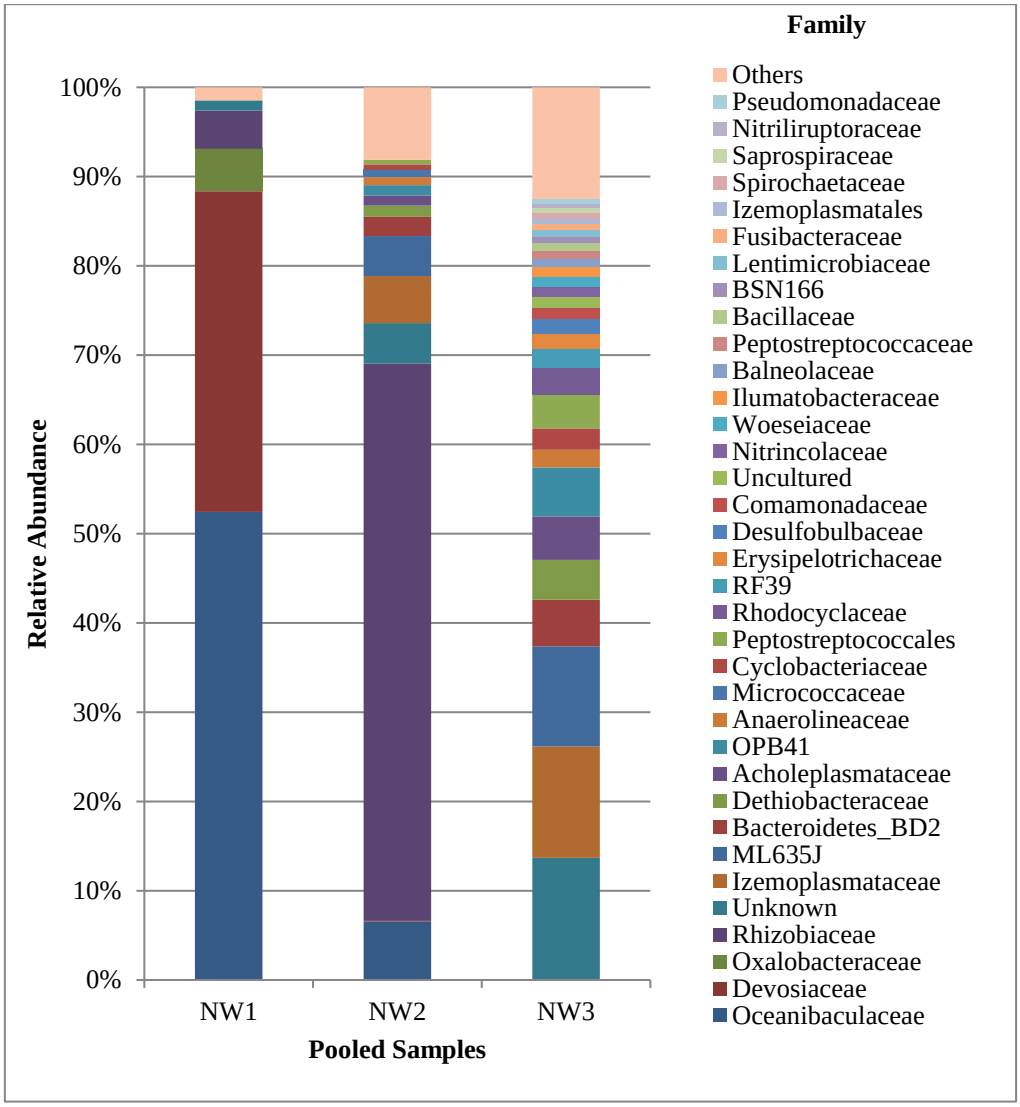

Figure 4: Classification at Family level of bacteria from the waters of Lake Natron in Arusha, Tanzania. The figure shows only the families with the relative abundances of 0.5%≥.

In family level the Phylum Actinomycetota species were chiefly represented by *Nocardioidaceae*. Species found in this family have previously been isolated from Momela lakes and are characterized by producing antimicrobial products and enzymes (Lema et al. 2022). Other notable Families were *Devosiaceae*, *Oxalobacteraceae*, *Rhizobiaceae*, *Izemoplasmataceae*, *Bacteroidetes*_BD2, *Dethiobacteraceae*, *Acholeplasmataceae*, *Peptostreptococcales*, *Rhodocyclaceae*, *Anaerolineaceae*, *Erysipelotrichaceae*, *Desulfobulbaceae*, *Comamonadaceae*, *Nitrincolaceae*, *Woeseiaceae* and those families belonging to "Candidate Phyla" namely ML635J, OPB41 and RF39. The "Candidate Phyla" refers to a group of bacterial lineages that are not yet fully characterized or cultivated in the laboratory. Their presence represents a frontier in microbiology, offering insights into the vast unseen microbial diversity that exists in various environments around the world including soda lakes. Collectively, members

of these families are ecologically involved in different nutrient cycling in soda lakes and produce a variety of products of biotechnological and industrial importance. The rest of the families had relative abundances of less than 0.5% (Appendix 3).

Bacterial diversity at the genus level

Lake Natron waters were shown to contain a number of diverse bacteria at the genus level. The results showed 28 different genera in NW1, 104 in NW2 and 110 in NW1. In total, 162 different genera of bacteria were retrieved in the Lake Natron waters. The dominant genera were *Oceanibaculum* (52.44%) in NW1, *Allorhizobium* (65.59%) in NW2 and *Izimaplasma* (12.50%) in NW3 (Figure 5). Genus *Oceanibaculum* contains the species which have been solely recovered from oceanic environments capable of reducing nitrate to nitrites (Lai et al. 2009). This is the first study revealing the presence of *Oceanibaculum* bacteria in a soda lake environment. Members of genus *Allorhizobium* are exclusively found in soil and responsible for nitrogen fixation in legumes and crown gall (Mousavi et al. 2015). Moreover, the presence of genus *Izimaplasma* have been reported in diverse environments, they are typical commensals or parasites in their eukaryotic hosts (Skennerton et al. 2016).

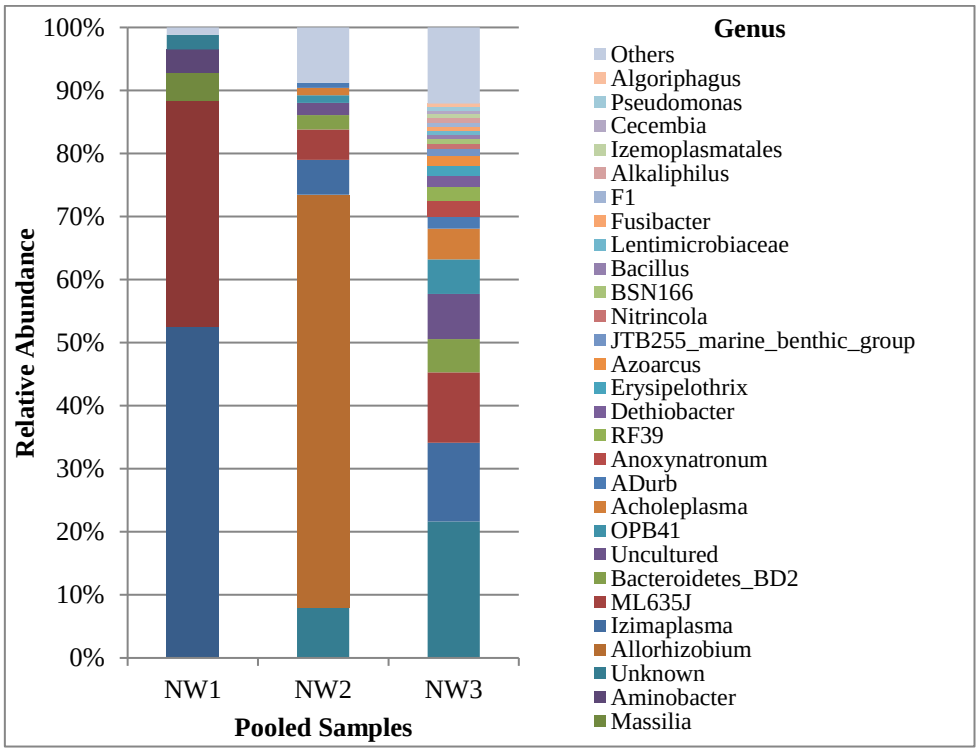

Figure 5: Classification at Genus level of bacteria from the waters of Lake Natron, Arusha, Tanzania. The Figure shows the genera with the relative abundances of 0.5%≥.

Other relatively abundant genera included *Devosia* and ML635J that have been reported to be potential in agriculture as competent biofertilizers (Chhetri et al. 2022). The "candidate phyla" genus ML635J was also relatively abundant and has already been reported to be dominant in Siberian soda lakes Vavourakis et al. (2019). Other notably genera found were *Massilia*, *Aminobacter*, *Bacteroidetes*_BD2, *Acholeplasma*, OPB41, *Anoxynatronum*, RF39, ADurb, *Dethiobacter*, *Erysipelothrix* and *Azoarcus*. Generally, the members of these genera are potential in the production of different biotechnological products and, ecologically, they are involved

in various nutrient cycling processes (Sorokin et al. 2014).

The representation of other genera was low as they had relative abundances of less than 0.5% (Appendix 4), but interestingly, Actinomycetota species were massively found at genus level as it included *Cellulomonas*, *Alkaliphilus*, *Agrococcus*, *Streptomyces*, *Nocardioides*, *Alkalitalea*, *Micromonospora*, *Agrococcus*, *Kocuria* and *Nitriliruptora*. Previously these species have been also found in Momela soda lakes (Kaale et al. 2022; Lema et al. 2022) hence they are both ubiquitous in Tanzania soda lakes and capable of surviving the alkaline environments with high pH and salinity.

Diversity at species level and presence of vast unknown and uncultured species

The results show that there were 18 species in NW1 (SAMN38511730), 75 species in NW2 (SAMN38511731), 91 species in NW3 (SAMN38511732) and in total 133 different species in all samples with notable dominance of unknown and uncultured species. Most of metagenomic studies often characterize the microbiome at least up to the genus level due to the high number of microbial unknown species (poorly characterized) at the species level within the environmental samples. Thus, further investigation of the unknown groups through NGS technologies is of importance as it gives more insight of soda lake microbial communities (Msaki et al. 2023). It was observed that as you go down the taxonomic ranks, the percentage of unknown bacteria significantly increased a trend evident across all samples (Figure 3−6); the trend was also observed in resulted uncultured bacteria. The unknowns were the most dominant species recovered from Lake Natron waters in NW1, NW2 and NW3 samples, with the relative abundances of 98.80%, 79.13% and 42.55%, respectively. This could be attributed by low representation of their sequences in reference databases (Msaki et al. 2023). Henceforth, the development of more bioinformatics tools to decode and characterize these unknown groups is a necessity (Vanni et al. 2022).

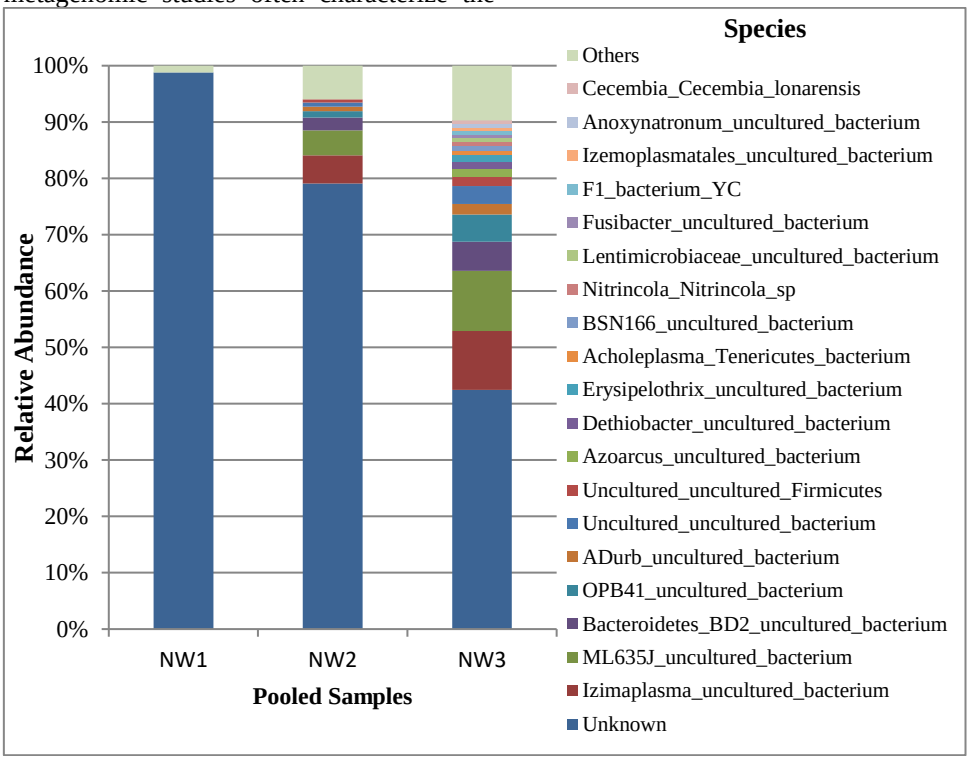

Figure 6: Speciation of different bacterial species of the waters of Lake Natron, Arusha, Tanzania. The figure shows the species with the relative abundances of 0.5%≥.

Other species found in Lake Natron waters were in very low abundances (Appendix 5) including *Cecembia lonarensis*, *Izimaplasma*, L635J, *Erysipelothrix*, *Dethiobacter*, *Bacteroidetes*_BD2, OPB41, ADurb, *Firmicutes*, *Nitrincola* sp. and *Acholeplasma tenericutes* (Figure 6). The presence of *Cecembia lonarensis* in Lake Natron waters makes this the second study to report it in soda lake habitat after Anil et al. (2012) reported it from haloalkaline water samples of Lonar Lake of Buldhana district, India. The ecological and biotechnological importance of this haloalkalitolerant bacterium is not well understood, thus further studies are needed in order to assess their role in soda lake water habitats.

Bacterial community diversity indices

In microbial ecology studies, the community diversity indices (CDI) are the measures of diversity, richness and evenness of biological distribution in a given community. So far there is no universal agreement on specific CDI to use when conducting microbiome studies. Thus, in this study we used Chao1, SDI, Simpson, Good's Coverage and dominance indices which are commonly used (Moris et al. 2014). Chao1 is the species richness estimator which expresses the total number of species present in a community by using the frequency of occurrence of rare operational taxonomic units (OTUs). As shown in Table 2, the Chao1 estimate for the NW2 is significantly higher than those in NW1 and NW3 samples, suggesting that the particular community may harbor a greater number of species or a higher diversity of rare species. The observation that Chao1 values are higher than OUT values in all samples indicates species richness is higher than the number of species actually observed as Chao1 calculates not only the observed OUTs but also rare species not commonly observed in the sample (Kim et al. 2017).

Table 2: Community Diversity indices

Pooled Sample Name	OTUs	Chao1	Shannon Index	Simpson Index	Good's coverage	Dominance
NW1	636	3093.02	2.52	0.64	0.99	0.39
NW2	918	4906.68	4.29	0.61	0.98	0.36
NW3	828	3097.24	8.98	0.99	0.95	0.01

The high value of Simpson Diversity Index and SDI in NW3 indicates a high number of species and high diversity of species in this community, respectively. The observation that all samples had low dominance index values (Table 2) stipulates individuals are evenly distributed in their communities, no single species dominates the whole community, higher biodiversity; the higher the value mean lower biodiversity. The Good's Coverage Index was high in all samples ranging from 95% (NW3) to 99% (NW1) indicating both the dominant phylotypes were well represented and more complete sampling of the microbial community was done (Bardenhorst et al. 2022). In general, the microbial CDI used in this study provided valuable insights into the diversity, richness, and evenness of the microbial communities found in Lake Natron waters. They indicated not only greater number of species or a higher diversity of rare species but also well representation of dominant phylotypes while insinuated more complete sampling of the microbial community was done.

Conclusions

Lake Natron exhibits a rich bacterial diversity in both its water with dominant populations in the phyla Proteobacteria, Firmicutes, Bacteroidota and Actinomycetota. These phyla are responsible for essential biogeochemical processes, significantly influencing nutrient cycling and organic matter degradation, thereby outcompeting other microbial groups present in the lake's extreme environment. The physicochemical parameters of the lake's water i.e. high pH and salinity insinuated the extremity of the habitat but supporting high bacterial species diversity. To the best of our knowledge, this is the first

full length 16S rRNA gene sequencing metagenomic research on soda lakes of Tanzania and the results corresponds with other findings elsewhere. Majority of the revealed species have already been reported in soda lake but this study reports the dominance of *Oceanibaculaceae*, *Rhizobiaceae* and *Izemoplasmataceae* for the first time in soda lake habitats. The high relative abundances of unknown, uncultured and candidate phyla species suggest new taxa and that only a fraction of the bacteria has been discovered. The community diversity index values exhibited a high diversity of rare species, indicating a rich ecological environment. The presence of a significant number of species and their even distribution across various sampling points suggests that the lake supports a balanced and diverse ecosystem.

Presence of Actinomycetota species in this study was of interest as it concurs with our earlier findings from Lake Natron and Momella soda lakes (Kaale et al. 2022, Lema et al. 2022, Kaale et al. 2024) that these soda lakes may be potential source of bacteria of biotechnological importance. The results of this study provide a foundation for those cultivation-dependent research focused on understanding the physiology and biochemistry of the as-yet uncultured taxa. While this study is the first to identify several bacterial species inhabiting Lake Natron, we acknowledge certain limitations. The sampling on the western side of Lake Natron only and pooling of DNA samples (due to financial constraints) may have led to the underrepresentation or overrepresentation of some species. We recommend for future studies to cover more geographical sites with more specific DNA samples to ensure a more accurate representation of each species. Additionally, the significant presence of unknown species could be addressed by incorporating both cultivation-dependent and cultivation-independent techniques, which would provide a more comprehensive understanding of the microbial diversity. The integration of both cultivation-dependent and cultivation-independent techniques will not only provide a comprehensive understanding of the specific microbial community but also help uncover rare and previously uncharacterized species. This combined approach enhances the depth of ecological insights, allowing for a more thorough exploration of microbial diversity. Moreover, the study of bacteria functional metagenomics is of importance as it complements the information about their roles in ecology.

Acknowledgments

This research was partially funded by The Nelson Mandela Institute of Science and Technology (Arusha, Tanzania) through the World Bank's Higher Education for economic Transformation Project (HEET) as part of Mr. SE Kaale's PhD studies.

Ethical statement

Neither endangered nor protected species found at Lake Natron were included in the study. The research permits number 2022-428-NA-2022-165 for field sampling were given by the Tanzania National Parks (TANAPA), Tanzania Wildlife Research Institute (TAWIRI), Tanzania Wildlife Management Authority (TAWA), and Tanzania Commission for Science and Technology (COSTECH).

Conflicts of interests

The authors have no relevant conflict of interests to disclose.

References

Ando N, Barquera B and Bartlett DH 2021 The molecular basis for life in extreme environments. *Annu. Rev. Biophys*. 50: 343–372.

Anil Kumar P, Srinivas TNR and Madhu S 2012 *Cecembia lonarensis* gen. nov., sp. nov., a haloalkalitolerant bacterium of the family Cyclobacteriaceae, isolated from a haloalkaline lake and emended descriptions of the genera Indibacter, Nitritalea and Belliella. *Int. J. Syst. Evol. Microbiol*. 62: 2252–2258.

Bardenhorst SK, Vital M, Karch A and Rübsamen N 2022 Richness estimation in microbiome data obtained from denoising pipelines. *Comput. Struct. Biotechnol. J*. 20: 508–520.

Bovio-Winkler P, Guerrero LD and Erijman L 2023 Genome-centric metagenomic insights into the role of Chloroflexi in anammox, activated sludge and methanogenic reactors.

BMC Microbiol. 23(1): 45.

Cabello-Yeves PJ, Picazo A and Roda-Garcia JJ 2023 Vertical niche occupation and potential metabolic interplay of microbial consortia in a deeply stratified meromictic model lake. *Limnol. Oceanogr.* 68(11): 2492–2511.

Caporaso JG, Kuczynski J and Stombaugh J 2010 QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods. 7*(5): 335–336.

Chavan V, Mulaje S and Mohalkar R 2013 A Review on Actinomycetes and their Biotechnological Application. *Intern. J. Pharm. Sci. Res.* 4: 1730–1742.

Chen YH, Chiang PW and Rogozin DY 2021 Salvaging high-quality genomes of microbial species from a meromictic lake using a hybrid sequencing approach. *Commun. Biol*. 4(1): 996.

Chhetri G, Kim I and Kang M 2022 *Devosia rhizoryzae* sp. nov., and *Devosia oryziradicis* sp. nov., novel plant growth promoting members of the genus *Devosia*, isolated from the rhizosphere of rice plants. *J. Microbiol.* 60: 1–10

Clarisse L, Van Damme M and Gardner W 2019 Atmospheric ammonia (NH_3) emanations from Lake Natron's saline mudflats. *Sci. Rep.* 9: 1–12.

Gopalakrishnan S, Sathya A and Vijayabharathi R 2015 Plant growth promoting rhizobia: challenges and opportunities. *3 Biotech.* 5: 355–377.

Hamisi MI, Lugomela C, Lyimo TJ, Bergman B and Díez B 2017 Plankton composition, biomass, phylogeny and toxin genes in Lake Big Momela, Tanzania *Afr. J. Aquat. Sci.* 42: 109–121.

Hilton SK, Castro-Nallar E and Pérez-Losada M 2016 Metataxonomic and metagenomic approaches vs. culture-based techniques for clinical pathology. *Front. Microbiol.* 7: 484.

Hur M and Park S 2019 Identification of Microbial Profiles in Heavy-Metal-Contaminated Soil from Full-Length 16S rRNA Reads Sequenced by a PacBio System. *Microorganisms*. *7*(9): 357.

Jeilu O, Gessesse A and Simachew A 2022 Prokaryotic and eukaryotic microbial diversity from three soda lakes in the East African Rift Valley determined by amplicon sequencing. *Front. Microbiol.* 13: 999876.

Kaale SE, Mahadhy A, Damas M, Mgina CA and Lyimo TJ 2022 Phylogenetic diversity of Actinobacteria from Momela soda lakes, Arusha National Park, Tanzania. *Afr. J. Aquat. Sci.* 47(2): 149–162.

Kaale SE, Machangu RS and Lyimo TJ 2024 Molecular characterization and phylogenetic diversity of actinomycetota species isolated from Lake Natron sediments at Arusha, Tanzania. *Microbiol. Res*. 278: 127543.

Kim BR, Shin J and Guevarra RB 2017 Deciphering diversity indices for a better understanding of microbial communities. *J. Microbiol. Biotechnol*. 27(12): 2089–2093.

Koziaeva VV, Sorokin DY, Kolganova TV and Grouzdev DS 2023 *Magnetospirillum sulfuroxidans sp*. nov., capable of sulfur-dependent lithoautotrophy and a taxonomic reevaluation of the order Rhodospirillales. *Syst. Appl. Microbiol*. 46(3): 126406.

Lai Q, Yuan J, Wu C and Shao Z 2009 Oceanibaculum indicum gen. nov., sp. nov., isolated from deep seawater of the Indian Ocean. *Int. J. Syst. Evol. Microbiol*. 59(7): 1733–1737.

Lema WS, Mahadhy A, Damas M, Munissi JJ and Lyimo TJ 2022 Characterisation and Antimicrobial Potential of Actinobacteria Isolated from Momela Soda Lakes, Tanzania. *Tanz. J. Sci*. 48: 607–622.

Luo W, Li H and Kotut K 2017 Molecular diversity of plankton in a tropical crater lake switching from hyposaline to subsaline conditions: Lake Oloidien, Kenya. *Hydrobiologia*. 788: 205–229.

Maghembe R, Michael A, Harish A, Nyandoro SS, Lyantagaye SL and Hati-Kaul R 2021 Draft Genome Sequence of *Limnospira* sp. Strain BM01, Isolated from a Hypersaline Lake of the Momela Ecosystem in Tanzania. *Microbiol. Resource Announ*. 10(16): 10–128.

Mgimwa EF, John JR and Lugomela CV 2021 The influence of physical–chemical variables on phytoplankton and lesser flamingo (*Phoeniconaias minor)* abundances in Lake Natron, Tanzania. *Afr. J. Ecol*. 59:667–675.

Msaki GL, Kaale SE, Njau KN and Lyimo TJ

2023 Bacterial communities' structure in constructed wetlands for municipal and industrial wastewater treatment in Tanzania. *Water Pract. Technol*. 20: 23155.

Morris EK, Caruso T and Buscot F 2014 Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. *Ecol. Evol*. 4: 3514–3524.

Mousavi SA, Willems A and Nesme X 2015 Revised phylogeny of *Rhizobiaceae:* proposal of the delineation of *Pararhizobium gen. nov.,* and 13 new species combinations. *Syst. Appl. Microbiol.* 38: 84–90.

Nonga HE, Mdegela RH and Sandvik M 2017 Cyanobacteria and cyanobacterial toxins in the alkaline-saline Lakes Natron and Momela, Tanzania. *Tanz. Vet. J*. 32: 108–116.

Oren A and Garrity GM 2021 Valid publication of the names of forty-two phyla of prokaryotes. *Int. J. Syst. Evol. Microbiol.* 71: 005056.

Omeroglu E, Sudagidan M and Yurt NZ 2021 Microbial community of soda Lake Van as obtained from direct and enriched water, sediment and fish samples. *Sci. Rep*. 11: 18364.

Palkova L, Tomova A and Repiska G 2021 Evaluation of 16S rRNA primer sets for characterisation of microbiota in paediatric patients with autism spectrum disorder. *Sci. Rep*. 11: 6781.

Philip N and Mosha S 2012 Salt Lakes of the African Rift System: A Valuable Research Opportunity for Insight into Nature's Concenrtated Multi-Electrolyte Science. *Tanz. J. Sci*. 38:1–13.

Rojas P, Rodríguez N, de la Fuente V, Sánchez-Mata D, Amils R and Sanz JL 2018 Microbial diversity associated with the anaerobic sediments of a soda lake (Mono Lake, California, USA). *Can. J. Microbiol*. 64:385–392.

Skennerton C, Haroon M and Briegel A 2016 Phylogenomic analysis of *Candidatus* 'Izimaplasma' species: free-living representatives from a *Tenericutes* clade found in methane seeps. *ISME J*. 10: 2679–2692.

Sorokin DY, Berben T and Melton ED 2014 Microbial diversity and biogeochemical cycling in soda lakes. *Extremophiles*. 18: 791–809.

Thomas T, Gilbert J and Meyer F 2012 Metagenomics - a guide from sampling to data analysis. *Microb. Inform. Exp*. 2: 3.

Vanni C, Schechter MS and Acinas SG 2022 Unifying the known and unknown microbial coding sequence space. *Elife*. 11: e67667.

Vavourakis CD, Andrei AS and Mehrshad M 2018 A metagenomics roadmap to the uncultured genome diversity in hypersaline soda lake sediments. *Microbiome*. 6:1–18.

Vavourakis CD, Ghai R and Rodriguez-Valera F 2016 Metagenomic insights into the uncultured diversity and physiology of microbes in four hypersaline soda lake brines. *Front. Microbiol.* 7: 211.

Vavourakis CD, Mehrshad M and Balkema C 2019 Metagenomes and metatranscriptomes shed new light on the microbial-mediated sulfur cycle in a Siberian soda lake. *BMC Biol.* 17: 69.

Wang M, Zhang X and Shu Z 2022 Bacterial and archaeal communities within the alkaline soda Langaco Lake in the Qinghai-Tibet Plateau. *Ann. Microbiol.* 72: 33.

Yakimov MM, Giuliano L and Chernikova TN 2001 *Alcalilimnicola halodurans* gen. nov., sp. nov., an alkaliphilic, moderately halophilic and extremely halotolerant bacterium, isolated from sediments of soda-depositing Lake Natron, East Africa Rift Valley. *Int. J. Syst. Evol. Microbiol.* 51: 2133−2143.

Yona C, Makange M and Moshiro E 2023 Water pollution at Lake Natron Ramsar site in Tanzania: A threat to aquatic life. *Ecohydrol. Hydrobiol*. 23: 98–108.

Zorz JK, Sharp C and Kleiner M 2019 A shared core microbiome in soda lakes separated by large distances. *Nat. Commun*. 10: 4230.

APPEDICES

Appendix 1: Classification at phylum level of bacterial communities retrieved from sample of Lake Natron water in Arusha, Tanzania (- = not retrieved).

Name	NW1		NW2		NW3	
Phyla Classification	**Read Count**	**%**	**Read Count**	**%**	**Read Count**	**%**
Proteobacteria	13763.0	98.46	2752.0	70.46	170.0	12.80
Actinomycetota	86.0	0.62	108.0	2.76	115.0	8.66
Bacteroidota	71.0	0.51	379.0	9.70	341.0	25.68
Firmicutes	19.0	0.14	433.0	11.09	469.0	35.32
Unknown	10.0	0.07	78.0	2.00	90.0	6.78
Chloroflexi	7.0	0.05	41.0	1.05	33.0	2.48
Spirochaetota	4.0	0.03	6.0	0.15	8.0	0.60
Desulfobacterota	4.0	0.03	49.0	1.25	51.0	3.84
Gemmatimonadota	3.0	0.02	4.0	0.10	7.0	0.53
Hydrogenedentes	3.0	0.02	-	-	2.0	0.15
Cloacimonadota	2.0	0.01	17.0	0.44	15.0	1.13
Nitrospirota,	2.0	0.01	1.0	0.03	-	-
Verrucomicrobiota	-	-	10.0	0.26	5.0	0.38
Bdellovibrionota	-	-	9.0	0.23	1.0	0.08
Halanaerobiaeota	-	-	7.0	0.18	1.0	0.08
SAR324_clade,	2.0	0.01	5.0	0.13	-	-
Fusobacteriota	-	-	1.0	0.03	4.0	0.30
Fibrobacterota,	-	-	2.0	0.05	3.0	0.23
Cloacimonadota,	2.0	0.01	-	-	-	-
Dependentiae	2.0	0.01	-	-	-	-
Deinococcota, FCPU426, Latescibacterota, Thermotogota, s	-	-	-		@2.0	@0.15
Thermotogota	-	-	2.0	0.05	-	-
Sva0485, LCP,	-	-	1.0	0.03	-	-
Armatimonadota, Methylomirabilota, Myxococcota, LCP, Cyanobacteria	-	-	-	-	@1.0	@0.08

Appendix 2: Classification at class level of bacterial communities retrieved from sample of Lake Natron water in Arusha, Tanzania (- = not retrieved).

Name	NW1		NW2		NW3	
Class classification	**Read Count**	**%**	**Read Count**	**%**	**Read Count**	**%**
Alphaproteobacteria	13078.0	93.59	2645.0	67.72	16.0	1.20
Gammaproteobacteria	684.0	4.90	104.0	2.66	154.0	11.60
Actinobacteria	74.0	0.53	52.0	1.33	16.0	1.20
Bacteroidia	62.0	0.44	352.0	9.01	316.0	23.80
Bacilli	12.0	0.09	317.0	8.12	308.0	23.19
Unknown	13.0	0.09	92.0	2.36	103.0	7.76
Clostridia	3.0	0.02	52.0	1.33	83.0	6.25

Coriobacteriia,	3.0	0.02	48.0	1.23	73.0	5.50
Dethiobacteria,	4.0	0.03	52.0	1.33	59.0	4.44
Anaerolineae	5.0	0.04	41.0	1.05	31.0	2.33
Desulfobulbia	-	-	22.0	0.56	24.0	1.81
Acidimicrobiia	7.0	0.05	2.0	0.05	23.0	1.73
Cloacimonadia,	3.0	0.02	17.0	0.44	15.0	1.13
Rhodothermia	5.0	0.04	16.0	0.41	13.0	0.98
Spirochaetia	4.0	0.03	6.0	0.15	8.0	0.60
Ignavibacteria,	3.0	0.02	11.0	0.28	10.0	0.75
Desulfobacteria	-	-	10.0	0.26	10.0	0.75
Desulfuromonadia,	3.0	0.02	13.0	0.33	8.0	0.60
Desulfovibrionia,	-	-	2.0	0.05	7.0	0.53
Other classes represented by less than 0.5% in all sites (Superscript 1,2,3 means present in 1 = NW1; 2 = NW2 and 3 = NW3)						
Hydrogenedentia[1], Nitrospiria[1], SAR324_clade[1,2], Thermoleophilia[1], Gemmatimonadetes[1], Babeliae[1], Halanaerobiia[2], Oligoflexia[2], BD2[2,3], Verrucomicrobiae[2], Limnochordia[2], Bdellovibrionia[2], MB[2], Thermoleophilia[2,3], Uncultured[2,3], Fibrobacteria[2,3], Thermotogae[2], Omnitrophia[2], Sva0485[2], Thermoanaerobacteria[2], LCP[2], Nitrospiria[2], Syntrophia[2], Fusobacteriia[2,3], Lentisphaeria[2], Verrucomicrobiae[3], Desulfitobacteriia[3], Gemmatimonadetes[3], Natranaerobiia[3], Deinococci[3], SJA[3], FCPU426[3], Thermotogae[3], Hydrogenedentia[3], Methylomirabilia[3], Incertae_Sedis[3], Polyangia[3], Halanaerobiia[3], Kiritimatiellae[3], Latescibacterota[3], Longimicrobia[3], Oligoflexia[3], KD4[3], Latescibacteria[3], Chloroflexia[3], Syntrophia[3], LCP[3], Cyanobacteriia[3]						

Appendix 3: Classification at family level of bacterial communities retrieved from sample of Lake Natron water in Arusha, Tanzania (- = not retrieved).

Name	**NW1**		**NW2**		**NW3**	
Family classification	**Read Count**	**%**	**Read Count**	**%**	**Read Count**	**%**
Oceanibaculaceae	7321.0	52.43	7.0	0.18	-	-
Izemoplasmataceae	-	-	219	5.61	166	12.5
Devosiaceae	5013.0	35.90	-	-	-	-
OPB41	2.0	0.01	48.0	1.23	73.0	5.50
Dethiobacteraceae	-	-	52.0	1.33	59.0	4.44
Acholeplasmataceae	4.0	0.03	48.0	1.23	65.0	4.89
Oxalobacteraceae	664.0	4.76	-	-	-	-
Rhizobiaceae	599.0	4.29	2602.0	66.62	-	-
Unknown	161.0	1.15	189.0	4.84	3.0	0.23
Nocardioidaceae	35.0	0.25	-	-	2.0	0.15
Cellulomonadaceae	31.0	0.22	-	-	-	-
ML635J	29.0	0.21	187.0	4.79	149.0	11.22
Peptostreptococcales	-	-	21.0	0.54	51.0	3.84
Rhodocyclaceae	-	-	13	0.33	39	2.94
RF39	-	-	-	-	29	2.18
Erysipelotrichaceae	-	-	-	-	22.0	1.66
Desulfobulbaceae	-	-	15.0	0.38	22.0	1.66
Comamonadaceae	-	-	8.0	0.20	17	1.28

Bacteroidetes_BD2	11.0	0.08	90.0	2.30	70.0	5.27
Anaerolineaceae	5.0	0.04	36.0	0.92	26.0	1.96
Uncultured	10.0	0.07	10.0	0.26	16.0	1.20
Micrococcaceae	-	-	35.0	0.90	-	-
Cyclobacteriaceae	-	-	25.0	0.64	32.0	2.41
Microbacteriaceae	7.0	0.05	-	-	-	-
Nitrincolaceae	-	-	-	-	15.0	1.13
Woeseiaceae	2.0	0.01	-	-	15.0	1.13
Ilumatobacteraceae	-	-	-	-	14.0	1.05
Balneolaceae	-	-	16	0.41	13.0	0.98
Peptostreptococcaceae	-	-	-	-	12.0	0.90
Bacillaceae	-	-	10.0	0.26	11.0	0.83
BSN166	-	-	-	-	10.0	0.75
Lentimicrobiaceae	6	0.04	-	-	10.0	0.75
Fusibacteraceae	-	-	-	-	9.0	0.68
Other families represented by less than 0.5% in all sites (Superscript 1,2,3 means present in 1 = NW1; 2 = NW2 and 3 = NW3)						
Parvibaculaceae[1], Lentimicrobiaceae[1], Anaerolineaceae[1], Nitrosomonadaceae[1], Balneolaceae[1], Saprospiraceae[1,2], Dethiobacteraceae1, Izemoplasmataceae[1,2], Spirochaetaceae[1,2], BSN166[1], Bacillaceae[1], Hydrogenedensacea[1], Xanthobacteraceae[1], Geoalkalibacteraceae[1,2,3] Cyclobacteriaceae[1], Acholeplasmataceae[1], Nitrospiraceae[1], SAR324_clade[1], Pseudohongiellaceae[1], Gemmatimonadaceae[1,3], Magnetospiraceae[1], Flavobacteriaceae[1,2,3], Vibrionaceae[2,3], Caulobacteraceae[2], MSBL8[2], Halobacteroidaceae[2,3], Desulfurivibrionaceae[2,3], RF39[2], Nitriliruptoraceae[2], 0319[2], Marinilabiliaceae[2,3], SAR324_clade[2], BD2[2], Fusibacteraceae[2], Nitrosomonadaceae[2,3], Hungateiclostridiaceae[2], Desulfosarcinaceae[2,3], Streptomycetaceae[2], Puniceicoccaceae[2], Desulfobacteraceae[2], Geodermatophilaceae[2], Lentimicrobiaceae[2], Desulfococcaceae[2], SBR1031[2], Bacteriovoracaceae[2], MB[2], Sphingomonadaceae[2,3], Acidithiobacillaceae[2,3], Geopsychrobacteraceae[2], Beijerinckiaceae[2,3], Fibrobacteraceae[2,3], Devosiaceae[2], Planococcaceae[2], Ignavibacteriaceae[2], 67[2], Methyloligellaceae[2,3], Kosmotogaceae[2,3], Omnitrophaceae[2], CW[2], A4b[3], Paenibacillaceae[3], AT[3], Hydrogenedensaceae[3], Desulfonatronaceae[3], Alcanivoracaceae1[3], Micromonosporaceae[3], Bacteroidetes_vadinHA17[3] Williamwhitmaniaceae[3], Desulfatiglandaceae[3], 37[3], Microtrichaceae[3], Aeromonadaceae[3], MSBL8[3], Rhodobacteraceae[3], Peptococcaceae[3] IMCC26256[3], Microbacteriaceae[3], Thioalkalispiraceae[3], Natranaerobiaceae[3], Alteromonadaceae[3], Trueperaceae[3], SJA[3], Desulfitobacteriales[3], 0319[3], Desulfomicrobiaceae[3], FCPU426[3], Caldilineaceae[3], Pseudohongiellaceae[3], Puniceicoccaceae[3], Planococcaceae[3], Moraxellaceae[3], BD2[3], Geopsychrobacteraceae[3], Neisseriaceae[3], Desulfonatronovibrionaceae[3], Micrococcaceae[3], NS11[3], Rhodanobacteraceae[3], MAT[3], Rokubacteriales[3], CMW[3], FTLpost3[3], Fibrobacterales[3], MidBa8[3], Kineosporiaceae[3], WCHB1[3], Latescibacterota[3], Longimicrobiaceae[3], Carnobacteriaceae[3], Halomonadaceae[3], Haloplasmataceae[3], KD4[3], Latescibacteraceae[3], Chloroflexaceae[3], Xanthobacteraceae[3], Hungateiclostridiaceae[3], RBG[3], LCP[3], SHA[3], Cyanobacteriaceae[3]						

Appendix 4: Classification at genus level of bacterial communities retrieved from sample of Lake Natron water in Arusha, Tanzania (- = not retrieved).

Name	**NW1**		**NW2**		**NW3**	
Genus classification	**Read Count**	**%**	**Read Count**	**%**	**Read Count**	**%**
Allorhizobium	-	-	2562.0	65.59	-	-
Oceanibaculum	7321.0	52.44	-	-	-	-
Devosia	5003.0	35.84	-	-	-	-
Izimaplasma	4.0	0.03	217.0	5.56	166	12.5
Massilia	637.0	4.56	-	-	-	-
Aminobacter	507.0	3.63	-	-	-	-
Unknown	326.0	2.34	312.0	7.99	288	21.69
Acholeplasma	4.0	0.03	47.0	1.20	65	4.89
OPB41	2.0	0.01	48.0	1.23	73	5.5
Uncultured	29.0	0.21	75.0	1.92	95	7.15
ADurb	-	-	32.0	0.82	25	1.88
Dethiobacter	-	-	19.0	0.49	24	1.81
ML635J	29.0	0.21	187.0	4.79	149	11.22
Anoxynatronum	-	-	11.0	0.28	34	2.56
RF39	-	-	6.0	0.15	29	2.18
Erysipelothrix	-	-	15.0	0.38	22	1.66
Azoarcus	-	-	-	-	20	1.51
JTB255_marine_benthic_group	2.0	0.01	10.0	0.26	15	1.13
Nitrincola	-	-	-	-	10	0.75
BSN166	-	-	-	-	10	0.75
Bacillus, Lentimicrobiaceae, Fusibacter, F1, Alkaliphilus	6.0	0.04	3.0	0.08	9	0.68
Izemoplasmatales	-	-	15.0	0.38	8	0.6
Cecembia	-	-	10.0	0.26	8	0.6
Pseudomonas	-	-	10.0	0.26	7	0.53
Algoriphagus	-	-	-	-	7	0.53
Bacteroidetes_BD2	11.0	0.08	90.0	2.30	70	5.27
Other genera represented by less than 0.5% in all sites (Superscript 1,2,3 means present in 1 = NW1; 2 = NW2 and 3 = NW3)						
Microbacterium[1], Noviherbaspirillum[1], Cellulomonas[1], Parvibaculum[1], MND11, BSN166[1], Hydrogenedensaceae[1], Geoalkalibacter[1,2,3], Izimaplasma[1], Acholeplasma[1], CK061, JTB255_marine_benthic_group[1], Nitrospira[1,2], SAR324_clade[1,2], Bacillus[1,2], Sediminispirochaeta[1,3], Pseudohongiella[1,3], BSN166[2], Thioalkalispira[2] Vibrio[2,3], Phenylobacterium[2], F1[2], MSBL8[2] Azoarcus[2], Desulfurivibrio[2,3], Oceanibaculum[2] Hydrogenophaga[2,3], Nitriliruptoraceae[2], 0319[2], Acetoanaerobium[2], BD2[2], Tindallia[2], Fusibacter[2], Algoriphagus[2], Streptomyces[2], Candidatus_Contubernalis[2,3], SBR1031[2], Peredibacter[2], MB[2], Nitrosomonas[2], UCG[2], Desulfonatronobacter[2], Novosphingobium[2], Alkaliphilus[2], TX1A[2], Desulfuromusa[2,3], Alkalitalea[2], Microvirga[2,3], Puniceicoccus[2], Mesorhizobium[2], Devosia[2], Lysinibacillus[2], Ignavibacterium[2], Spirochaeta[2], 67[2], Mariniradius[2], Alkalispirochaeta[2], Candidatus_Omnitrophus[2,3], Anaerobacillus[2,3], Aquiflexum[2], Desulfotignum[2], CW[2], Sva0485[2], Planktosalinus[1,2,3], Desulfobotulus[2],						

Romboutsia[2], Longispora[2,3], MAT[2], Marinospirillum[2,3], Rhodohalobacter[2], Desulfomicrobium[2,3], SRB2[2], oc32[2], Blastococcus[2], Denitratisoma[2], Aeromonas[2], DMI[2], Desulfonatronum[2,3], CMW[2], Aquimonas[2,3], LCP[2], Sva0996_marine_group[2], Nocardioides[2,3], Bradymonadales[2], Hypnocyclicus, Lenti, Geodermatophilus, Alcanivorax[3], ST[3], Parvibaculum[3], 37[3], Verruc[2,3], Marinobacter[2], Sulfurifustis[2], PB19[2], Desulfatitalea[2,3],Aeromonas[3], MSBL8[3], Hypnocyclicus[3], Nitriliruptoraceae[3], 37[3], Blvii28_wastewater[3], Desulfatiglans[3], Alcanivorax[3], Bacteroidetes_vadinHA17 CK06[3], Alishewanella[3], Truepera[3], SJA[3], TC1[3], 319[3], FCPU426[3], Desulfobulbus[3], Porphyrobacter[3], Nitrosomonas[3], Acetoanaerobium[3], BD2[3], Spirochaeta_2[3], Candidatus_Accumulibacter[3], A4b[3], Ammoniphilus[3], AT[3], Hydrogenedensaceae[3], Natronohydrobacter[3], Thauera[3], Kocuria[3], NS11[3], MAT[3], Wenzhouxiangella[3], Rokubacteriales[3], Ilumatobacter[3], CMW[3], FTLpost3[3], Dechlorobacter[3], Micromonospora[3], BBMC[3], MidBa8[3], Quadrisphaera[3], Candidatus_Berkiella[3], WCHB1[3], Latescibacterota[3], Longimicrobiaceae[3], Alkalibacterium[3], Lunatimonas[3], Haloplasma[3], KD4[3], TX1A[3], FFCH7168[3], Egicoccus[3], Pseudorhodoplanes[3], Alkanindiges[3], RBG[3], Paenisporosarcina[3], LCP[3], SHA[3], Geminocystis_PCC[3], IMCC26256[3], Agrococcus[3], Thioalkalispira[3], Azonexus[3], Lentimicrobium[3], IMCC26207[3]

Appendix 5: Speciation of bacterial communities retrieved from sample of Lake Natron water in Arusha, Tanzania.

Name	**NW1**		**NW2**		**NW3**	
Species classification	**Read Count**	**%**	**Read Count**	**%**	**Read Count**	**%**
Unknown	13795.0	98.80	3091.0	79.13	565.0	42.55
Izimaplasma_uncultured_bacterium			196.0	5.02	139.0	10.47
Oceanibaculum_uncultured_bacterium	32.0	0.23				
ML635J_uncultured_bacterium	29.0	0.21	173.0	4.43	142.0	10.69
Bacteroidetes_BD2_uncultured_bacterium	11.0	0.08	90.0	2.30	69.0	5.20
OPB41_uncultured_bacterium	2.0	0.01	43.0	1.10	64.0	4.82
Uncultured_bacterium	8.0	0.06	28.0	0.72	42.0	3.16
ADurb_uncultured_bacterium			32.0	0.82	25.0	1.88
Uncultured_Firmicutes			22.0	0.56	22.0	1.66
Erysipelothrix_uncultured_bacterium			14.0	0.36	16.0	1.20
Dethiobacter_uncultured_bacterium			16.0	0.41	17.0	1.28
Azoarcus_uncultured_bacterium	6.0	0.04			19.0	1.43
Other cultured species represented by less than 1% in all sites (Superscript 1,2,3 means present in 1 = NW1; 2 = NW2 and 3 = NW3)						
Cellulomonas_sp.[1], Devosia lucknowensis[1], Cecembia lonarensis[2,3], Vibrio metschnikovii[2], Lunatimonas lonarensis[3], Desulfurivibrio alkaliphilus[2], Aquiflexum_aquiflexum_sp[2], Streptomyces flavoviridis[2], Acholeplasma tenericutes[3], Bacillus mannanilyticus[3], Acholeplasma vituli[3], Agrococcus lahaulensis[3], Ammoniphilus sp.[3], Nitrincola sp.[3]						
Other uncultured species represented by less than 1% in all sites (Superscript 1,2,3 means present in 1 = NW1; 2 = NW2 and 3 = NW3)						
Lentimicrobiaceae[1,2,3], Massilia[1], BSN166[1], Hydrogenedensaceae[1,3], SAR324[1], Sediminispirochaeta[1], OPB41[1], CK06[1,3], ADurb[2], Izemoplasmatales[2], Erysipelothrix[2,3], BSN166[2], Thioalkalispira[2], F1[2], MSBL8[2,3], Phenylobacterium[2], ML635J Bacteroidetes[2,3], Anaerobic_digester[2,3], Hydrogenophaga[2,3], Geoalkalibacter[1,2,3], Acetoanaerobium[2], BD2[2], Tindallia_anaerobic_digester[2], Azoarcus[2], Fusibacter[2,3],						

Algoriphagus_sp.[2], Acholeplasma[2,3], Contubernalis[2,3], MB[2], TX1A[2], Desulfuromusa[2,3] Alkalitalea saponilacus[2], Dethiobacter[2,3], Anoxynatronum[2,3], UCG[2], Alkaliphilus_bacterium_YC[2,3], Ignavibacterium[2], Desulfonatronobacter[2], Mariniradius saccharolyticus[2], Alkalispirochaeta[2], Omnitrophus[2], CW Spirochaetes[2], Sva0485[2], Desulfobotulus_sapovorans[2], 0319[2], MAT[2], Marinospirillum_gamma_proteobacterium[2,3], Peredibacter[2,3], Allorhizobium[2], SRB2[2], oc32[2], JTB255[2], Denitratisoma[2], DMI[2], CMW[2], LCP[2], Oceanibaculum indicum, Bradymonadales[2], SAR324[2], Lenti[2], ST_uncultured_Cytophagales[2], 37[2], PB19[2], Desulfatitalea[2,3], BSN166[3], Izemoplasmatales[3], Hypnocyclicus[3]37[3] Sediminispirochaeta[3], Bacteroidetes vadinHA17[3], Blvii28[3], Truepera[3], SJA_Ignavibacteriales_bacterium[3], TC1[3], FCPU426[3], Desulfobulbus[3], Verruc[3], Acetoanaerobium[3], BD2[3], Candidatus_accumulibacter_uncultured_Denitratisoma[3], Longispora[2,3], AT[3], NS11[3], Desulfatiglans[3], Aquimonas[3], MAT[3], Wenzhouxiangella[3], Nitriliruptoraceae[3], Rokubacteriales[3], CMW[3], 0319[3], FTLpost3[3], Dechlorobacter[3], BBMC_bacterium_enrichment[3], Pseudohongiella pseudohongiella spirulinae[3], Quadrisphaera_quadrisphaera_granulorum[3], Candidatus_berkiella[3], WCHB1[3], Latescibacterota[3], Longimicrobiaceae[3], Haloplasma[3], KD4[3], TX1A[3], Anaerobacillus[3], FFCH7168[3], Alkanindiges[3] , Spirochaeta[3], Geminocystis_PCC_Cyanobacterium_sp[3], IMCC26256 Actinomycetales[3], Azonexus[3], A4b[3], LCP[3]