Archives of Microbiology
https://doi.org/10.1007/s00203-021-02227-4

ORIGINAL PAPER

In silico determination of nitrogen metabolism in microbes from extreme conditions using metagenomics

Lulit Tilahun[1] · Asfawossen Asrat[2] · Gary M. Wessel[3] · Addis Simachew[1]

Received: 10 August 2020 / Revised: 21 December 2020 / Accepted: 11 February 2021

Abstract

The acid ponds of the Danakil Depression in northern Ethiopia are polyextreme environments that exceed the normal physicochemical limits of pH, salinity, ion content, and temperature. We tested for the occurrence of DNA-based life in this environment using Metagenomic Shotgun DNA sequencing approaches. The obtained sequences were examined by the bioinformatic tools MetaSpades, DIAMOND and MEGAN 6-CE, and we were able to bin more than 90% of the metagenomics contigs of Dallol and Black Water to the *Bacteria* domain, and to the *Proteobacteria* phylum. Predictions of gene function based on SEED disclosed the presence of different nutrient cycles in the acid ponds. For this study, we focused on partial or completely sequenced genes involved in nitrogen metabolism. The KEGG nitrogen metabolism pathway mapping results for both acid ponds showed that all the predicted genes are involved directly or indirectly in the assimilation of ammonia and no dissimilation or nitrification process was identified. Furthermore, the deduced nitrogen fixation in the two acid ponds based on SEED classification indicated the presence of different sets of nitrogen fixing (nif) genes for biosynthesis and maturation of nitrogenase. Based on the in silico analysis, the predicted proteins involved in nitrogen fixation, especially the cysteine desulfurase and [4Fe-4S] ferredoxin, from both acid ponds are unique with less than 80% sequence similarity to the next closest protein sequence. Considering the extremity of the environmental conditions of the two acid ponds in the Danakil depression, this metagenomics dataset can add to the study of unique gene functions in nitrogen metabolism that enable thriving biocommunities in hypersaline and highly acidic conditions.

Keywords Extremophiles · Evolution · Bacteria

Introduction

The repertoires of habitats for biological occupancy are vast and varied. Environmental conditions are measured generally by physicochemical metrics such as acidity/alkalinity, salinity, temperature, concentrations of heavy metals, etc. Human environmental requirements for survival and proliferation are relatively narrow, and they are greatly assisted by clothing and accessories. Most organisms instead rely on gene sets to enable survival and/or proliferation in their environment. The relationship between organism and environment in this environment is of great importance for understanding evolution, gene function, and biotechnological applications.

Solfataric fields are poly-extreme environments found all over the world including in North America, Europe, Asia, and Africa and are characterized by low pH and high temperatures (Spear et al. 2006; Mayer 2017; Crognale et al. 2018). The Danakil depression is located in the northern

Communicated by Erko Stackebrandt.

✉ Gary M. Wessel
rhet@brown.edu

Lulit Tilahun
lulit.tilahun@aau.edu.et

Asfawossen Asrat
asfawossen.asrat@aau.edu.et

Addis Simachew
addis.simachew@aau.edu.et; addissimachew@gmail.com

[1] Institute of Biotechnology, Addis Ababa University, P.O. Box 1176, Addis Ababa, Ethiopia

[2] School of Earth Sciences, Addis Ababa University, P.O. Box 1176, Addis Ababa, Ethiopia

[3] Department of Molecular and Cell Biology and Biochemistry, Brown University, Box G, 185 Meeting Street, Providence, RI 02912, USA

Published online: 07 March 2021

part of the larger Afar Depression of northeastern Ethiopia, and is one of the most extreme of quantitated environments on this planet with recorded temperatures reaching 50–60 °C (120–140°F). The depression is a segment of proto-oceanic crust where lowland plains are 116–125 m below sea level. These plains are split by fault blocks and are dotted with shield volcanoes including the active Ert'ale volcano and the quasi-active, subaerial Dallol volcano buried under relatively thick sediment covers (Illsley-Kemp 2017) and references therein). Most hydrothermal-geothermal activities in the Danakil depression are confined to the Dallol summit crater and to hyper-saline brine called the "Black Water pond", located about ~ 2 km southwest of mount Dallol in an area locally called "Black Mountain" (Franzson et al. 2015; Asrat 2016). Nearly permanent, hyper-acidic and hypersaline hot brine springs and ponds are manifestations of ongoing volcanic- hydrothermal activity beneath the Dallol summit (Asrat 2016).

The Danakil depression, particularly the Dallol summit crater and the satellite eruption of Black Water, has an extraterrestrial appearance and quality. These sites could rightly be considered inhospitable to living beings. Yet both molecular based as well as lipid biomarker and carbon stable isotope studies conducted on different Dallol hydrothermal systems and solfataric fields signal the presence of life (Carrizo 2019) (Gómez 2019). Alternative to the postulation of the starting of life in deep-sea hydrothermal vents, volcanic pools and hot springs on land also provides basic nutrients, energy and conditions as well as reaction conditions to create complex molecules as precursors of life (Damer 2016; Kranendonk et al. 2017). Hence, the acid ponds scattered in the Danakil depression are ideal places to study how the poly-extreme environments affect organic chemistry and life on Earth, or on other sites in the solar system (Wächtershäuser 2006; Barbieri and Cavalazzi 2014).

From a biotechnological point of view, extreme environments are impactful since life adapted to such conditions likely means a plethora of genes were modified for survival in this environment. Extremophiles evolve rapidly to adjust to the dynamic and extreme conditions and thus become pioneers in producing novel bio-products (Burg 2003; Hamdi 2007; Kikani and Shukla 2010; Börne 2013; Li 2014). Hence, exploring the acid ponds in the Danakil depression is important to investigate the diversity and functionality of extremophiles, and for discovering bio-resources of use to humans. Metagenomic methods are preferred in such conditions to culture-based studies since only 1% of the estimated total microorganisms on earth are cultivable by normal means (Schmeisser et al. 2007). The development of different genome sequencing platforms and bioinformatic tools enabled the direct analysis of natural microbial communities (Su 2012). Furthermore, the integration of in silico analyses to metagenomics allows holistic investigation of extremophiles with biotechnological potential. Synergistically implementing multiple methodologies instead of a single approach can increase the rates of discovering target genes and metabolic pathways from metagenomes (Barriuso and Martínez 2015; Kodzius and Gojobori 2015; Alves et al. 2017). Hence, cataloging the ORFs (open reading frames) present in the metagenome of extremophilic communities based on metabolic function can be a valuable source for synthetic biology and genetic engineering to create or improve bio-resources (Kodzius and Gojobori 2015).

Furthermore, the Dallol hot springs and surrounding hydrothermal fields (Black and Yellow Lakes also known as Black Water acid pond and Gaet'ale by the locals) are known to contain hydrocarbon fluids, mixed with sulfides/sulfates and different types of industrial minerals and excess elemental nitrogen gas N_2 (Gebresilassie et al. 2011; Kotopoulou 2019). Carrizo et al. (Carrizo 2019) predicted the possible pathways of carbohydrate and sulfate in autotrophs by measuring $\delta^{13}C$ and $\delta^{34}S$ ratios in Dallol hot springs respectively. Though biological transformation of N_2 to its oxidized or reduced forms in various acidic environments has been reported (Hamilton 2011), efficient NH_3/NH_4 recycling reflecting sufficient nitrogen supply for sustaining autotrophs primary production has not been inferred in hot springs of Danakil depression. Therefore, the objective of this study was to study the biological communities, in the brine of the two most extremely acidic ponds of the Danakil depression, Dallol and Black Water, soon after the 2015 phreatic eruption, using metagenomic shotgun sequencing approaches. In addition, to predict the nutrient cycle, especially the nitrogen cycle, the metagenomes of the two ponds were examined as a self-contained, open system with a distinct functional profile. Further emphasis was given to the translated products of Open Read Frames (ORFs) for nitrogen fixation and to predicting the structure and model of selected proteins involved in nitrogen fixation.

Materials and methods

Description of the study site

This study was conducted on acid ponds at the Dallol crater and the Black Water located in the Danakil depression of northern Ethiopia (Fig. 1). Brine samples were collected on February 4, 2015, only a few weeks after the January 2015 phreatic eruption at the summit crater. This phreatic eruption led to the draining and drying up of most of the active springs at the crater and one pond (located at UTM 0639983E, 1574575 N, 95 m Below Sea Level) was available for sampling at the time (Fig. 2a). Other samples were collected the same day from the Black Water

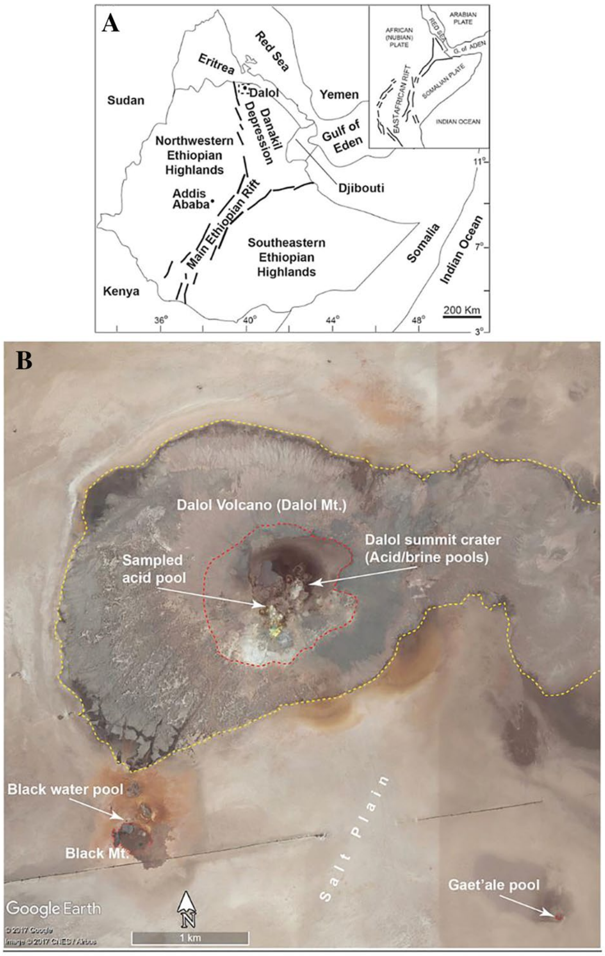

Fig. 1 Location of Danakil depression on map of Ethiopia (**a**) and Satellite image of Dallol crater, Gaet'ale and Black water ponds (**b**)

pond (Fig. 2b) located at UTM 0638790E, 1572643 N, 114 below sea level, just SW of the crater.

Samples were collected randomly in triplicates from accessible parts of the ponds considering controlled minimal human and animal contacts to reduce chances of contamination. All the sample areas are small ponds and were designated as Dal (Dallol) and Bla (Black Water). Onsite measurements of pH, temperature and conductivity could not be done due to the unsafe conditions for the measuring devices on the sample sites after the phreatic eruption at the Dallol summit crater. Hence, measurement of pH was done using 0–14 pH indicator paper sticks (Fisherbrand™) while onsite measurements of temperature and conductivity were avoided. Salinity was measured using a refractometer (DIGIT-0120 ATC, VWR) by diluting the brine samples 1 to 10× with deionized water. The color, texture and other physicochemical characteristics of the samples are listed in Table 1.

Sample collection

Brine samples were collected from each sample site on February 4, 2015 (cool season) using sterile High Density Polyethylene (HDPE) bottles for cell harvesting. Additional 50 ml of brine samples were collected in conical plastic tubes (BD Falcon TM, Greiner Bio-One, Germany), for other biological and physicochemical analysis. The samples were stored in an icebox containing cooling elements and transported immediately to a temporary working space at Yara Dallol BV, Dallol Camp for processing and cell harvesting.

A total of 2000 ml of brine sample from Dallol (200 ml per filter) was filtered without dilution using a 0.22 µm GE® polycarbonate filter membranes. However, due to its high viscosity, only 500 ml of brine (50 ml per filter) from Black Water was filtered without dilution using a 0.22 µm GE® polycarbonate filter membranes. Then, the filter membranes were stored in a 2 ml Lasany® internal treaded Cryo-vials and sucrose lysis buffer was added for preserving the cells as previously described (Mitchell and Takacs-Vesbach 2008).

For physicochemical analysis, 200 ml of the filtrate was placed in sterile containers without acid treatment. All the processed samples (the filter membranes for metagenomic DNA extraction and the filtrate brine for physicochemical analysis) were transported to the Microbial Biotechnology laboratory at Addis Ababa University in iceboxes containing cooling elements and then stored at -20 °C and 4 °C, accordingly until further analysis.

Physico-chemical analysis and water isotope measurement

Stable isotope compositions of oxygen and hydrogen from site samples were determined at the School of Earth Sciences, Isotope Hydrology Laboratory, Addis Ababa University, using the Los Gatos Research Off-Axis Integrated Cavity Output Spectroscopy (OA-ICOS), following standard procedures as recommended (Emanuelsson 2015). However, only partial hydrochemical analysis was feasible to conduct at the Laboratory of Addis Ababa Environmental Protection Authority. The content of magnesium (Mg^{2+}) was determined using an atomic absorption spectrometer and the graphite method after 10,000X dilution. Total phosphorus (TP), nitrate (NO_3^-), nitrite (NO_2^-) and sulfate (SO_4^{2-}) were assayed according to the Standard Methods for the Examination of Water and Wastewater (APHA, 1998).

The Chemical Oxygen Demand (COD) was measured using a Hach COD reactor 45,600, USA. Chloride (Cl^-) testing was accomplished using argentometric titration where potassium chromate was used as indicator. Organic nitrogen load in both sample sites was not measured.

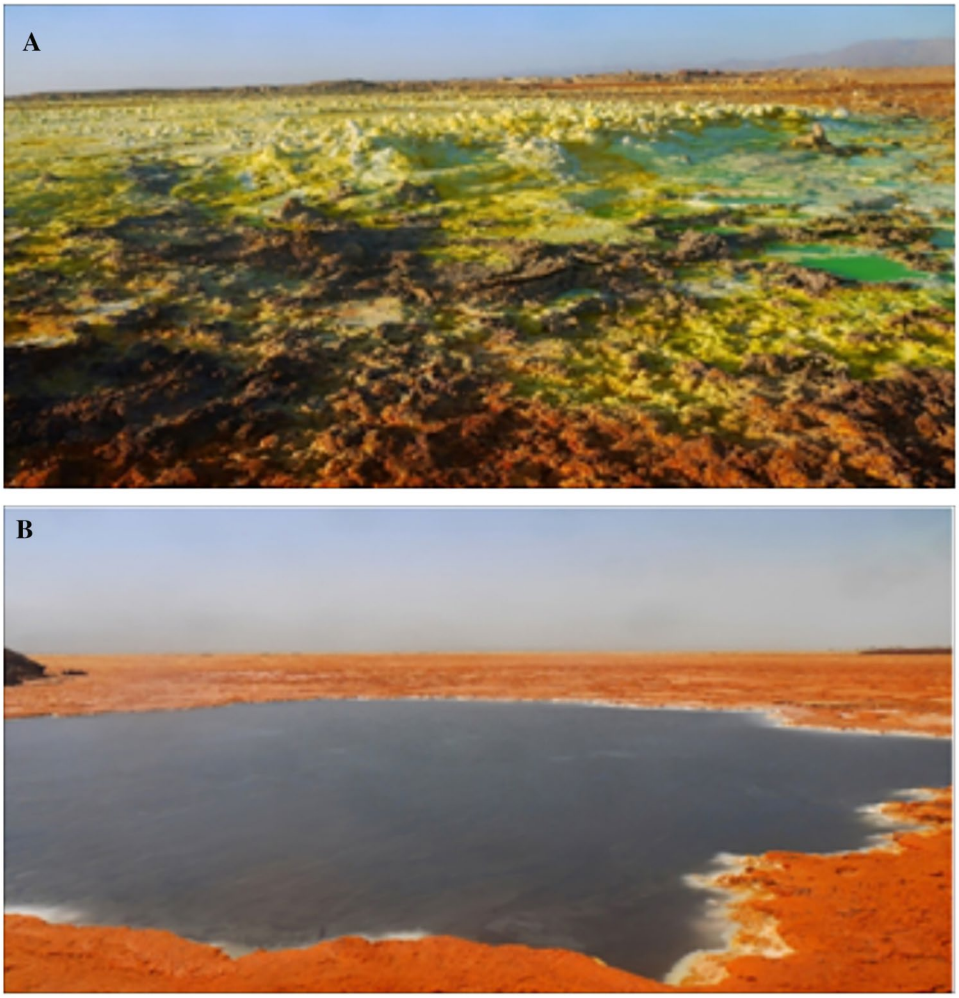

Fig. 2 Acidic pond on top of the Dallol crater (**a**); Black water pond (**b**). (Photo by Lulit, 2015)

DNA extractions

Environmental DNA (eDNA) was extracted at the Microbial Biotechnology Laboratory (Addis Ababa University) and at the PrIMO Laboratory (Brown University, Providence RI) using a modified CTAB method adapted from Zhou et al. (1996) and Mitchell and Takacs-Vesbach (2008). For Lake As'ale and Muda'ara pond, eDNA extraction was accomplished by directly following the optimized CTAB method from Zhou et al. (1996). The 1% CTAB-SDS DNA extraction method from Mitchell and Takacs-Vesbach (2008) was used for optimum eDNA extraction from Gaet'ale. This method was proved to be ideal for extracting eDNA from acidic hot springs at Yellowstone National Park, USA. Nonetheless for Dallol sample, immediate oxidation occurs upon addition of equal amount of phenol:chloroform and the color of the solution was observed to change into deep purple red. The results from NanoDrop spectrometer also confirmed high presence of contaminants where the 260/230 values were lower than commonly acceptable range of 2.0–2.2.

In order to check whether the method adapted from Mitchell and Takacs-Vesbach (2008) was optimally lysing the cells to release genomic materials, 1 ml of 'Dallol sample containing 1% CTAB-SDS buffer' was taken at the end of the last incubation period and was directly stained using FM™ 1–43 Dye (N-(3-Triethylammoniumpropyl)-4-(4-(Dibutylamino) Styryl) Pyridinium Dibromide),Invitrogen™ and Hoechst 33,342 Solution, Thermo Scientific™, in accordance to the manufacturer's instructions. According to the catalog supplied by the manufacturer, Hoechst 33,342 is a cell-permeable DNA stain used for specifically staining the nuclei of living or fixed cells, and that is excited by ultraviolet light and emits blue fluorescence at 460 to 490 nm. On the other hand, FM™ 1–43 Dye, Invitrogen™ is known to insert into the outer leaflet of the cell membrane and small secretory vesicles where it becomes intensely fluorescent. Cellular imaging was done using Zeiss Axioplan Fluorescence Microscope with Hamamatsu orca ER optical camera C4742-95 at PrIMO lab. Images of cells stained by both fluorescence dyes is shown in Supplementary Fig. 4 and Fig. 5.

Therefore, the optimized eDNA extraction methods employed for Dallol acid pond is as followed:

- 1% CTAB buffer solution was prepared (1% CTAB, 0.75 M NaCl, 50 mM Tris pH 8, 10 mM EDTA)
- The 0.22 µm GE® polycarbonate filter membranes with the non toxic Sucrose lysis buffer was transfered into 50 ml Sterile, certified RNase-/DNase-free and non-pyrogenic conical tubes (Corning™ Centrifuge Tubes, USA)
- 1% CTAB buffer solution was added in to the 50 ml tube and wash the filter membrane by vortexing
- 2 ml cell containing buffer solution was distributed into a 15 ml Sterile, certified RNase-/DNase-free and non-pyrogenic conical tubes (Corning™ Centrifuge Tubes, USA) and place them in Bioruptor® Sonication System (a cycling parameter of 30 s on and 30 s off for 10 min)
- Proteinase K (10 mg/ml) final concentration was added and incubate for 2 h at 60 °C
- 20% SDS (final concentration 2%) was added and incubate for 1–2 h at 60 °C and then centrifuged at 6000×*g* for 10 min
- Top phase of the supernatant was transferred into a 2 ml micro centrifuge tube then equal amount phenol:chloroform was added and centrifuged at 4000×*g* for10 min
- Top phase of the supernatant was transferred into a new 2 ml micro centrifuge tube equal amount chloroform was added and centrifuged at 4000×*g* for10 min
- Top phase of the supernatant was transferred into a new micro centrifuge tube then 0.6 volume of Isopropanol was added and placed at 4 °C overnight then centrifuged 16,000×*g* for 10 min at 4 °C
- Wash the precipitated pellet with 70% ethanol and resuspend with TE buffer
- DNA Clean & Concentrator™-5 (Zymo Research) was used to purify and concentrate the DNA per the manufacturer's instructions.

Since the amount of brine sample filtered from Black water pond was small due to high viscosity, the optimized extraction method for Dallol sample was adopted for extracting eDNA from Black Water. Considering the extremity of the environments where the samples were taken, intensive optimization for extracting, cleaning, and concentrating of DNA was needed. All eDNA extractions were performed in triplets and the extracted DNA from each buffer was later pooled prior to metagenome sequencing. The quantity and quality of eDNA from all brine samples were checked using PicoGreen Assay and Thermo Scientific NanoDrop 3300 Fluorospectrometer.

Table 1 Geohydro-chemical and stable water isotope measurements of the sample sites

Sample sites	Sample code	GPS Location		Altitude (meter)	Liquid color/ texture	Average pH	Average Salinity	Mg^{+2} (g/l)	NO_3^- (g/l)	NO_2^- (g/l)	TP (g/l)	COD	SO_4^{-2} (g/l)	Cl^- (g/l)	δ^{18} O (/ml)	δ^2H (/ml)
		Easting	**Northing**													
Dallol	DAL	0639983E	1574575 N	− 95	Light green/ oily	< 1	34.4%	3.7	1.94	0.76	1.2	12,150	5.5	234.3	8.56	-5.89
Black Water	BLA			− 114	Dusky/ oily	< 1	60.0%	300.4	0.2	0	0.13	14,150	0	311.1	-6.22	45.17

Metagenomics and sequence processing

Fragmentation of isolated DNA was performed using a Covaris™ S220 Ultra-Sonicator and fragments sizes were checked using the Fragment Analyzer™ Automated CE System and DNF-486–33 High Sensitivity NGS Fragment Analysis Kit (35 bp–6000 bp) at the Genomics Core facility (Brown University, Providence RI; Supplementary figures Fig. S4, Fig. S5 and Fig. S6). Four dual indexed libraries were prepared using TruSeq NANO DNA LT library prep kit (Illumina, San Diego, CA, USA) per the instruction manual. Libraries were barcoded, and combined into a single group and sequenced on one lane of a flow cell using a 100 bp paired-end run on a HiSeq 2500 instrument (Illumina) at the Genomics Core Facility (Brown University, Providence RI).

Reads from DNA sequencing were demultiplexed using Cassava v.2.0 and the detected barcodes in the report were well balanced. Sequence composition of the raw data was identified using FastQC and Trimmomatic was used to remove any adapter contamination. The software settings for Trimmomatic (Bolger et al. 2014) were ‘ILLUMINACLIP:adapters.fa:2:30:10′ to remove the TruSeq adapter sequences and ‘LEADING:3 TRAILING:3 SLIDINGWINDOW:4:10′ to improve mean sequence quality by trimming leading and trailing bases with $Q < 3$ and any sequences with a four-base sliding window mean below Q20. The sequences were then aligned to the human (GRCh37) and the phage PhiX genome (a standard for Illumina sequencing) to find and filter any non-microbial contamination.

Taxonomic assignment of non-assembled reads

For non-assembly approach, the direct quality-checked sequence reads of the samples from the two runs were merged and put into MetaPhlAn 2 bioinformatics tool which maps direct sequence reads against a reduced set of clade-specific marker sequences (Segata 2012; Truong 2015; Waterhouse 2018). Microbial relative abundance profiles were generated with MetaPhlAn2 (Segata 2012) using default parameters and bowtie2 alignment. The MetaPhlAn2 reference database consisted of clade-specific marker genes from ~ 17 000 reference genomes (79% bacteria/archea, 20.4% viral and 0.6% eukaryotic). The marker genes in the database are genes that are unambiguously characterized in a taxonomic clade as they are always present in the sequenced isolates of that clade and never present in any other sequenced organism. Operational Taxonomic Unit (OTU) profiles from the two samples were merged with the script ‘merge_metaphlan_tables.py’ included with the MetaPhlAn2. Distribution and heatmaps were generated with ‘metaphlan_hclust_heatmap.py’ script using default options and the ‘-d braycurtis’, ‘-minv 0.01′ flags.

Due to the extensively conducted optimization procedures of DNA extraction, the microbial relative abundance profile table resulted from MetaPhlAn2 was converted to a ‘Binary or presence-absence’ data to analyze with optimal degree of precision in these two particular ecological phenomena (Legendre and Legendre 1998). Hence, the dissimilarity index was calculated Bray Curtis dissimilarity using vegdist function in vegan and package on R Studio. For calculating Bray–Curtis dissimilarity, the default for the function “vegdist” was changed to “binary = TRUE” (i.e. $(b + c)/(2a + b + c)$, where a denoted the number of species shared between two sites while b and c indicated unique species). In binary terms, the shared component is number of shared species, and totals are numbers of species on sites. Plotting of result to dissimilarity was done using the function “ordiplot”.

Taxonomic and functional gene assignment of assembled contigs

Quality sequence reads were assembled using metaSPADes with a flag ‘meta’ and kmers 21, 33 and 55 (Nurk 2017). The resulting metagenome contigs were aligned against NCBI non-redundant protein database using Double Index Alignment of Next Generation Data (DIAMOND) v0.9.24; BlaSTx with the sensitive mode, frameshift alignment for longer sequences and a default e-value cut-off of 0.001 (Buchfink et al. 2015). The taxonomic assignment was performed using MEtaGenome analyzer 6 Community Edition (MEGAN6 CE) (Bağcı 2019; Huson 2018; Huson, et al. 2016). One contig may contain several protein coding genes (ORFs). Hence, during the filtration process, each gene was considered separately, where alignments that overlap significantly were grouped into segments, denoting different genes. The top segments within the best scoring alignments were taken into account. Therefore, MEGAN6 CE program placed the annotated reads onto the NCBI taxonomy tree using settings of Lower Common Ancestor (LCA) algorithm for long read adjusted as follows: min score- 100.0; max expected- 0.01; min percent identity- 50; top percent- 10 and LCA coverage 80%. The min support percent was adjusted to 0.02 so that taxa that obtained at least 0.02% of all aligned is reported. This adjustment will increase the ‘level of detection” and improve sensitivity for low-abundance species. After the initial automatic binning step, additional manual inspection was performed. Contigs with uncertain taxonomic association, characterized by mixed Blastx hits were moved to the ‘Unassigned’ bin.

Taxonomy comparison was computed based on the previously set algorithm used to annotate reads onto the NCBI taxonomy tree on MEGAN 6 CE. In addition, MEGAN6 CE was used to map the RefSeq ids to SEED functional roles using the ‘seed2ncbi.gz’ file from the SEED server (Bağcı

2019; Huson 2018; Huson et al. 2016). Functional gene evaluation and protein identification for nutrient cycles focusing on nitrogen metabolism was performed using MEGAN-SEED. In order to identify possible pathways of nitrogen utilization and flux, the Enzyme Commission (EC) number, or the gene sequences for key proteins of nitrogen cycle were retrieved from SEED and converted to KEGG Orthology identifier (KO). The website available KEGG mapping tool (Kanehisa and Sato 2020) was used for automatic assignment of KO identifiers to KEGG molecular networks and KEGG pathway maps of nitrogen metabolism for the studied sample sites.

Homology modeling of predicted nitrogen fixing (Nif) proteins

MEGAN's-Alignment was used to compute and view multiple sequence alignment of all contigs that have significant matches to reference sequences associated with nitrogen fixation from SEED's nitrogen metabolism class. Consensus translated ORFs mapped to respective reference protein sequences were exported in fasta format for further in silico analyses. Primary sequence analysis was performed on the consensus protein sequences using ProtParam tool available at (www.expasy.org/tools/protparam) (Gasteiger et al. 2005). Sequence similarity of translated ORFs grouped as a nitrogen fixing class was analyzed using NCBI Blastp (cut-off value 1e-10) taking as reference Nif protein sequences well characterized from different families. Query translated ORFs with more than 95% coverage of residues of the complete homologous reference's protein sequence were considered as near complete sequences for this study. Both complete translated ORFs and near complete ORFs were selected for further structural analyses. Accession numbers of the reference proteins are showed in Supplementary Table 1.

Prediction of protein pattern, structural domain, active site and their related function of the selected translated ORFs were performed using InterProscan at (http://wwwdev.ebi.ac.uk/interpro/). In addition, secondary structures were predicted using GOR4, SOPMA available at ExpaSy (http://www.expasy.org/tools) using default settings and in accordance with the prediction methods (Geourjon and Deléage 1995; Garnier et al. 1996). Results of these predictions were compiled and compared (Table 2).

3-D modeling was performed only for predicted proteins common in both sample sites that fulfill the above criteria. Hence, common query translated complete or near complete ORFs to Dal and Bla were submitted to Swiss Model (http://swissmodel.expasy.org) under automatic mode settings (Waterhouse 2018). The server automatically selects the template from experimentally solved protein structures thereby generating the best possible model according to homology modeling method. The predicted 3-D models were then accessed in the form of PDB files (Supplementary PDB file).

Table 2 Overview of metagenomics

	DAL	BLA
Library insert average	~500 bp	~500 bp
Final DNA concentration for sequencing	NA	NA
Length of single read	100 bp	100 bp
Total number of reads	3,480,089	6,147,722
GC content (%)	54	58

Results

Geochemical and physicochemical properties

The Dallol (Dal) and Black Water (Bla) ponds are categorized as poly-extreme environments with hyper saline and hyper-acidic conditions (Table 1). The salinity of Black Water and Dallol ponds were 60% and 34% respectively, with pH values less than 1 for both sample sites.

The two ponds showed distinct variations in the measured ion contents (Table 1). The amount of Mg^{2+} ion in the Bla was by far the highest (81 times more than Dal). The presence of large amounts of oxidizable organic material in the acid ponds (12,150 and 14,150 for Dal and Bla respectively) was indicated by the results of the COD measurements. Sufficient amounts of NO_2^- and NO_3^- were measured for Dal (1.94 g/l and 0.76 g/l respectively) but the amount of NO_2^- in Bla was zero even if the quantity of NO_3^- was significant (0.2 g/l). The total phosphorous (TP) recorded in the ponds was very high (0.1–1.2 g/l) while SO_4^{2-} was only measurable in the Dal (5.5 g/l). The oxygen and hydrogen isotope analysis showed that the Bla is characterized by positive δ^2H (45.17/ml) but depleted in $\delta^{18}O$ (-6.22/ml). On the other hand, Dal is characterized by strongly enriched $\delta^{18}O$ (8.56/ml) and slightly depleted δ^2H (-5.89/ml), suggesting a strong hydrothermal input to the brines.

Metagenomic sequence analysis and taxonomic profiling of assembled and non-assembled reads

Quantification of double stranded DNA from Dal and Bla samples using the PicoGreen Assay and Thermo Scientific NanoDrop 3300 Fluorospectrometer was possible only with the Bla samples. This is because, in the case of Dallol sample, there was an anomalous compound or contaminant associated with the double stranded DNA that interfered with the readings (Supplementary Fig. S6). Yet, after sequencing, a total of 9.6 million reads with average GC content of 56% were generated (Table 3). The read quality was average in the leading and trailing ends of the reads, and very good at all positions within the reads. The overall sequence result was enough to analyze existing biota in detail (Table 3). After removing the adaptor sequences and checking the quality, all reads were 100 base pair (bp) long. When MetaPhlAn2 was used to perform taxonomic profiling from shotgun sequences, significant differences in OTU composition were identified between the two sites. Based on the result of the calculated binary dissimilarity index (i.e. 0.53), the two sample sites contain unique OTUs and shared only 20 OTUs. More than 80% and 97% of the OTUs profiled are grouped under the domain Bacteria for Dal and Bla respectively, while small percentage of viral OTUs were encountered in both sites. As highlighted in the heat map generated (Fig. 3), the *Paraburkholderia_fungorum* from phylum Proteobacteria was seen with more abundance relative to other OTUs in both Dal and Bla.

Individual metagenome assemblies were generated from quality-trimmed metagenome reads using MetaSpades where a total of 85,336 and 109,720 contigs with weighted average length of 508 and 1126 bp were obtained for Dal and Bla respectively (Table 4). DIAMOND alignment against the NCBI-nr database (downloaded September 2019) resulted in protein alignments of 68,991 contigs (27.3 Mb) for Dal and 80,595 contigs (45.4 Mb) for Bla (Table 4). More than 99.7% of the total assigned reads were binned to *Bacteria* in both Bla and Dal (Table 4). Of these, more than 16 Mb and 1 Mb were assigned to the three classes of phylum Proteobacteria (Alphaproteobacteria, Betaproteobacteria and Gammaproteobacteria) and Actinobacteria respectively in both Dal and Bla (Fig. 4). Like the result generated by MetaPhlAn2, *Paraburkholderia_fungorum, Escherichia_coli* and *_Bradyrhizobium_sp_DFCI* are among the top 10 OTUs with high number of aligned bases at species rank level based on MEGAN 6CE (Supplementary Table 2).

Functional placement of assembled contig reads

Only 26% and 28% the total assigned reads were functionally annotated in Dal and Bla respectively and of these functionally assigned reads, only 17% are binned to SEED subsystems of nutrient metabolisms (carbohydrate, nitrogen, phosphorus, potassium and sulfur) (Table 2). The highest number of reads was assigned to carbohydrate metabolism (11.8% for Dal and 11.2% for Bla) and the least were assigned to phosphorus and potassium metabolisms to both sample sites (only 1% each) (Table 5). For nitrogen and sulfur metabolisms, less than 2% the total reads were parsed for both sample sites (Table 2). In both sample sites, more than 80% of the aligned bases binned to the subsystems of the five nutrient metabolisms are from the taxonomic classes Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria and Actinobacteria (Supplementary Table 5). Contigs parsed under *Rhizobiales*, *Burholderiales* and *Propionibacteriales* are the three primary sources of Open reading frames for prediction protein families in the carbohydrate, nitrogen, phosphorus, potassium and sulfur metabolism subsystems in both Dal and Bla (Table 6). Generally, predicted protein families in Carbohydrate metabolism, in nitrogen metabolism, in sulfur metabolism, in phosphorus metabolism and in potassium metabolism are listed in Supplementary Tables 6 and 7.

Furthermore, MEGAN-SEED was used to scan the whole collection of contigs and to find ORFs grouped under nitrogen metabolism. Complete or partial translated ORFs, with different percentages of identity to homologous proteins involved in nitrogen metabolism were predicted. MEGAN-Aligner also helped in aligning and connecting two or more contigs to predict complete or near complete ORFs involved in nitrogen metabolism. Relatively numerous contigs with predicted genes were binned in the classes of ammonia assimilation and nitrate /nitrite ammonification. Several contigs were also grouped under the nitrogen fixing class in both Dal and Bla of nitrogen metabolism.

Functional orthology was searched and matched for 92% and 81% predicted ORFs involved in nitrogen metabolism for

Table 3 Summerized MetaPhlAn 2 microbial relative abundance profile at levels of Domain and Phylum

	Dal (LTW0001)	Bla (LTW0003)
k__Archaea	**0.04398**	**0**
k__Archaea|p__Euryarchaeota	0.04398	0
k__Bacteria	**82.0232**	**97.51716**
k__Bacteria|p__Actinobacteria	15.22829	2.59224
k__Bacteria|p__Bacteroidetes	1.19922	0.54672
k__Bacteria|p__Deinococcus_Thermus	0.03643	2.70664
k__Bacteria|p__Firmicutes	2.76014	3.30194
k__Bacteria|p__Proteobacteria	62.79912	88.36962
k__Viruses	**17.93282**	**2.48284**
k__Viruses|p__Viruses_noname	17.93282	2.48284

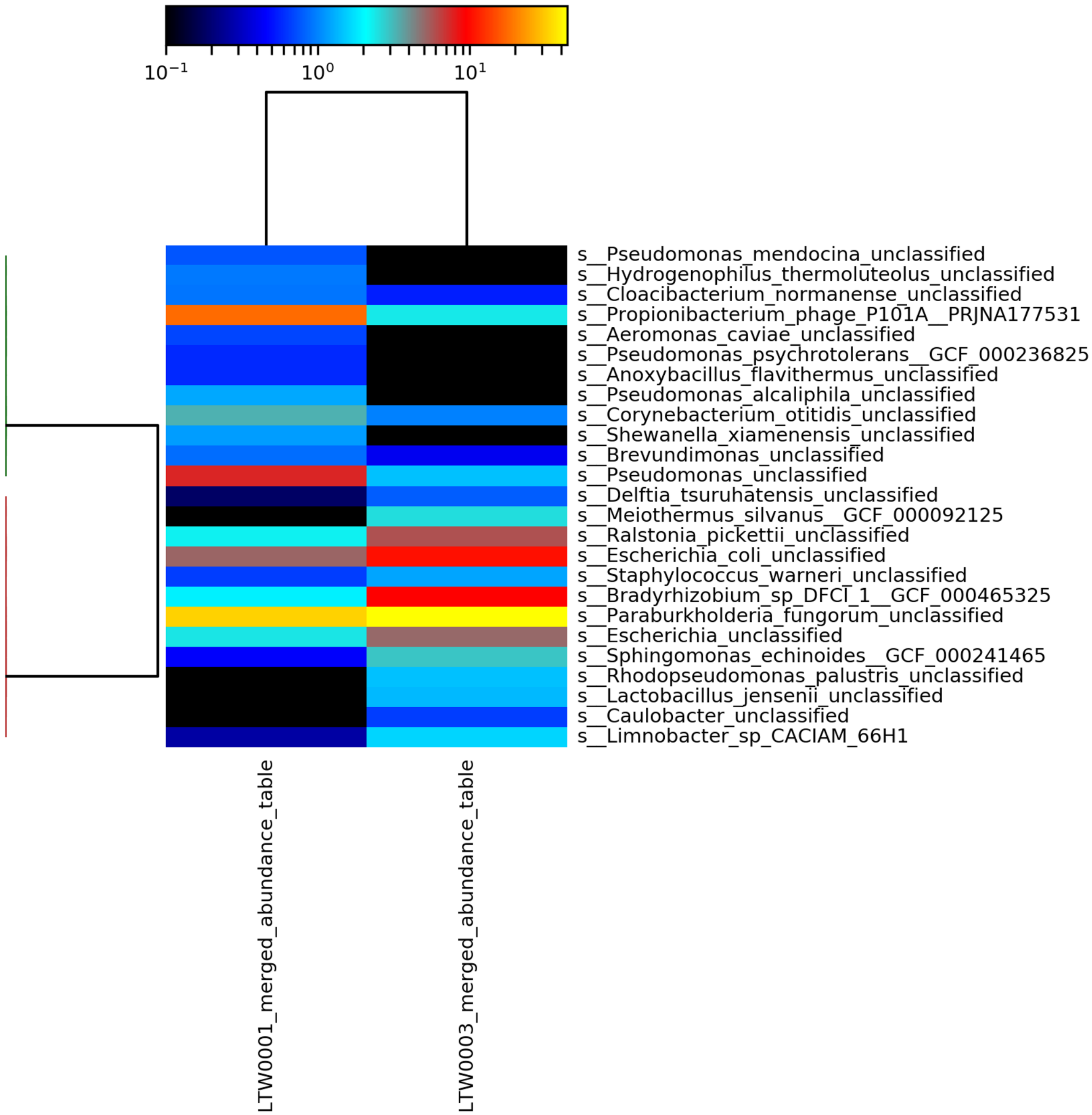

Fig. 3 Heat Map of top 25 abundant OTUs generated by MetPhlAn2

Dal and Bla respectively. According to the metabolic pathway mapping using the KEGG database, all the predicted genes are directly or indirectly involved in the nitrate ammonification process in both Dal and Bla (Figs. 4, 5). Net nitrification was absent in the two acid ponds. All important genes for a complete denitrification process were predicted in the Bla metagenome. As for the Dal metagenome, all genes for denitrification except the final NO reductase expression regulator (NosR) were predicted, which led to incomplete mapping of denitrification. In general, different types of enzymes, transporter proteins and transcription factors involving in nitrogen metabolism were all predicted from these data.

Prediction of 1°, 2° structure and 3-D modeling of selected nitrogen fixing proteins

From the Dal assembled metagenome, the translation products of 2 complete or near complete ORFs (homologous to

Table 4 Metagenome assemblies with MetaSPAdes and taxonomic assignment of contigs with DIAMOND/MEGAN against the NCBI *GenBank nr protein database*

	DAL (LTW0001)	BLA (LTW0003)
Metagenome size (bp/read)	38,505,818/ 3,480,089	77,553,274/ 6,147,722
No. of contigs	85,336	109,720
N50	508	1126
Max. contig length (bp)	48,588	67,786
Total number of contigs aligned by DIAMOND	68,991	80,595
Total number of aligned bases	27,258,425	45,365,792
Number of reads with no hits	721 (0.003%)	767 (0.002%)
Number of reads not assigned	6,580,771(24.1%)	10,350,335(22.8%)
Number of aligned bases in assigned contigs	20,676,933 (75.9%)	35,014,895 (77.2%)
Root	36,256 (0.18%)	57,626 (0.16%)
Bacteria	20,615,436 (99.7%)	34,938,932 (99.8%)
Archaea	6,792 (0.03%)	9,234 (0.03%)
Virus	18,449(0.09%)	9,102 (0.03%)

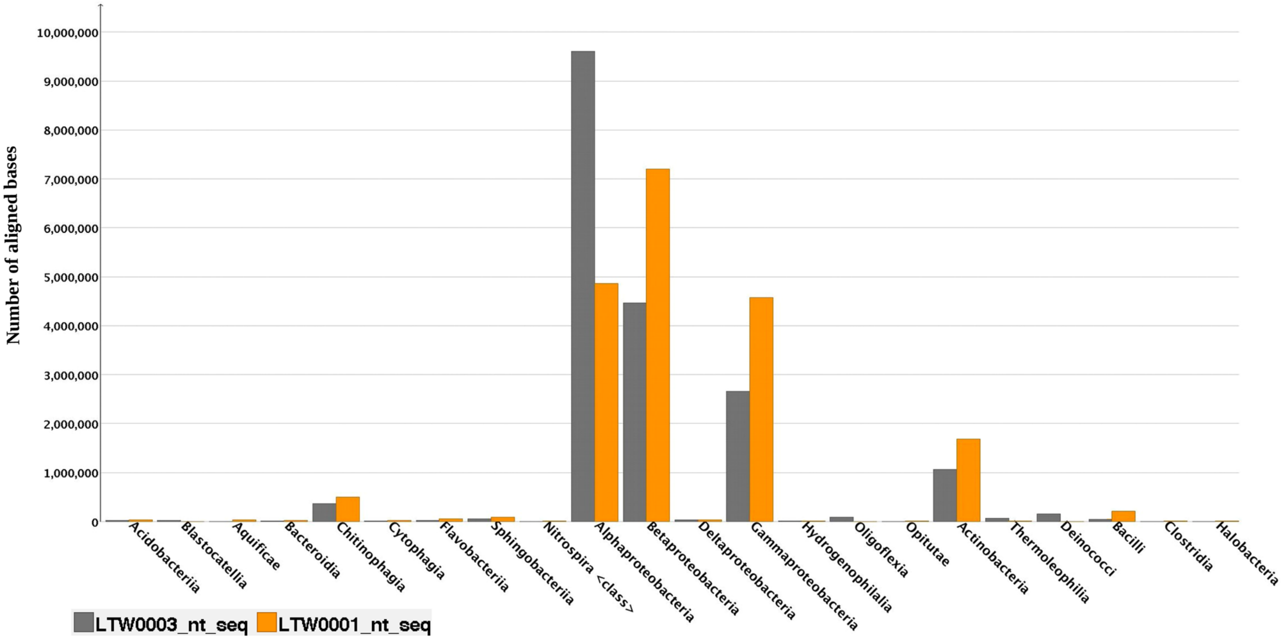

Fig. 4 Class level taxonomic classification of metagenome assembled reads from Dallol (LTW0001_nt_seq) and Black Water (LTW0003_nt_seq)

Table 5 Functional annotation of reads based on SEED database

	Dal (LTW0001)	Bla (LTW0003)
Total assigned reads using SEED	27,294,000 bp	45,418,728 bp
Total functionally annotated reads	7,087,814 bp	12,825,621 bp
Total # of predicted genes	2297	2568
Carbohydrate metabolism	836,574 bp	1.436,088 bp
Nitrogen metabolism	105,279 bp	214,768 bp
Phosphorus metabolism	76,818 bp	132,127 bp
Sulfur metabolism	112,683 bp	198,011 bp
Potassium metabolism	71,609 bp	136,990 bp

4Fe-4S ferredoxin and cysteine desulfurase) and 8 partial ORFs (homologous to nifK, nifH, nifB, nifX, nifZ, nifQ, nifV, nifE) were predicted. On the other hand, from Bla, translation products of 5 complete or near complete ORFs (homologous to 4Fe-4S ferredoxin and cysteine desulfurase, NifB, NifT and NifW) and 8 partial ORFs (homologous to NifD, NifK, NifA, NifH, NifX, NifE, NifN and NifV) were predicted. The results of prediction of the primary and secondary structures of the translated products of the complete or near complete ORFs are listed in Supplementary Table 3.

The amino acid identity of the selected protein sequences in the assembled metagenomes was checked against the

Table 6 Number of aligned bases assigned in kilo base (kb) to the five nutrient metabolism subsystems of the top five tax*onomic Orders*

	Nitrogen	Sulfur	Phosphorus	Carbohydrate	Potassium
Dallol (DAL)					
Rhizobiales	17.8 kb	31.6 kb	12.5 kb	143.7 kb	15.6 kb
Burholderiales	25.7 kb	26.5 kb	17.9 kb	192.8 kb	13.8 kb
Pseudomonadales	3.6 kb	6.5 kb	13.5 kb	41.8 kb	3.9 kb
Entrobacteriales	7.5 kb	6.4 kb	2.9 kb	55.0 kb	4.1 kb
Propionibacteriales	24.2 kb	13.9 kb	11.6 kb	163.9 kb	10.3 kb
Black Water (BLA)					
Rhizobiales	69.9 kb	78.9 kb	50.1 kb	458.8 kb	59.0 kb
Burholderiales	28.8 kb	31.5 kb	18 kb	198.1 kb	15.9 kb
Pseudomonadales	0.2 kb	0.2 kb	0.2 kb	1.0 kb	0
Entrobacteriales	25.1 kb	12.0 kb	7.0 kb	156.0 kb	14.1 kb
Propionibacteriales	20.6 kb	16.1 kb	15.0 kb	181.0 kb	12.0 kb

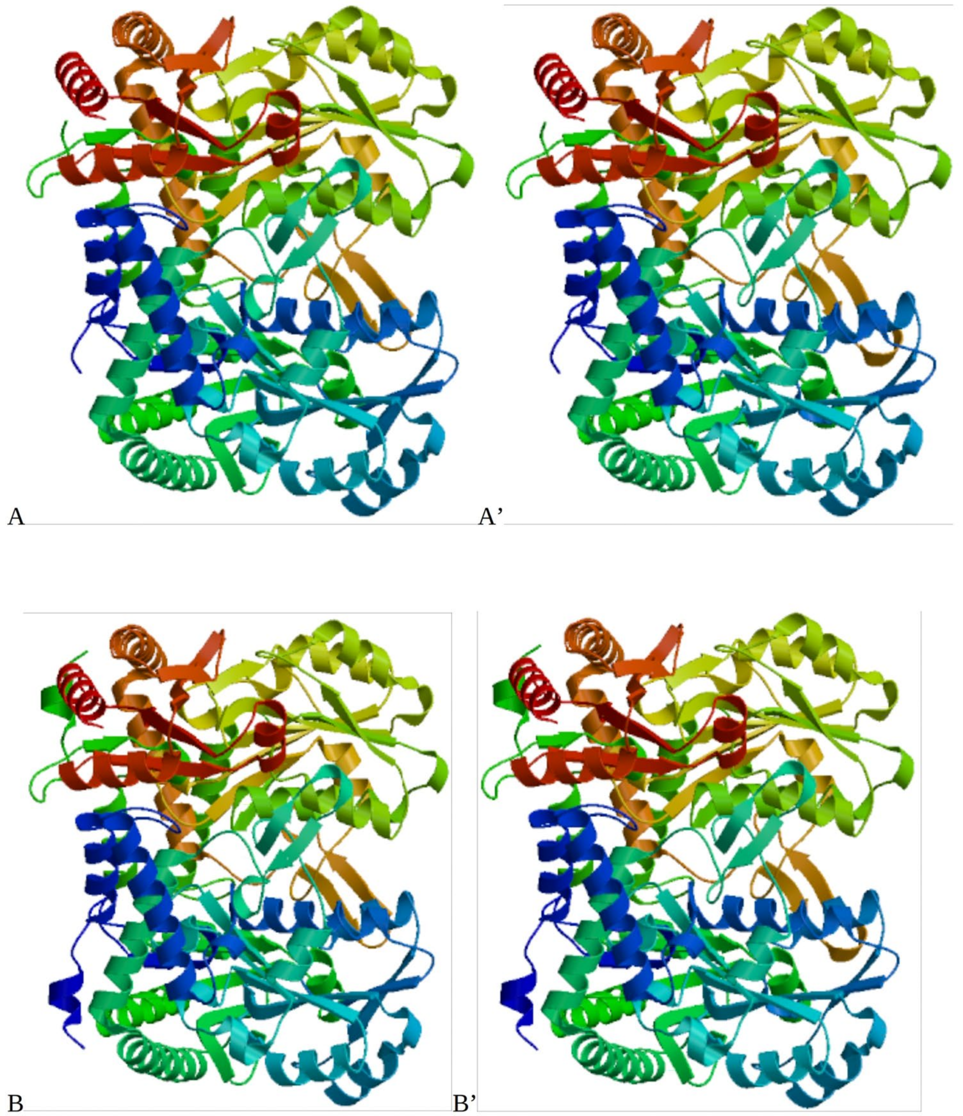

Fig. 5 3-D modeling of cysteine desulfurase from Dallol (**A**) with reference protein WP_149425408.1 (**A′**) and cysteine desulfurase from Black Water (**B**) with reference homologous protein WP_149425408.1 (**B′**)

reference proteins. The maximum percentage of similarity obtained was for the nitrogen fixing protein NifT from Bla at 91% identity to *Pseudolabrys sp. FHR47*. The rest of selected predicted protein sequences have less than 80 percent identity to their respective reference proteins (Supplementary Table 1). Sequence identity searches using blastp and primary sequence analysis performed using ProtParam were found to be contigs encoding proteins involved in nitrogen fixation. Their respective reference proteins are found in Supplementary Table 1.

The primary sequence analyses for all predicted proteins in both Bla and Dal showed a difference in composition of amino acids to their respective reference proteins. The number of negatively charged amino acids was predicted to be higher than the number of positively charged amino acids in all selected proteins as in the case of their respective reference proteins. The results of secondary structure predictions obtained from the two servers were different for all query protein sequences (Table 7). While Gor4 was only able to predict the three basic secondary structural elements (helix, sheet and coil), SOPMA further predicted beta turns and the results are shown in both numeric and graphical forms. The difference among the secondary structure prediction results produced by the respective servers could probably result from the use of different indices in making the prediction by the servers.

The 3-D models of the two commonly found query proteins and their respective references are shown (Figs. 6, 7). According to prediction outputs, the Swiss Modeler used cysteine desulfurase from *Legionella pneumophila Philadelphia 1* and 2[4FE-4S] ferredoxin from *Pseudomonas aeruginosa* as templates (Table 8). In all the models predicted, the structures are in agreement with the results of the secondary structure earlier predicted for the query as well as the reference sequences. Alpha helix and random coils dominated the structure of the predicted cysteine desulfurase while random coils and extended sheets dominated the structure of the predicted [4Fe-4S] ferredoxin. Functional motifs within the cysteine desulfurase and NifB protein families were identified. The Pyridoxal 5′-phosphate binding motif within the cysteine desulfurase and FeS/SAM binding motif with in FeMo cofactor biosynthesis protein (nifB-like) were determined. The predicted positions of conserved and active sites of the query proteins with the corresponding amino acid residues are listed in Table 5.

Discussion

Cycles of environmental and geological changes in natural systems over different periods of time and amplitudes affect the microbial composition, genetic repertoire and activity of microbial community members, and community function

Table 7 Comparison of predicted secondary structures of query proteins versus reference proteins using two different servers

Sample site	Contig's node number	Amino acid length	Predicted protein	GOR4				SOPMA			
				α helix	Extended sheet	β turn	Random coil	α helix	Extended sheet	β turn	Random coil
BLA	Node_4344	551	Cysteine desulfurase	38.48%	10.53%	0.00%	51.00%	34.30%	13.79%	8.17%	43.74%
Reference [a]	WP_149425408176_609	534	Cysteine desulfurase	41.01%	9.18%	0.00%	49.81%	33.90%	12.92%	6.93%	46.25%
DAL	Nodes_5315_35911	393	Cysteine desulfurase	44.78%	13.74%	0.00%	41.48%	40.97%	17.30%	8.91%	32.82%
Reference [b]	WP_149425408176_609	393	Cysteine desulfurase	46.56%	10.69%	0.00%	42.75%	42.24%	16.79%	8.14%	32.82%
BLA	Node_27591	72	4Fe-4S Ferredoxin	0.00%	19.44%	0.00%	80.56%	2.78%	29.17%	6.94%	61.11%
DAL	Node_15557	72	4Fe-4S Ferredoxin	0.00%	19.44%	0.00%	80.56%	2.78%	29.17%	6.94%	61.11%
Reference [c]	WP_063195668.1:	72	4Fe-4S Ferredoxin	0.00%	27.78%	0.00%	72.22%	1.39%	26.39%	5.56%	66.67%
BLA	Node_7493_18684	457	NifB	36.54%	16.63%	0.00%	46.83%	40.70%	16.41%	7.22%	35.67%
Reference [d]	WP_091676580	457	NifB	31.29%	19.04%	0.00%	49.67%	43.11%	14.88%	7.00%	35.01%
BLA	Node_62099	69	NifT	14.49%	31.88%	0.00%	53.62%	11.59%	31.88%	13.04%	43.48%
Reference [e]	WP_137042982	69	NifT	14.49%	31.88%	0.00%	53.62%	8.70%	36.23%	7.25%	47.83%
BLA	Node_21959	109	NifW	57.80%	5.50%	0.00%	36.70%	62.39%	7.34%	1.83%	28.44%
Reference [f]	WP_016918478.1	109	NifW	53.21%	5.50%	0.00%	41.28%	66.06%	6.42%	3.67%	23.85%

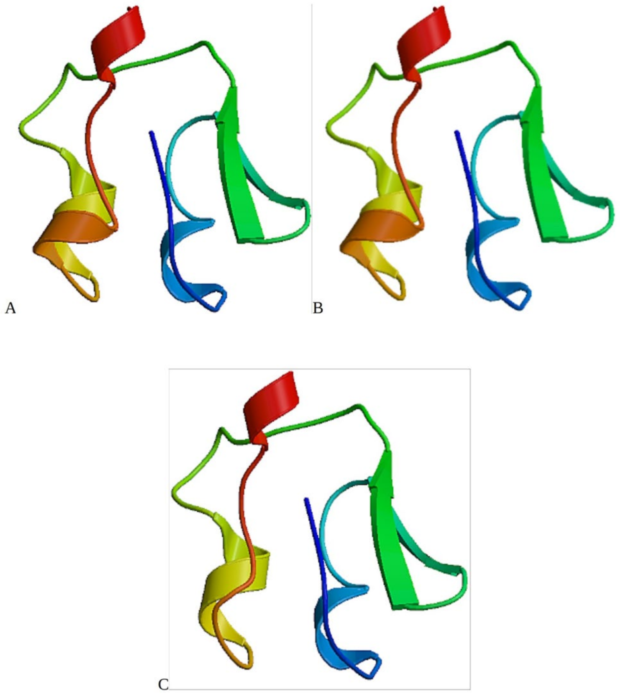

Fig. 6 3-D modeling of [4Fe-4S] ferredoxin from Dallol (**a**), Black Water (**b**) and reference homologous protein WP_0631956681172 (**c**)

as a whole (Ghebrezgabher et al. 2016; Klotz 2016). The year 2015 was special for investigating the acid ponds of Dallol and Black Water because for the first time since the early 1900's (Waterhouse 2018) a phreatic eruption of the Dallol summit was observed (Asrat 2016). Though recent magmatic eruptions to the surface of "Dallol mountain" have not been recorded, numerous phreatic eruptions including the 1926 eruption, which formed a 30 m wide crater at the top of the mound, the 2014 phreatic eruption of bischofite (hydrous magnesium chloride) rich hot brine, just southwest of the Dallol summit, as well as the minor phreatic eruption of January 2015 at the summit crater, are well recorded (Asrat 2016). These phreatic eruptions, particularly the most recent eruption of 2015, led to the drying up of many brine/acid ponds and the appearance of a few new ones at the summit crater (Asrat 2016).

The acid ponds of Danakil depression, whose size and volume vary with the periodic phreatic eruptions, are composed of fluids derived mainly from a mixture of groundwater and hydrothermal-geothermal up-flow (Franzson et al. 2015; Asrat 2016; Nobile, et al. 2004). At the time of sampling in 2015, the isotopes δ^{18}O and sδ^{2}H measured showed that the sources of water for Dallol and Black Water were different and the Dallol brines had strong volcanic signature while Black Water was fed by non-volcanic, dominantly rainfall originated groundwater. In addition, the two acid ponds were different in regard to the concentration of ions present. For example, the concentration of Mg^{2+} in Black Water was the highest ever recorded (~ 12 M) with an almost 1:1 ratio to Cl^{-}, which could be considered as inhibitory to many cellular systems (Hallsworth 2007). This anomalous concentration of Mg^{2+} in the Black Water pond is possibly related to the bischofite (hydrous MgCl$_2$) eruption in close proximity to the Black Water pond in 2014. Such super saturation of salts can only be achieved under high temperature (> 560 °C) and very acidic conditions (Herbstein et al. 1982; HCl Leaching and Acid Regeneration 2011), which are the main characteristics of the two acid ponds. Furthermore, the two acid ponds were also rich in oxidizable organic material (COD), that can be attributed to the extreme concentration of hydrothermal gases and fluids such as water vapor, CO$_2$, CH$_4$, H$_2$S, ethane, benzene and so on. These compounds are typical to solfatara floors, geothermal hot springs and soils after condensation of steam at shallow depth subsequent to phreatic eruption (Mayer 2017; Crognale, et al. 2018).

The difficulty of studying extremophiles using culture-based methods has been well recognized. Though molecular methods have shown supremacy over culture-based diversity study of extremophiles, DNA extraction techniques and sequencing strategy play significant role in defining environmental microbial communities (Roux 2011; Poretsky 2014; Ranjan 2016; Bag 2016; Ketchum et al. 2018). The major challenge of metagenome study is optimizing proper lysis of heterogeneous community microbial cells without damaging their genomes. Combinations of physical, chemical, and mechanical methods for proper lysis of microbial inhabitants and extraction of community DNA from different environments have been proven suitable for in-depth analysis (Bag 2016). In this study, combined physical and chemical lysis techniques for heterogeneous microbial community cells and whole community metagenome sequencing were effectively used to capture taxa that are traditionally difficult to lyse and excessively rare to detect respectively.

Recently published studies regarding life in the acid ponds and hot springs of Danakil depression based on 16S rDNA investigation reported contrasting results (Gómez 2019; Belilla 2019). While Gómez et al. (2019) confirmed the presence of members of the *Nanohaloarchaea* group from the salt precipitates at Dallol hydrothermal fluid source, Belilla et al. (2019) strongly suggested absence of active life forms in Dallol and Black Water acid ponds. However, Carrizo et al. (2019) explored signs of life in Dallol sulfur springs by studying the patterns of lipid biomarkers and stable isotope composition and suggested the presence of microbial community largely composed of thermophilic members of the Aquificae, Thermotogae, Chloroflexi and Proteobacteria phyla. Both non-assembled and assembled metagenome sequence analysis methods applied (MetaPhlAn2 and MEGAN 6 CE) showed that unclassified *Paraburkholderia fungorum,* unclassified *Escherichia_coli_*and

Table 8 Predicted protein motifs, patterns and profiles (InterProScan searches)

Sample site	Contig's node number & protein coding region	Overlapping protein entries & regions	Biological function	Binding sites[a] / molecular function	
BLA	Node_4344 (147–547)	IPR015424 PyrdxlP-dep_Trfase: 4–403 IIPR000192 Aminotransferase class V: 26–395 IPR015422 PyrdxlP-dep_Trfase_dom1: 9–398 IPR015421 PyrdxlP-dep_Trfase_major: 37–297 IPR010970 Cys_dSase_SufS: 8–403	Cysteine metabolic process (GO:0,006,534)	K [227] /catalytic residue (GO:0,003,824) A,T[95–96], I[99], H[124], D[201], A,Q[203–204], S[224], H,K[226–227]/ Pyridoxal 5′-phosphate binding pocket (GO:0,030,170)	
DAL	Nodes_5315_35911(1–393)	IIPR000192 Aminotransferase class V: 17–386 IPR015421 PyrdxlP-dep_Trfase_major: 28–288 IPR010970 Cys_dSase_SufS: 1–392 IPR015422 PyrdxlP-dep_Trfase_dom1: 1–389 IPR015424 PyrdxlP-dep_Trfase: 1–392	Cysteine metabolic process (GO:0,006,534)	K[218]/catalytic residue (GO:0,003,824) A,T(Kiyasu 2000; Egener, et al. 2001), I[90], H[115], D[192], A,Q [194–195], S[215], H,K[217–218]]/ Pyridoxal 5′-phosphate binding pocket (GO:0,030,170)	
BLA	Node_27591(1–72)	SSF54862 4Fe-4S Ferredoxins: 4Fe-4S_Fe-S-bd	NA	NA	NA
DAL	Node_15557 (1–72)	SSF54862 4Fe-4S Ferredoxins: 4Fe-4S_Fe-S-bd	NA	NA	NA
BLA	Node_7493_18684 (35–434)	IPR005980 Nitrogenase cofactor biosynthesis protein NifB: rSAM	coenzyme biosynthetic process (GO:0,009,108)	C[67], I[69], C(Ivleva 2016), Y,C,N[73 – 75], L[115], G,D[121 – 122], S,P,H[149 – 151], T[173], V[228], N,I[254 – 255]	Catalytic activity (GO:0,003,824), iron-sulfur cluster binding (GO:0,051,536), metal ion binding (GO:0,046,872),4 iron, 4 sulfur cluster binding (GO:0,051,539)
BLA	Node_62099 (1–69)	NifT/FixU (IPR009727) NA	Nitrogen fixation (GO:0,009,399)	NA	NA
BLA	Node_21959 (1–109)	Nitrogen fixation protein NifW (IPR004893) NA	nitrogen fixation (GO:0,009,399)	NA	NA

[a]Positions of binding sites on assembled contigs with the corresponding residues

Bradyrhizobium sp. DFCI-1 are among the top 10 abundant OTUs at species rank level commonly found in Dal and Bla (Supplementary Tables 2-9). Furthermore, assembled sequences from Dal and Bla were parsed to Aquificae and Chloroflexi phyla respectively as predicted by Carrizo et al. (2019) but not to Thermotogae. Large number of assembled contigs was also binned to the phylum Proteobacteria, similar to most extreme acidic environments studied so far (Spear et al. 2006; Johnson 2001; Méndez-García 2015; Mesa 2017; Schuler et al. 2017). Based on the metagenome data, Alphaproteobacteria was signified in Black Water while Betaproteobacteria and Gammaproteobacteria were dominantly presented in Dallol.

The best option to improve our understanding on microbial ecology and metabolism is to justify the functional potential of the extremophiles through metagenomic sequence data since the advantages of whole genome shotgun sequencing are to increase detection of microbial diversity and prediction of genes (Schmeisser et al. 2007; Su 2012; Ranjan 2016; Abreu 2018). Understanding the stress response of microbially mediated nutrient cycling processes in hot springs and volcanic pools through analysis of pathways of formation of complex polymers from simpler elements or monomers is important to search for life in different places in the Solar System as well as to develop better mechanisms to adapt to existing changes in the intensity and frequency of global climate extremes (Li et al. 2018). They performed in silico investigation of gene functions on 30% of the total metagenome, and indicated the four important biogeochemical cycles for sustaining life in the acid ponds of DAL and BLA. As carbon is the main constituent of living organisms and the essential component for all organic polymers, large number of contigs with translated ORFs encoding enzymes, transcription factors and different proteins for the carbon metabolism were predicted. Furthermore, genes important for converting different forms of biologically important nutrients such as phosphorous, potassium, sulfur and nitrogen were identified (Supplementary Tables 6-7).

The nitrogen cycle is the second most important nutrient cycle to organisms because it provides the building blocks of DNA and proteins (Widdison and Burt 2008) and the contigs with protein coding genes for nitrogen metabolism were predicted in both the Dallol and Black Water metagenomes. Vital steps in the transformation and movement of nitrogen in the two environments were mapped based on the translated ORFs present in the metagenome. The predicted paths in the nitrogen cycle of the two acid ponds showed that the main biochemical reaction is reduction of nitrogen. The distinct absence of a nitrogen oxidation pathway and the prediction of allantoin utilization can be indicative of the strict nitrogen budget in the two extreme environments and the necessity for the residing extremophiles to maintain the bioavailablity of nitrogen (Vogels and Drift 1976). Further deduction of metabolism of ammonia to L-glutamate also showed that glutamate is the main precursor which contributes amine groups for the biosynthesis of amino acids, nucleic acid bases, and aminated carbohydrates in both sample sites (Kim and Gadd 2008; Meng 2016).

Inhibition of nitrification in hyper-saline environments suppresses denitrification and anammox, resulting in efficient ammonium recycling and also enriches ammonium in surface brine due to the release of ^{15} N-depleted NH3 gas (Isaji 2019). Proteins involved in nitrification were not predicted in this study. Partial and/or complete ORFs for all important orthologous proteins involved in denitrification were predicted in the metagenomes of both ponds. But in the Dallol metagenome, crucial protein coding genes for the final reduction of NO to N_2 were missing, causing incomplete mapping of denitrification. Still, it is not uncommon among extremophiles, especially thermophiles to undergo incomplete denitrification (Chen et al. 2002; Hedlund et al. 2011). On the other hand, protein coding genes, homologous to copper containing nitrite reductase (NirK) and nitrite reductase associated c-type cytochorome (NirN) were only predicted in the Dallol metagenome. The NirN has been proven to be an essential cofactor for the cytochrome associated nitrite reductase NirS, which is very important for catalyzing the reduction of NO_2^- to NO in denitrification respiration (Adamczack 2014; Ward 2001).

One of the most important functional annotations performed for the metagenome reads from both acid ponds is the deduction of nitrogen fixation. In principle, the prospect of genetic manipulation of plants to fix nitrogen independently and fulfill the energy demand has prompted researchers to explore different genetic resources and methods (Dixon 1997; Ivleva 2016; Li 2016; Burén et al. 2018). Fixing atmospheric nitrogen by the nitrogenase enzyme is a characteristic present only in diazotrophs and it influences the amount of nutrient present within an ecosystem (Hamilton 2011; Jones 2016; Messer 2016; Lesser 2018). In this study, in both acid ponds, the genes involved in nitrogen fixation were predicted only from *Proteobacteria* OTUs, particularly, from *Alphaproteobacteria.* As in the case of many diazotrophic bacteria, either in poly-extreme environment or not, the predicted nitrogenase from the metagenomes of Dallol and Black Water has two components; the NifH-[4Fe-4S] cluster and nitrogen fixation (nif) genes for the biosynthesis of Mo–Fe cofactor (Dixon 1997; McGlynn 2012).

Molecular and biochemical data are limited to construct the best model for biosynthesis of nitrogenase for all heterotrophic free-living nitrogen fixing bacteria (Lesser 2018). Still, a common understanding exists on how the environment determines the quantity and organization of nitrogen fixing genes in different diazotrophs (Li 2016; Messer 2016; Lesser 2018; Pedersen 2018; Pérez 2017). The broadly

Table 9 Swiss model programe parameters and Ramachandran Plot output of the two commonly found query proteins involved in nitrogen fixation

		Protein Model and PDB ID	Target and Template Identity (%)	Ramachandran plot distribution of residue		
				Most Favored (%)	Allowed (%)	Outliner (%)
Black Water	Node_4344_Cysteine_desulfurase	Cysteine desulfurase from *Legionella pneumophila Philadelphia* 1(6c9e.1.A)	53.88%	95.72%	0.76% [A463 VAL, B463 VAL, B199 HIS, A199 HIS, A464 PRO, B464 PRO]	0.32% [A385 GLU, B480 SER]
	WP_149425408.1:76–609 SufS family cysteine desulfurase	Cysteine desulfurase from *Legionella pneumophila Philadelphia 1*(6c9e.1.A)	51.13%	95.47%	0.88% [A448 VAL, B448 VAL, B184 HIS, A184 HIS, A513 PHE, A449 PRO, B449 PRO]	0.63% [A370 GLU, B465 SER, A419 LEU, B419 LEU]
	Node_27591_Ferredoxine	2[4FE-4S] ferredoxin from *Pseudomonas aeruginosa* (2fgo.1)	48.39%	88.71%	0.0%	0.0%
	WP_063195668.1:1–72 ferredoxin	2[4FE-4S] ferredoxin from Pseudomonas aeruginosa (2fgo.1)	53.23%	88.71%,	1.61% [A23 VAL]	3.51% [A33 VAL, A62 VAL]
Dallol	Nodes_5315_35911_Cysteine_desulfurase	Cysteine desulfurase from *Legionella pneumophila* Philadelphia 1(6c9e.1.A)	54.96%	95.65%	0.77% [A311 VAL, B311 VAL, B47 HIS, A47 HIS, A312 PRO, B312 PRO]	0.32% [A233 GLU, B328 SER]
	WP_149425408.1:213–605 SufS family cysteine desulfuras	Cysteine desulfurase from *Legionella pneumophila Philadelphia 1*(6c9e.1.A)	51.15%	96.55%	0.64% [A47 HIS, B47 HIS, B311 VAL, A312 PRO, B312 PRO]	0.97% [B213 VAL, A213 VAL, A233 GLU, B328 SER, A282 LEU, B282 LEU]
	Node_15557_Ferredoxine	2[4FE-4S] ferredoxin from Pseudomonas aeruginosa (2fgo.1)	48.39%	88.71%	0.0%	0.0%
	WP_063195668.1:1–72 ferredoxin	2[4FE-4S] ferredoxin from Pseudomonas aeruginosa (2fgo.1)	53.23%	88.71%	1.61% [A23 VAL]	3.51% [A33 VAL, A62 VAL]

accepted minimal gene cluster for Mo-dependent nitrogenase biosynthesis consists of nifDKH and nifBEN sets (McGlynn 2012; Pérez 2017; Black and Santos 2015; Wang 2013). However, the predicted sets of nif genes in the two metagenomes are quite different from each other even if both sample sites are hyper saline and highly acidic.

Differences among the types of accessory proteins involved in maturation of nitrogenase and trafficing molybdenum to nitrogenase was also observed in the two acid ponds. For instance, partial ORFs of nifQ and nifZ were predicted only in Dallol. The NifQ and NifZ proteins are known to be involved in trafficking molybdenum to nitrogenase and maturation of the P-clusters contained within the Mo–Fe protein, respectively (Meng 2016; Isaji 2019; Chen et al. 2002; Hedlund et al. 2011). On the other hand, complete ORFs of nifT and nifW and partial ORFs of nifA and nifN genes were predicted only in the Black Water. The NifW protein is involved in the maturation processes of the NifD protein, whereas NifA is an enhancer-binding protein that binds to specific DNA sequences upstream of nif genes and NifT has some suppressive effects on nitrogenase (Morett et al. 1991; Nonaka 2019).

Both blastx and blastp results showed that many of the predicted Nif proteins in both acid ponds are highly divergent with less than 85% average similarity of translated ORF sequences to orthologs of Nif proteins. As a principle to sequence analysis, more distant species will have more variable sequences of orthologous proteins (Koonin et al. 2003). Thus, divergent groups of free-living nitrogen fixing bacteria may inhabit these two extreme environments of the Danakil depression.

Further structural examination on translated ORFs for cysteine desulfurase and ferredoxin revealed that the predicted hypothetical proteins from both acid ponds are quite different from their respective orthologous reference proteins (Supplementary Table 3). Cysteine desulfurase (NifS/SufS), a homodimeric enzyme, mobilizes sulfur from cysteine via a pyridoxal 5′-phosphate (PLP) dependent mechanism and is known to be contributing in many biosynthesis reactions (Black and Santos 2015; Poudel et al. 2018; Kiyasu 2000). If NifS is absent, the activities of both nitrogenase component proteins can be negatively impacted (Poudel et al. 2018). Scrutiny of the pairwise amino acid sequence alignments with the reference protein along with the analysis of the corresponding structural models showed that functionally important residues in the predicted enzymes were conserved. In particular, the lysine residue with catalytic activity and other residues interacting as binding sites for the Pyridoxal 5′-phosphate are identically conserved in both predicted proteins as well as the reference (Table 7). The [4Fe-4S] ferridoxines predicted from both Dallol and Black Water have identical ORFs and the closest homologous protein is the ferredoxin of *Bradyrhizobium sp. AT1*(WP_063195668.1) with only 83% amino acid sequence identity. This protein is an important, but not the primary, electron donor for dinitrogenase reductase (Poudel et al. 2018; Egener et al. 2001). Yet, it is known to have an effect on the rapid "switch-off" of nitrogenase activity in response to ammonium (Egener et al. 2001).

The nodes with ORFs, representing query proteins cysteine desulfurase and ferredoxin from Dal have more than 80% identity with conserved active sites to the reference proteins, while 69.41% and 83.3% similarity was observed between the query proteins from Bla and their respective reference proteins (Supplementary Table 1). Normally, amino acid sequence similarity less than 80% is more difficult to reliably infer similar function (Pearson 2013). However, according to the secondary and tertiary structure analyses, all the important residues on the sites of major twists and turns as well as the active sites are present to maintain the structural and functional integrity of the predicted proteins (Tables 5, 7, 8). Hence, regardless of the apparent dissimilarity of primary amino acid sequences of the query proteins to the reference proteins, the similarities of the secondary and 3-D structures of the predicted cysteine desulfurase and ferredoxin from Dallol and Black Water to the reference proteins clearly indicated functional orthology (Pearson 2013) (Table 9). The Black water metagenome provided us with many complete ORFs to further study possible novel Nif proteins. The predicted proteins (NifB, NifT and NifW) amino acid sequences were dissimilar to the closest homologous proteins (Supplementary Tables 1 and 3). Just like in the case of cysteine desulfurase, functionally important residues in the predicted NifB were conserved (Table 7).

Conclusion

Combined techniques of physical and chemical cells lysis for a heterogeneous microbial community and whole community metagenome sequencing were effective to capture and identify extremophilic inhabitants of Dallol and Black Water acid ponds. The two acid ponds (Dallol and Black Water) are as different in their microbial community composition as they are physicochemically distinct. Each of the ponds have distinct biotia flourishing in them but *Proteobacteria* is dominant. We have identified unique biota associated with each pond, suggesting that these extreme environments may require unique gene sets as in the case especially of nitrogen metabolism. By and large, from the metagenome data, it is logical to deduce the presence of different strategies adopted by the extremophiles for trafficking metals and maturation of nitrogenase enzymes. Therefore, the modeling of these enzymatic pathways from these two extreme environments will be a great input for genetic engineering or synthetic biology. These data support the value of extremophile analysis in naturally occurring poly-extreme environments and suggests that additional analyses of samples from these regions both bioinformatically as well as in vitro, are warranted.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00203-021-02227-4.

Acknowledgements GMW acknowledges support from the National Institutes of Health 1R01GM132222-01.

Author contributions LT designed experiments, collected samples, analyzed data, and wrote drafts of the manuscript. AA, and AS helped designed experiments, collected samples, and edited drafts of the manuscript. GMW helped in DNA isolation, sequencing, and analysis and helped in writing and editing the manuscript.

Data availability The data underlying this article will be shared on reasonable request to the corresponding author.

Compliance with ethical standards

Conflict of interest The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this work.

References

Spear JR, Walker JJ, Pace NR (2006) Microbial ecology and energetics in yellowstone hot springs. Yellowstone Sci 14(1):17–24

Mayer K et al (2017) Phreatic activity and hydrothermal alteration in the Valley of Desolation, Dominica, Lesser Antilles. Bull Volcanol 79(12):82

Crognale S et al (2018) Microbiome profiling in extremely acidic soils affected by hydrothermal fluids: the case of the Solfatara Crater (Campi Flegrei, southern Italy). FEMS Microbiol Ecol 94(12)

Illsley-Kemp F et al (2017) Local Earthquake Magnitude Scale and b-Value for the Danakil Region of Northern Afar. Bull Seismol Soc Am pp 107

Franzson H, Helgado HM, Oskarsson F (2015) Surface Exploration and First Conceptual Model of the Dallol Geothermal Area, Northern Afar, Ethiopia. In: Proceedings World Geothermal Congress

Asrat A (2016) The Danakil depression: an exceptional place where different types of extreme environments coexist. pp 189–196

Carrizo D et al (2019) Lipid biomarker and carbon stable isotope survey on the dallol hydrothermal system in Ethiopia. Astrobiology 19(12):1474–1489

Gómez F et al (2019) Ultra-small microorganisms in the polyextreme conditions of the Dallol volcano, Northern Afar, Ethiopia. Sci Rep 9:7907. https://doi.org/10.1038/s41598-019-44440-8

Damer B (2016) A field trip to the Archaean in search of Darwin's Warm Little Pond. Life (Basel, Switzerland) 6(2):21

Van Kranendonk MJ, Deamer DW, Djokic T (2017) Life springs. Sci Am 317(2):28–35

Wächtershäuser G (2006) *From volcanic origins of chemoautotrophic life to Bacteria, Archaea and Eukarya.* Philos Trans R Soc Lond B Biol Sci 361(1474):1787–806; discussion 1806–8

Barbieri R, Cavalazzi B (2014) How Do Modern Extreme Hydrothermal Environments Inform the Identification of Martian Habitability? The Case of the El Tatio Geyser Field. Challenges 5:430–443

van den Burg B (2003) Extremophiles as a source for novel enzymes. Curr Opin Microbiol 6(3):213–218

Hamdi M (2007) Microbial Resources and Industrial microbial processes design and behavior. Current Research, Technology and Education Topics in Applied Microbiology and Microbial Biotechnology, A. Mendez-Vilas, Editor. 2007.

Kikani B, Shukla R (2010) and S. Biocatalytic potential of thermophilic bacteria and actinomycetes, Singh, pp 1000–1007

Börne RA (2013) Exploring anaerobic bacteria for industrial biotechnology: diversity studies, screening and biorefinery applications. Lund University, Sweden

Li SJ et al (2014) Microbial communities evolve faster in extreme environments. Sci Rep 4:6205

Schmeisser C, Steele H, Streit WR (2007) Metagenomics, biotechnology with non-culturable microbes. Appl Microbiol Biotechnol 75(5):955–962

Su C et al (2012) Culture-independent methods for studying environmental microorganisms: methods, application, and perspective. Appl Microbiol Biotechnol 93(3):993–1003

Barriuso J, Martínez MJ (2015) In silico metagenomes mining to discover novel esterases with industrial application by sequential search strategies. J Microbiol Biotechnol 25(5):732–737

Kodzius R, Gojobori T (2015) Marine metagenomics as a source for bioprospecting. Mar Genom 24(Pt 1):21–30

Alves L, Silva-Rocha R, Guazzaroni M-E (2017) Enhancing metagenomic approaches through synthetic biology. In: Charles T, Liles M, Sessitsch A (Eds) Functional Metagenomics: Tools and Applications, Springer, New York, pp 75–94

Gebresilassie S, Gebretsadik HT, Kabeto K (2011) Preliminary study on geology, mineral potential and characteristics of hot springs from Dallol area, Afar rift, northeastern Ethiopia: Implications for natural resource exploration. Momona Ethiopian J Sci. https://doi.org/10.4314/mejs.v3i2.67710

Kotopoulou E et al (2019) A polyextreme hydrothermal system controlled by iron: the case of Dallol at the Afar triangle. ACS Earth Space Chem 3(1):90–99

Hamilton TL et al (2011) Biological nitrogen fixation in acidic high-temperature geothermal springs in Yellowstone National Park, Wyoming. Environ Microbiol 13(8):2204–2215

Mitchell KR, Takacs-Vesbach CD (2008) A comparison of methods for total community DNA preservation and extraction from various thermal environments. J Ind Microbiol Biotechnol 35(10):1139–1147

Emanuelsson BD et al (2015) High-resolution continuous-flow analysis setup for water isotopic measurement from ice cores using laser spectroscopy. Atmos Meas Tech 8(7):2869–2883

Zhou J, Bruns MA, Tiedje JM (1996) DNA recovery from soils of diverse composition. Appl Environ Microbiol 62(2):316–322

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120

Segata N et al (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 9(8):811–814

Truong DT et al (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 12(10):902–903

Waterhouse A et al (2018) SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 46(W1):W296-w303

Legendre P, Legendre L (1998) Numerical ecology, 2nd edn. Elsevier, Amsterdam

Nurk S et al (2017) metaSPAdes: a new versatile metagenomic assembler. Genome Res 27(5):824–834

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Methods 12(1):59–60

Bağcı C et al (2019) Introduction to the analysis of environmental sequences: metagenomics with MEGAN. Methods Mol Biol 1910:591–604

Huson DH et al (2018) MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. Biol Direct 13(1):6

Huson DH et al (2016) MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput Biol 12(6):e1004957

Kanehisa M, Sato Y (2020) KEGG Mapper for inferring cellular functions from protein sequences. Protein Sci 29(1):28–35

Gasteiger E et al (2005) *Protein Identification and Analysis Tools on the ExPASy Server*. In: Walker JM (ed) Humana Press, pp 571–607

Geourjon C, Deléage G (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput Appl Biosci 11(6):681–684

Garnier J, Gibrat JF, Robson B (1996) GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 266:540–553

Ghebrezgabher MG, Yang T, Yang X (2016) Long-term trend of climate change and drought assessment in the Horn of Africa. Adv Meteorol 2016:8057641

Klotz MG et al (2016) Editorial: systems biology and ecology of microbial mat communities. Front Microbiol 7:115

Nobile A et al (2012) Dike-fault interaction during the 2004 Dallol intrusion at the northern edge of the Erta Ale Ridge (Afar, Ethiopia). Geophys Res Lett 39(19).

Hallsworth JE et al (2007) Limits of life in MgCl2-containing environments: chaotropicity defines the window. Environ Microbiol 9(3):801–813

Herbstein FH, Kapon M, Weissman A (1982) X-ray diffraction as a tool for studying stoichiometry and kinetics of solid state thermal decomposition reactions. Application to the Thermal Decomposition of Bischofite MgCl2 · 6H2O. Israel J Chem 22(3):207–213

HCl Leaching and Acid Regeneration (2011) Using MgCl2 Brines and Molten Salt Hydrates, in EPD Congress 2011, pp 520–528

Roux S et al (2011) Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (Bacteria and Archaea) in ecosystems. FEMS Microbiol Ecol 78(3):617–628

Poretsky R et al (2014) Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PLoS ONE 9(4):e93827

Ranjan R et al (2016) Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun 469(4):967–977

Bag S et al (2016) An improved method for high quality metagenomics DNA extraction from human and environmental samples. Sci Rep 6:26775

Ketchum RN et al (2018) DNA Extraction Method Plays a Significant Role When Defining Bacterial Community Composition in the Marine Invertebrate Echinometra mathaei. Front Mar Sci **5**(255).

Belilla J et al (2019) Hyperdiverse archaea near life limits at the polyextreme geothermal Dallol area. Nat Ecol Evol 3(11):1552–1561

Johnson DB et al (2001) Isolation and phylogenetic characterization of acidophilic microorganisms indigenous to acidic drainage waters at an abandoned Norwegian copper mine. Environ Microbiol 3(10):630–637

Méndez-García C et al (2015) Microbial diversity and metabolic networks in acid mine drainage habitats. Front Microbiol 6:475

Mesa V et al (2017) Bacterial, archaeal, and eukaryotic diversity across distinct microhabitats in an acid mine drainage. Front Microbiol 8:1756

Schuler CG, Havig JR, Hamilton TL (2017) Hot spring microbial community composition, morphology, and carbon fixation: implications for interpreting the ancient rock record. Front Earth Sci 5(97)

Abreu F et al (2018) Culture-independent characterization of a novel magnetotactic member affiliated to the Beta class of the Proteobacteria phylum from an acidic lagoon. Environ Microbiol 20(7):2615–2624

Li, Y., et al., *Coupled Carbon, Sulfur, and Nitrogen Cycles Mediated by Microorganisms in the Water Column of a Shallow-Water Hydrothermal Ecosystem.* Frontiers in Microbiology, 2018. **9**(2718).

Widdison, P.E. and T.P. Burt, *Nitrogen Cycle*, in *Encyclopedia of Ecology* S.E. Jørgensen and B.D. Fath, Editors. 2008, Academic Press. p. 2526–2533.

Vogels GD, Van der Drift C (1976) Degradation of purines and pyrimidines by microorganisms. Bacteriol Rev 40(2):403–468

Kim, B.H. and G.M. Gadd, *Degradation of nucleic acid bases*, in *Bacterial Physiology and Metabolism* 2008, Cambridge University Press.: New York. p. 135- 223.

Meng L et al (2016) Effects of sucrose amendment on ammonia assimilation during sewage sludge composting. Bioresour Technol 210:160–166

Isaji Y et al (2019) Efficient recycling of nutrients in modern and past hypersaline environments. Scientific Reports 9(1):3718

Chen, M.Y., et al., *Pseudoxanthomonas taiwanensis sp. nov., a novel thermophilic, N2O-producing species isolated from hot springs.* Int J Syst Evol Microbiol, 2002. **52**(Pt 6): p. 2155–61.

Hedlund, B.P., et al., *Potential role of Thermus thermophilus and T. oshimai in high rates of nitrous oxide (N2O) production in ~80 °C hot springs in the US Great Basin.* Geobiology, 2011. **9**(6): p. 471–80.

Adamczack J et al (2014) NirN protein from Pseudomonas aeruginosa is a novel electron-bifurcating dehydrogenase catalyzing the last step of heme d1 biosynthesis. J Biol Chem 289(44):30753–30762

Ward TW et al (2001) Characterization of the Structural Gene Promoter of Aedes aegypti Densovirus. J Virol 75(3):1325–1331

Dixon R et al (1997) Nif gene transfer and expression in chloroplasts: Prospects and problems. Plant Soil 194(1):193–203

Ivleva NB et al (2016) Expression of Active Subunit of Nitrogenase via Integration into Plant Organelle Genome. PLoS ONE 11(8):e0160951

Li XX et al (2016) Using synthetic biology to increase nitrogenase activity. Microb Cell Fact 15:43

Burén S, López-Torrejón G, Rubio LM (2018) Extreme bioengineering to meet the nitrogen challenge. Proc Natl Acad Sci U S A 115(36):8849–8851

Jones FP et al (2016) Novel European free-living, non-diazotrophic Bradyrhizobium isolates from contrasting soils that lack nodulation and nitrogen fixation genes - a genome comparison. Scientific reports 6:25858–25858

Messer LF et al (2016) High levels of heterogeneity in diazotroph diversity and activity within a putative hotspot for marine nitrogen fixation. Isme j 10(6):1499–1513

Lesser MP et al (2018) Diazotroph diversity and nitrogen fixation in the coral Stylophora pistillata from the Great Barrier Reef. Isme j 12(3):813–824

McGlynn SE et al (2012) Classifying the metal dependence of uncharacterized nitrogenases. Front Microbiol 3:419

Pedersen JN et al (2018) Diazotrophs and N(2)-Fixation Associated With Particles in Coastal Estuarine Waters. Front Microbiol 9:2759

Pérez CA et al (2017) Biological nitrogen fixation in a post-volcanic chronosequence from south-central Chile. Biogeochemistry 132(1):23–36

Black KA, Dos Santos PC (2015) Shared-intermediates in the biosynthesis of thio-cofactors: Mechanism and functions of cysteine desulfurases and sulfur acceptors. Biochim Biophys Acta 1853(6):1470–1480

Wang S et al (2013) Transcriptome sequencing of Zhikong scallop (Chlamys farreri) and comparative transcriptomic analysis with Yesso scallop (Patinopecten yessoensis). PLoS ONE 8:e63927

Morett E, Fischer HM, Hennecke H (1991) Influence of oxygen on DNA binding, positive control, and stability of the Bradyrhizobium japonicum NifA regulatory protein. J Bacteriol 173(11):3478–3487

Nonaka A et al (2019) Accessory Proteins of the Nitrogenase Assembly, NifW, NifX/NafY, and NifZ, Are Essential for Diazotrophic Growth in the Nonheterocystous Cyanobacterium Leptolyngbya boryana. Front Microbiol 10:495

Koonin EV, Galperin MY, in Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. (2003) Kluwer Academic Copyright © 2003. Kluwer Academic, Boston

Poudel, S., et al., *Electron Transfer to Nitrogenase in Different Genomic and Metabolic Backgrounds.* J Bacteriol, 2018. **200**(10).

Kiyasu T et al (2000) Contribution of cysteine desulfurase (NifS protein) to the biotin synthase reaction of Escherichia coli. J Bacteriol 182(10):2879–2885

Egener, T., et al., *Role of a ferredoxin gene cotranscribed with the nifHDK operon in N(2) fixation and nitrogenase "switch-off" of Azoarcus sp. strain BH72.* J Bacteriol, 2001. **183**(12): p. 3752–60.

Pearson, W.R., *An Introduction to Sequence Similarity ("Homology") Searching.* Current Protocols in Bioinformatics, 2013. **42**(1): p. 3.1.1–3.1.8.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Terms and Conditions